Development of Annotation Guidelines, Datasets and Deep Networks for Palm Leaf Manuscript Layout Understanding

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Electronics and Communication Engineering by Research

by

Sowmya Aitha
2018702007
sowmya.aitha@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2023

Copyright © Sowmya Aitha, 2023

All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Development of Annotation Guidelines, Datasets and Deep Networks for Palm Leaf Manuscript Layout Understanding**" by Sowmya Aitha, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Ravi Kiran Sarvadevbhatla

To my dearest Family and Friends!

Acknowledgments

I would like to take this chance to express my heartfelt thanks to everyone who helped me along the way in any manner. To begin with, I would want to express my sincere gratitude to my professor Dr. Ravi Kiran for leading me in every step and helping me stay on the right path despite any challenging circumstances I encountered. I shall treasure the lessons I have learned while being under his direct supervision for the rest of my life.

Secondly, I want to express special thanks to my team, without whom I could not have accomplished this. I thank my co-authors Abhishek Prusty, Abhishek Trivedi, Sharan, and Sindhu's cooperation and support in getting the papers published. I am delighted I had the opportunity to work with you. I have learned many new things from each of you. I recognise and thank everyone who were with me during the journey, especially Aaron, Abhinav, Pranav, Niharika, Khadiravana, Manaswini, Surendra, Mourya, and Vamshi. It was laborious to annotate so many documents. You guys helped me meet the deadline by being immensely understanding and supportive of me at this time.

In addition, I want to thank my friends Prathyusha, Kiruthika, Mythri, Tanu, Anoop and others for making my college experience unforgettable and exciting. I am very grateful to have found friends like you who have helped, encouraged, and supported me in my coursework and research. I did learn a lot from you guys as well. The times we spent together have always been wonderful to me.

Finally, I would like to express my gratitude to my family for all of their efforts, love, and support. I can never thank them enough for what they have done. I am eternally thankful to my amma, nanna, anna, and chelli. Without their help, I could not have progressed this far or attained the things that I have. I am sure there are a lot of people other than mentioned, who have directly or indirectly helped me on this path. I am sincerely grateful to each of them for their support.

Abstract

Ancient paper documents and palm leaf manuscripts from the Indian subcontinent have made a significant contribution to the world literary and culture. These documents often have complex, uneven, and irregular layouts. The process of digitization and deciphering the content from these documents without human intervention pose difficulties in a broad range of areas, including language, script, layout, elements, position, and number of manuscripts per image.

Large-scale annotated Indic manuscript image datasets are needed for this kind of research. In order to meet this objective, we present Indiscapes, the first dataset containing multi-regional layout annotations for ancient Indian manuscripts. We also adapt a fully convolutional deep neural network architecture for fully automatic, instance-level spatial layout parsing of manuscript images in order to deal with the challenges such as presence of dense, irregular layout elements, pictures, multiple documents per image and the wide variety of scripts. Eventually, We demonstrate the effectiveness of proposed architecture on images from the Indiscapes dataset.

Despite advancements, the segmentation of semantic layout using typical deep network methods is not resistant to the complex deformations that are observed across semantic regions. This problem is particularly evident in the domain of Indian palm-leaf manuscripts, which has limited resources. Therefore, we present Indiscapes2, a new expansive dataset of various Indic manuscripts with semantic layout annotations, to help address the issue. Indiscapes2 is 150% larger than Indiscapes and contains materials from four different historical collections. In addition, we propose a novel deep network called Palmira for reliable, deformation-aware region segmentation in handwritten manuscripts. As a performance metric, we additionally report a boundary-centric measure called Hausdorff distance and its variations. Our tests show that Palmira offers reliable layouts and outperforms both strong baseline methods and ablative versions. We also highlight our results on Arabic, South-East Asian and Hebrew historical manuscripts to showcase the generalization capability of PALMIRA.

Even though we have reliable deep-network based approaches for comprehending manuscript layout, these models implicitly assume one or two manuscripts per image during the process, whereas in a real-world scenario there are often cases where multiple manuscripts are typically scanned together into a scanned image to maximise scanner surface area and reduce manual labour. Now, making sure that each individual manuscript within a scanned image can be isolated (segmented) on a per-instance basis became the first essential step in understanding the content of a manuscript. Hence, there is a need for a precursor system which extracts individual manuscripts before downstream processing. The highly curved and deformed boundaries of manuscripts, which frequently cause them to overlap with each other, introduce another complexity when confronting issue. We introduce another new document image dataset named IMMI (Indic Multi Manuscript Images) to address these issues. We also present a method that generates synthetic images to augment sourced non-synthetic images in order to boost the efficiency of the dataset and facilitate deep network training. Adapted versions of current document instance segmentation frameworks are used in our experiments. The results demonstrate the efficacy of the new frameworks for the task. Overall, our contributions enable robust extraction of individual historical manuscript pages. This in turn, could potentially enable better performance on downstream tasks such as region-level instance segmentation, optical character recognition and word-spotting in historical Indic manuscripts at scale.

Contents

Chapter Page

List of Figures

Figure Page

List of Tables

Table Page

Chapter 1

Introduction

Ancient palm leaf manuscripts are historical artefacts that hold a plethora of data. They are an effective means of preserving cultural heritage as they represent a multitude of traditions, history, and information. This is especially true for manuscripts from the Indian subcontinent and South-East Asian countries [43]. These are referred to as “Indic Manuscripts” throughout this thesis. While previous conservation efforts have focused on digitising these fragile texts, the work presented in this thesis takes it a few steps ahead by creating large-scale datasets of such manuscript images and developing deep learning models that can automatically identify and label different regions of the document, such as character line segments, library markers, pictures, and textual and non-textual elements. We also made an effort to determine the page boundaries from a single scanned image containing multiple manuscripts and were successful to a greater extent.

1.1 Indic Manuscripts and Importance

The Indic manuscript collection is considered one of the largest manuscript collections globally, estimated at around ten million manuscripts. Although Ayurveda and Yoga have become the most influential and popular Indian exports to the West and other parts of the globe, there are many more buried and preserved shastras, or texts, on a wide range of subjects, spanning from astrology, literature, science and technology, to wellness and ecology. Unfortunately, the early means of writing down such a large amount of information was done on materials such as stone, parchment, birch bark, and palm leaves, all of which are now fragile and fractured due to weather and time.

In comparison to modern or recent era documents, such manuscripts are significantly more fragile, susceptible to degradation from natural elements, and have a relatively short shelf life [47, 67, 74]. In particular to Indic documents, the manuscripts exist in multiple languages and scripts based on various factors such as geography and timelines. The number of domain experts who can comprehend such literature is shrinking rapidly. Hence, it is very vital to achieve access to the information contained in these manuscripts before it is permanently lost.

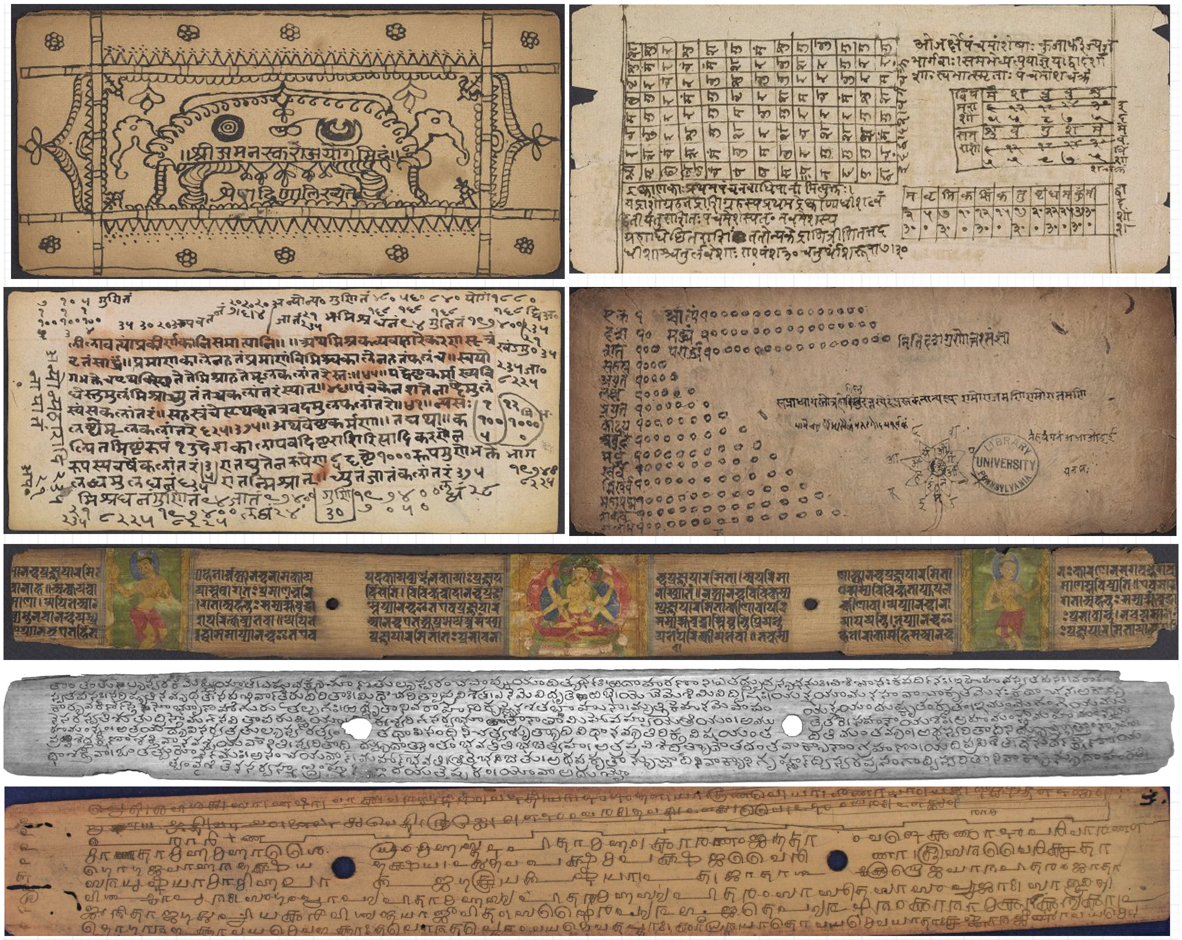


Figure 1.1: Sample images of ancient handwritten documents and palm leaf manuscripts

1.2 Importance of Developing Systems for Studying Indic Manuscripts

Given the multitude of languages, scripts, and non-textual regional elements found in Indian manuscripts, spatial layout parsing is crucial for enabling downstream applications such as optical character recognition (OCR), word-spotting, style-and-content based retrieval, and clustering. For this reason, we first tackle the problem of creating a diverse, annotated spatial layout dataset. Creating such a dataset enables progress and bypasses the hurdle of language and script familiarity for annotators.

In general, manuscripts from Indian subcontinent pose many unique challenges. To begin with, the documents exhibit a large multiplicity of languages. This is further magnified by variations in intra-language script systems. Along with text, manuscripts may contain pictures, tables, non-pictorial decorative elements in non-standard layouts (refer to figure 1.1). A unique aspect of Indic and South-East Asian manuscripts is the frequent presence of holes punched in the document for the purpose of binding [47, 74, 89]. These holes cause unnatural gaps within text lines. The physical dimensions of the manuscripts are typically smaller compared to other historical documents, resulting in a dense content layout. Sometimes, multiple manuscript pages are present in a single image. Moreover, imaging-related factors such as varying scan quality play a role as well. Given all of these challenges, it is important to develop robust and scalable approaches for the problem of layout parsing. In addition, given the typical non-technical nature of domain experts who study manuscripts, it was also important to develop easy-to-use graphical interfaces for annotation, post-annotation visualization and analytics.

1.3 Works on Semantic and Instance Segmentation

A number of research groups have invested significant efforts in the creation and maintenance of annotated, publicly available historical manuscript image datasets [39, 62, 68, 71, 75, 82, 83]. Other collections contain character-level and word-level spatial annotations for South-East Asian palm-leaf manuscripts [40, 86, 89]. In these latter set of works, annotations for lines are obtained by considering the polygonal region formed by union of character bounding boxes as a line. While studies on Indic palm-leaf and paper-based manuscripts exist, these are typically conducted on small and often, private collections of documents [5,19,61,76,77,80,85]. No publicly available large-scale, annotated dataset of historical Indic manuscripts exists to the best of our knowledge. In contrast to existing collections, our proposed dataset contains a much larger diversity in terms of document type (palm-leaf and early paper), scripts and annotated layout elements (see Tables 3.3,3.4 and 3.5). An additional level of complexity arises from the presence of multiple manuscript pages within a single image.

A number of contributions can also be found for the task of historical document layout parsing [16, 17, 94, 95]. Wei et al. [94] explore the effect of using a hybrid feature selection method while using autoencoders for semantic segmentation in 5 historical English and Medieval European manscript datasets. Chen et al. [17] explore the use of Fully Convolutional Networks (FCN) for the same datasets. Barakat et al. [12] propose a FCN for segmenting closely spaced, arbitrarily oriented text lines from an Arabic manuscript dataset. The mentioned approaches, coupled with efforts to conduct competitions on various aspects of historical document layout analysis have aided progress in this area [3, 4, 42]. A variety of layout parsing approaches, including those employing the modern paradigm of deep learning, have been proposed for Indic [73, 76, 80, 85] and South-East Asian [16, 43, 63, 86, 91] palm-leaf and paper manuscript images. However, existing approaches typically employ brittle hand-crafted features or demonstrate performance on datasets which are limited in terms of layout diversity. Similar to many recent works, we employ Fully Convolutional Networks in our approach. However, a crucial distinction lies in our formulation of layout parsing as an *instance* segmentation problem, rather than just a *semantic* segmentation problem. This avoids the problem of closely spaced layout regions (e.g. lines) being perceived as contiguous blobs.

The ready availability of annotation and analysis tools has facilitated progress in creation and analysis of historical document manuscripts [24, 28, 32]. The tool we propose here contains many of the features found in existing annotation systems. However, some of these systems are primarily oriented towards single-user, offline annotation and do not enable a unified management of annotation process and monitoring of annotator performance. In contrast, our web-based system addresses these aspects and provides additional capabilities. Many of the additional features in our system are tailored for annotation and examining annotation analytics for documents with dense and irregular layout elements, especially those found in Indic manuscripts. In this respect, our annotation system is closer to the recent trend of collaborative, cloud/web-based annotation systems and services [2, 33, 96].

1.4 Missing Resources

Despite the great significance of Indic manuscripts, there was no large scale dataset existing in the community to work with, when it came to obtaining access to the content. Due to the unique challenges that these manuscripts pose, there was no existing system that could overcome them, and parse the scanned manuscript document to retrieve the content from it. Hence there was a need for establishment of large scale datasets of Indic manuscript images and deep learning models to parse these images.

1.5 Thesis Contributions

In this work, we established standardised datasets following set of guidelines with Indic manuscript images and explored semantic instance segmentation based works for layout parsing of the same images.The contribution of the thesis are outlined as below.

- Development of guidelines for annotating the dataset with an emphasis on ensuring consistent annotations throughout.
- Datasets (Indiscapes, Indiscapes2 and IMMI) involving all the major possible real time challenges for layout parsing of the Indic Historical Manuscripts.
- Deep learning based Instance segmentation models for Layout parsing of the Historical Documents.

1.6 Thesis Organisation

The thesis is organised as follows. The requirements and considerations for developing a spatial layout annotation system are briefly summarised in chapter 2. It also includes a description of HInDoLA, a web-based system that constitutes a large-scale annotation tool, dashboard analytics, and machine learning engines. The proposed databases of Indic Manuscript images, Indiscapes and Indiscapes2, are discussed in chapter 3. The motivation for collecting the IMMI dataset and the analysis of the dataset are summarised in chapter 4. Deep networks for manuscript region instance segmentation, such as Mask R-CNN and Palmira, are presented in chapter 5, while deep networks for multi-manuscript document image segmentation are presented in chapter 6. Finally, the contributions, findings, experiments and analysis are summarised in chapter 7.

Chapter 2

Spatial Layout Annotation

Considering spatial layout annotation of manuscripts do not require any specific skill for identifying a language or script, unlike text annotation, setting up a diverse annotated spatial layout dataset overcomes the barrier of language and script familiarity for annotators. Although annotators do not need to be script or language experts, they should be able to recognise and categorise the most prominent and content-rich components such as character line segments, character components, pictures, and others precisely. To assist with this annotation, a web-based online annotation tool called HInDoLA, introduced by Trivedi et al. [88] was utilised. A set of guidelines were established while annotating to ensure that they stayed meaningful and consistent throughout the process. In the following sections, a quick overview of the tool and the established guidelines will be discussed. All annotation choices, such as manuscript sources, region types, and number of manuscripts, will also be provided in detail in the upcoming sections.

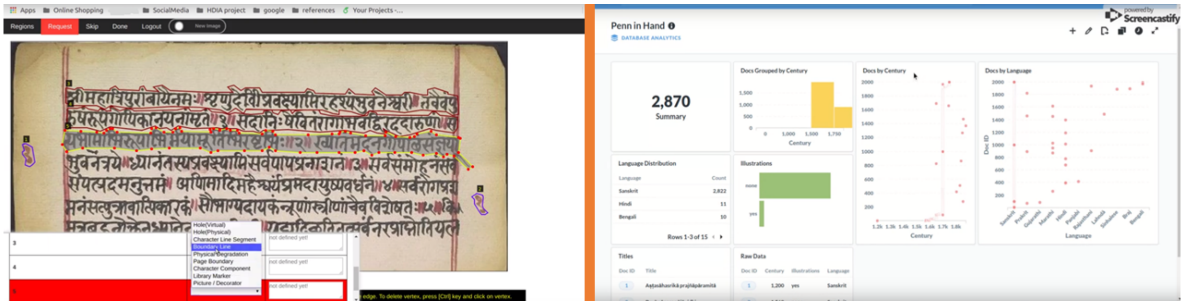
2.1 Background Work

2.1.1 HInDoLA

The availability and accessibility of annotation and analytic tools have greatly aided in the creation and study of historical manuscript annotations [24, 28, 32]. HInDoLA (Historical Intelligent Document Layout Analytics), the adopted annotation tool and analytics system, by Trivedi et al [88] incorporates many features available in existing annotation systems. However, most systems are built for single-user online annotation and do not offer centralized annotation management or annotator performance monitoring. HInDoLA, on the other hand, expands the tool's capabilities by resolving the concerns raised above and focusing on a more efficient annotation process and statistical analysis for texts with dense and irregular layout features, such as those found in Indian manuscripts. HInDoLA is more in line with the current trend of collaborative, cloud-based annotation systems and services [2, 33, 96] because it is entirely open source and comes with clear, step-by-step instructions for streamlined installation and usage. The Annotation Tool, ML Engine and Dashboard Analytics are crucial components of HInDoLA's overall architecture.

2.1.1.1 ML Engines

HInDoLA is set up to interact with machine-learning models that use deep networks. One of the models used here is fully automated and uses Mask R-CNN, a state-of-the-art object-instance segmentation framework. This model outputs a layout estimate at the instance level for a given document. A semi-automatic model is one in which the annotator supplies partial information in the form of the region's bounding box, and the model predicts a tight region boundary estimate. This bounding box supervision model learns the edge features and generates the boundary around the region. To access these intelligent models in the tool, a toggle button must be enabled. Both fully automatic and semi-automatic models become active once this toggle button is enabled after receiving the image from the tool. A sub-system provides control points on the boundaries of regions after the fully automatic model predicts masks for distinct regions in the input image. These control points are superimposed on the input image, allowing the annotator to tweak the prediction for a tighter region boundary. The control points can be adjusted by annotators to create a tight bounding box around the region. In the semi-automatic model, annotators draw a bounding box around a region in a freshly requested image, which is then automatically forwarded to the back-end edge model. A convolution neural network predicts the region's edge features and edge-logits (pixel values ranging from 0-1) and generates a concave-hull boundary from the most prominent features. Similar to the instance segmentation model boundary prediction, the boundary is then combined with control points, allowing the region boundary annotation to be served on the annotation tool. This feature saves users and experts time that would otherwise be spent annotating an entire document image from scratch.



edge. To delete vertex, press [Ctrl] key and click on vertex.

Language Distribution

Language	Count
Sanskrit	2,822
Hindi	11
Bengali	10

Rows 1-3 of 15

Titles

Doc ID	Title
1	Aṣṭasāhasrikā prajñāpāramitā

Raw Data

Doc ID	Century	Illustrations	Language
1	1,200	yes	Sanskrit

Figure 2.1: Screenshots of our web-based annotator (left) and analytics dashboard (right).

2.1.1.2 Annotation Tool

The Annotation tool can be used as a standalone application or as a web service. The standalone version can be used offline for sand boxing and enhancing the tool's capabilities. It is also used when the server-based components of HInDoLA cannot be installed owing to access constraints. The web service, on the other hand, supports distributed parallel sessions by registered annotators, as well as live session monitoring via the dashboard and shows various annotation-related statistics.

The web-based tool service allows users to engage in the interactive annotation of manuscripts. Users need to complete a one-time registration to track their annotation performance at multiple levels. Along with the standard polygon and rectangle drawing features, the free-hand drawing option is also introduced for quick annotations on irregular, small, and extremely big regions in manuscripts, giving the maximum annotation versatility possible. A single annotation is created by starting with a single mouse click and running the cursor smoothly along the component boundary, with a few more clicks in between the movement to create vertices of the boundary.

The regions in Indic manuscripts are often densely populated, making the annotators' jobs challenging. Therefore, features such as ultra zoom and spanning features were introduced to overcome this challenge and aid in pixel-level accuracy of annotations. Once a boundary is drawn around the region, it can be altered later by adding or removing the vertices with a single mouse click and the control button on the vertex. After drawing the boundary, a pop-up annotation window opens to label the region. Ten different class regions were identified and considered for annotation as character line segment, character component, hole, page boundary, library marker, decorator, picture, physical degradation, and boundary line considering the whole Indic manuscript dataset. The annotations for different class regions are shown in different vibrant colours for better visualisation and understanding(refer fig 2.3). The JSON format is used to extract all the annotations from the tool. Along with the region coordinates and labels, the JSON file also stores a time-stamp for each region annotated, which serves as a metric to support the training of intelligent models for automatic annotation.

The annotation manager was developed to manage and update the database in the backend while handling distributed concurrent sessions of registered annotators. It has a request queue that can handle several user sessions on the system simultaneously, and it can load unannotated images in normal mode and annotated images in correction mode. Expert annotators can use the "Correction" mode toggle to improve the quality of their annotations by correcting them. If the image is excessively corrupted or the annotator is unsure about the proper layout parsing of newly introduced components, the "skip button"" allows them to skip the served image. The annotation JSON file is saved in the backend folder when the "done" button is clicked.

2.1.2 Dashboard Analytics

The dashboard-style analytics includes many annotation management services, such as displaying graphs for the annotation stats, the annotators' progress reports, and the database viewer. It allows users to keep track of live annotation sessions and interactively explore document annotation statistics. The manuscripts are organized and displayed by annotated time, languages, and number of regions. There are various types of histograms and pie charts on the dashboard that represent the number of experiments added and completed document annotations. The viewer section of the dashboard enables the annotators to check the quality of completed annotations. Visit http://ihdia.iiit.ac.in/ for more information.

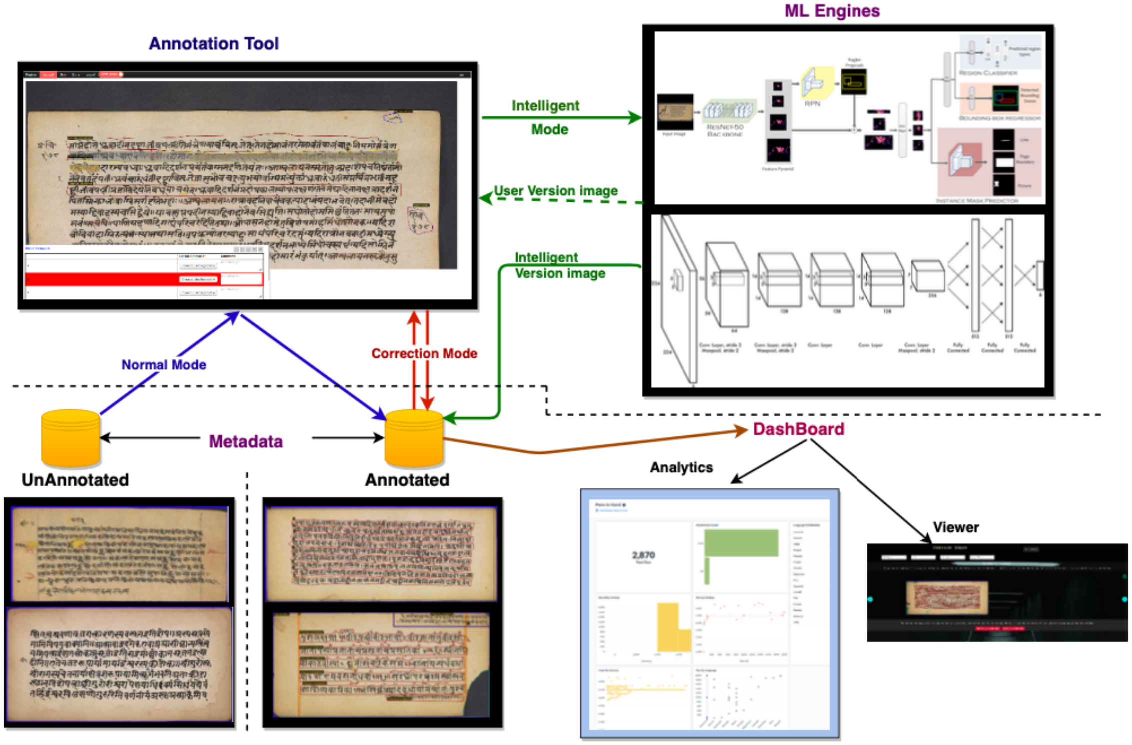


Figure 2.2: Architecture of HInDoLA

2.2 Considerations for Spatial Layout Annotations

Spatial layout parsing is critical for downstream applications such as OCR, word-spotting, and style- and content-based retrieval clustering, as the Indic manuscripts include a wide range of language, script, and non-textual regional features. Hence, we approach the problem of establishing a diversified, annotated spatial layout dataset because it has the advantage of immediately overcoming the barrier of language and script familiarity for annotators. Spatial annotations do not demand any expertise from annotators, unlike text annotations.

Surprisingly, there are currently no large-scale annotated Indic manuscript image datasets available to researchers in the community, which would be incredibly helpful for analyzing the layout of these manuscripts.As a result, we took a substantial step in addressing these gaps by creating a diverse, annotated spatial layout dataset.

In order to create such a dataset, multiple factors need to be considered in data sourcing, identifying and prioritising the significant regions, categorising the classes, and determining the region types. After assessing this, a comprehensive set of annotation guidelines must be established to avoid ambiguity among annotators and preserve consistency throughout the data set. The following sections dwell into the details of the annotation challenges and guidelines.

2.3 Annotation Challenges

In the context of annotating Indic manuscript layouts, there are many unique challenges that they pose when compared to regular documents. The constraints stem from three primary sources.
Content: The manuscripts are written in a series of diverse Indian languages. Some languages even have script variations within themselves. To ensure efficient annotation of lines and character components, a large pool of annotators versed with the languages and scripts contained in the corpus is required.
Layout: Indic manuscripts, unlike other datasets, include non-textual components such as color drawings, tables, and document decorations in non-standard layouts, frequently interspersed with text. Many manuscripts have one or more physical holes designed to allow a thread-like material to pass through and bind the leaves together to form a book. The spatial location, number, and diameter of such holes vary. At times, a "virtual" hole-like gap also occurs, possibly for punching a hole at a later time. When holes appear in the centre of a document, they cause a break in the line continuity. In many cases, the spacing between lines in manuscripts is often exceedingly small, making them dense. Because of the handwritten character of these texts and the uneven surface material, very precise and slow annotation is required, making the annotator's job challenging and time-consuming. If multiple manuscript images are present, the stacking order could be horizontal or vertical. Overall, the sheer variety in layout elements poses a significant challenge, not only for annotation, but also for automated layout parsing.
Degradations: Historical Indic manuscripts are fragile and prone to deterioration from a variety of sources, including wood-and-leaf-boring insects, humidity seepage, inappropriate storage and handling, and so on. While certain degradations cause the document's edges to fray, others result in oddly shaped perforations in the document's core. Before undertaking lexically-focused tasks such as OCR or word-spotting, it may be necessary to identify such degradations.

2.4 Annotation Guidelines

As the annotations are done by multiple annotators, there is a considerable risk of making mistakes and being inconsistent if there are no regulations. Therefore, a well-thought-out and published set of guidelines has been established to ensure uniformity and correctness throughout the dataset. Initially, depending on the data to be handled and the content to be accessed, a list of class types is defined, and images are chosen accordingly. Only images with the predefined class type are chosen randomly and annotated. Considering the whole dataset, we have identified 10 class types that constitute the majority of the documents, and they are as follows.

1. Holes(physical)

2. Holes(Virtual)

3. Boundary Line

4. Page Boundary

5. Picture

6. Decorator

7. Character Line segment

8. Character Component

9. Library marker

10. Physical Degradation

The HInDoLA Annotation tool is used to annotate all of these region types. As mentioned in previous sections, the documents have holes in them for a thread to run through, and these are labelled as **Holes (physical)**. The space is left for them to be punched in a few exceptional cases, but they remain unpunched eternally. These are known as **Holes (virtual)**. However, they are in relatively limited quantities. **Character line segments** are textual lines that run from one end of the page to the other, whereas **Character components** are short words that can occur anywhere on the page. The two region types that make up the document's main textual content are character line segments and character components. There are also some beautiful illustrations depicting the event or context, referred to as **Pictures**. A few documents contain floral art added by the authors, generally at the beginning or conclusion of the book, to make the collection look prettier. These are labelled as **Decorators**. **Boundary lines** are the ones that are drawn on the left and right sides of the page to maintain alignment. These can also be found at the top and bottom of the page. A **Page boundary** is a region where boundary around the document is drawn to ensure the clear separation between several pages in a single image or some foreign textual components. Refer to figure 2.3 for more details.

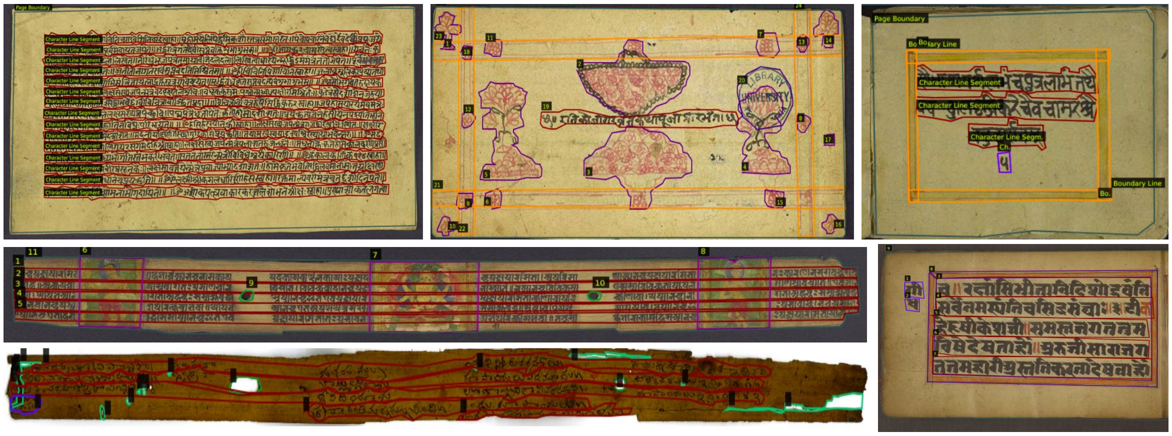


Figure 2.3: Sample class regions annotated on the manuscripts using HInDoLA

Chapter 3

Overview of the Datasets

3.1 Indiscapes: The Indic Manuscript Dataset

Indiscapes is the first-ever historical Indic manuscript dataset with detailed spatial layout annotations. It is composed of manuscript images acquired from two primary sources. The first source is the University of Pennsylvania's Rare Book and Manuscript Library's publicly accessible Indic manuscript collection, popularly known as **Penn-in-Hand** [1]. We picked 193 manuscript images for annotation from the 2,880 Indic manuscript book sets[1]. Our curated selection aims to maximise the dataset's diversity in terms of various parameters such as the extent of document degradation, script language, presence of non-textual elements (e.g. pictures, tables) and the number of lines. Some images contain multiple manuscript pages stacked vertically or horizontally (see the bottom-left image in Figure 3.1).

Bhoomi, a scattered collection of 315 images gathered from several Oriental Research Institutes and libraries across India, is our dataset's second source for manuscript images. We chose a subset from the whole dataset to optimise the dataset's overall diversity, just as we did with the previous collection. However, the latter collection of images is distinguished by a lower document quality, the inclusion of numerous languages, and the presence of long, densely and irregularly spaced text lines, binding holes, and degradations (Figure 3.1). We do not attempt to split the image into multiple pages, even if it contains multiple documents. While this makes automatic image processing and annotation difficult, keeping such images in the dataset avoids the need for manual or semi-automatic intervention.

In aggregate, we have 508 annotated Indic manuscripts in our collection. The table 3.1 depicts some critical aspects of the dataset, while the figure 3.1 depicts a pictorial representation of the layout regions. Unlike existing historical document datasets, which typically contain disjoint region annotations, multiple regions might overlap in our case.

[1] A book-set is a sequence of manuscript images.

	Character Line Segment (CLS)	Character Component (CC)	Hole (H)	Page Boundary (PB)	Library Marker (LM)	Decorator (D)	Picture (P)	Physical Degradation (PD)	Boundary Line (BL)
PIH	2401	494	–	256	32	59	94	34	395
BHOOMI	2440	210	565	316	133	–	–	2078	–
Combined	4841	704	565	572	165	59	94	2112	395

Table 3.1: Counts for various annotated region types in INDISCAPES dataset. The abbreviations used for region types are given below each region type.

	Train	Validation	Test	Total
PIH	116	28	49	193
BHOOMI	236	59	20	315
Total	352	87	69	508

Table 3.2: Dataset splits used for learning and inference.

Script	Source	Document Count
Devanagari	PIH	193
Nandinagari	BHOOMI	2
Telugu	BHOOMI	75
Grantha	BHOOMI	238

Table 3.3: Scripts in the INDISCAPES dataset.

3.2 Motivation for Indiscapes2

Among the varieties of historical manuscripts, many from the Indian subcontinent and South-east Asia are written on palm leaves. These manuscripts pose significant and unique challenges to the problem of layout prediction. The digital versions often reflect multiple degradations of the original. Also, a large variation exists in terms of the script language, aspect ratios and density of text and non-text region categories. The Indiscapes, Indic manuscript dataset and the deep-learning-based layout parsing model by Prusty et al. [65] represent a significant first step towards addressing the concerns mentioned above in a scalable manner. Although Indiscapes is the largest available annotated dataset of its kind, it contains a relatively small set of documents sourced from two collections. The deficiency is also reflected in the layout prediction quality of the associated deep learning model.

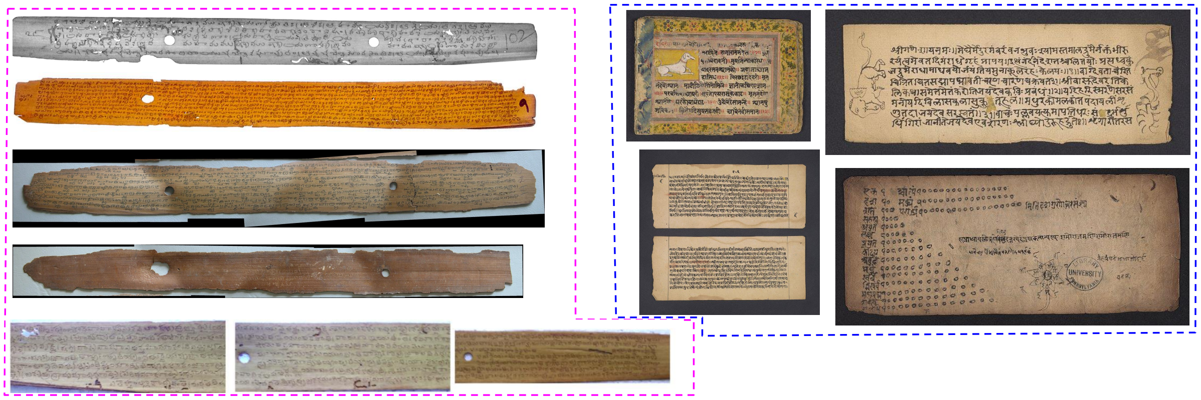


Figure 3.1: The five images on the left, enclosed by pink dotted line, are from the BHOOMI palm leaf manuscript collection while the remaining images (enclosed by blue dotted line) are from the 'Penn-in-Hand' collection (refer to Section 3.2.1). Note the inter-collection differences, closely spaced and unevenly written text lines, presence of various non-textual layout regions (pictures, holes, library stamps), physical degradation and presence of multiple manuscripts per image. All of these factors pose great challenges for annotation and machine-based parsing.

To address these shortcomings, we introduce the Indiscapes2 dataset as an expanded version of Indiscapes (Sec. 3.1). Indiscapes2 is **150**% larger compared to its predecessor and contains two additional annotated collections which greatly increased qualitative diversity (see Fig. 3.2, Table 3.4).

3.2.1 Indiscapes2

Despite Indiscapes being the first large-scale Indic manuscript dataset, it is relatively small by dataset standards. Indiscapes2 was developed on Indiscapes to alleviate this issue and enable advanced layout segmentation of deep networks. We used HInDoLA [88], a multi-feature annotation and analytics tool for historical manuscript layout processing for annotation. The fully automatic layout segmentation approach from Prusty et al. [65] is available as an annotation feature in HInDoLA. The annotators use the same to get an initial estimate and alter the output, reducing the time and effort required compared to pure manual annotation. HInDoLA also has a visualisation interface for evaluating annotation accuracy and labelling documents for correction.

In the new dataset, we introduce additional annotated documents from the Penn-in-Hand and Bhoomi book collections mentioned previously. Above this, we also add annotated manuscripts from two new collections - ASR and Jain. The ASR documents are from a private collection and contain 61 manuscripts written in Telugu language. They contain $18-20$ densely spaced text lines per document (see Fig. 3.2). The Jain collection contains 189 images. These documents contain $16-17$ lines per page and include early paper-based documents in addition to palm-leaf manuscripts.

	Train	Validation	Test	Total	Indiscapes (old)
PIH	285	70	94	**449**	193
BHOOMI	408	72	96	**576**	315
ASR	36	11	14	**61**	−−
JAIN	95	40	54	**189**	−−
Total	824	193	258	**1275**	508

Table 3.4: Collection level stats.

	Character	Character	Hole	Hole	Page	Library	Decorator/	Physical	Boundary
	Line Segment	Component	(Virtual)	(Physical)	Boundary	Marker	Picture	Degradation	Line
	(CLS)	(CC)	(Hv)	(Hp)	(PB)	(LM)	(D/P)	(PD)	(BL)
PIH	5105	1079	–	9	610	52	153	90	724
BHOOMI	5359	524	8	737	547	254	8	2535	80
ASR	673	59	–	–	52	41	–	81	83
JAIN	1857	313	93	38	166	7	–	166	292
Combined	12994	1975	101	784	1375	354	161	2872	1179

Table 3.5: Region count statistics.

Indiscapes2 has 1275 documents, representing a 150 per cent increase over the previous Indiscapes dataset. Additional statistics about the datasets can be found in Tables 3.4,3.5, and representative images can be seen in 3.2. Overall, Indiscapes2 provides more qualitative and quantitative coverage throughout the spectrum of historical documents.

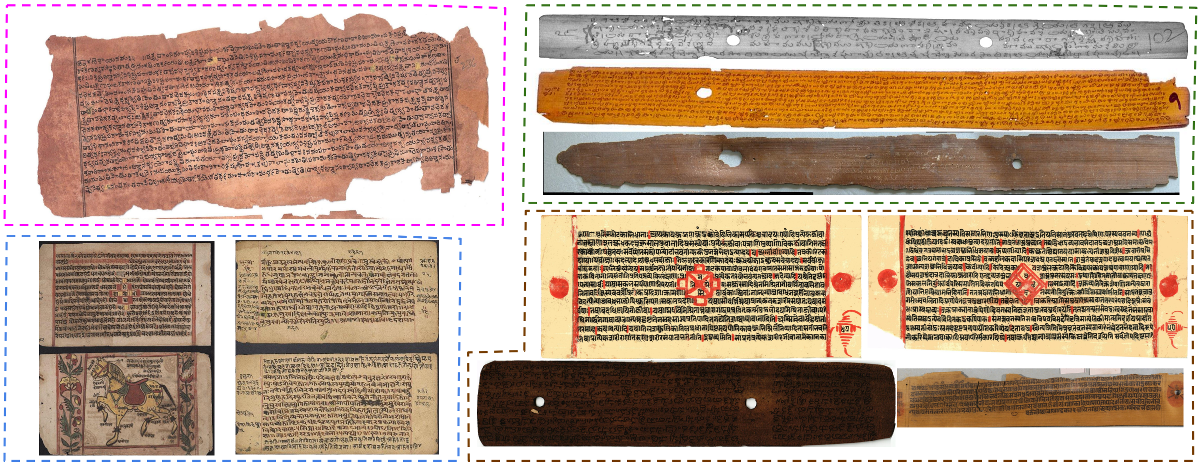


Figure 3.2: Representative manuscript images from Indiscapes2 - from newly added ASR collection (top left, pink dotted line), Penn-in-Hand (bottom left, blue dotted line), Bhoomi (green dotted line), newly added Jain (brown dotted line). Note the diversity across collections in terms of document quality, region density, aspect ratio and non-textual elements (pictures).

Chapter 4

IMMI Dataset

4.1 Motivation for Collection of IMMI Dataset

Handwritten historical manuscripts are a valuable source of tradition, history, and knowledge for preserving cultural heritage. This is especially true for manuscripts from the Indian subcontinent and South-East Asian countries [43]. Sourcing, preserving, restoring and digitizing the manuscripts are few significant steps in gaining access to content from the documents. In particular, accessing semantic content from digitized (scanned) manuscript images itself involves multiple steps and unique challenges. These aspects have been explored by several researchers in the community. The steps and challenges mentioned are usually in the form of structural layout parsing (semantic instance segmentation) [65,78] and optical character recognition (OCR) [21,57,84,93].

The first step towards accessing semantic content from historical palm-leaf manuscripts is scanning, i.e. obtaining a photo-like record of the physical manuscript item. A large diversity exists within this process due to the variety in capture mechanisms (flatbed scanners, handheld cameras). Often, in order to maximally utilize the area of scanning and minimize the manual labor involved, multiple manuscripts are scanned together. Therefore, the first crucial task is to ensure that each of the individual manuscripts within a scanned image can be isolated (segmented) on a per-instance basis.

This process of segmentation needs to tackle the diversity in appearance, capture quality, physical dimensions of the manuscripts. One particularly challenging aspect is overlap. Due to the physical attributes of the manuscripts (elongated and curved) or due to oversight by the person conducting the scanning, individual manuscripts can overlap with each other. This poses a challenge for image segmentation approaches which assume non-overlapping instances.

To address this gap, we first introduce a new document image dataset called IMMI (Indic Multi Manuscript Images). Each image in IMMI contains pixel level annotation of individual manuscript boundaries (Fig 4.3). Images from our dataset can be used to train deep networks for multi-page document image segmentation. The IMMI dataset covers a practical range in terms of number of manuscripts per image. However, the frequency distribution is imbalanced since there are fewer documents with a large number of manuscripts per image.

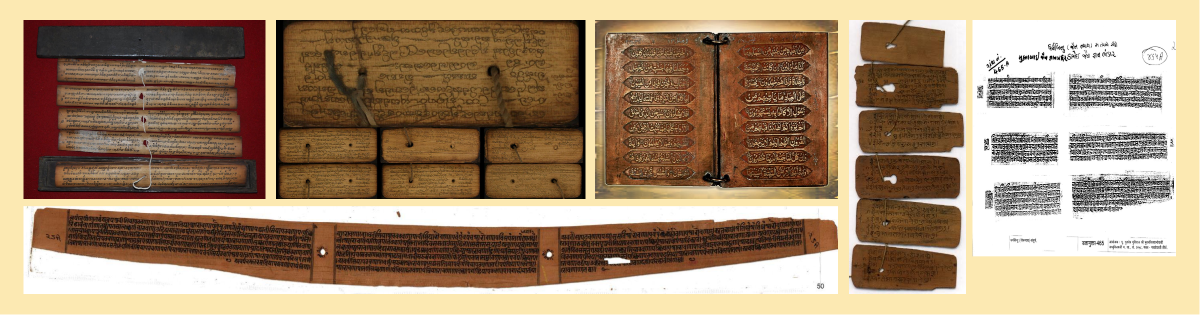


Figure 4.1: Examples of non-synthetic (real) document images. The challenging aspects such as variation in background, aspect ratio, layout, size of the image, number of manuscripts per document is evident.

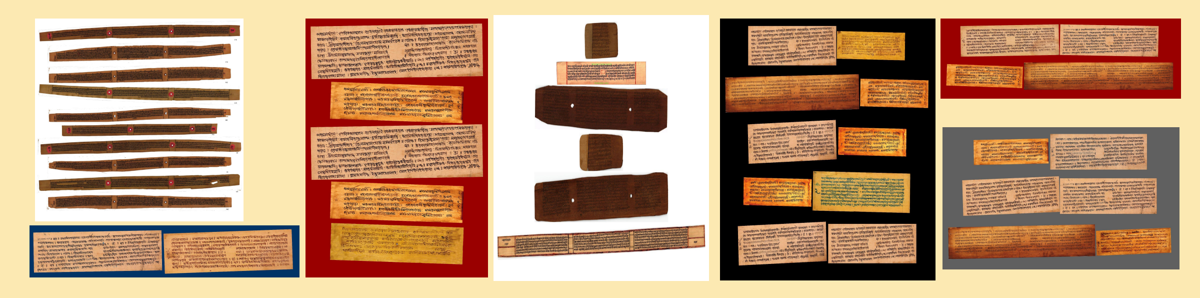


Figure 4.2: Examples of synthetic document images generated by our procedure (Section 4.2.2). Please compare with real (non-synthetic) document images in Figure 4.1.

Synthetic data generation is a popular option to compensate for lack of sufficient data when training deep learning approaches. Garai et al. [31] introduce a framework for creating synthetically warped images from flatbed scanned document images. Kieu et al. [44] employ a 3D approach to replicate the geometric distortions found in real documents as part of the new document synthesis process. Karpinski and Belaid [38] introduce a fully automatic method to generate synthetic historical document images by extracting and mixing Text-only images with Background-only images. We utilize a copy-paste augmentation strategy involving random mixing of document regions (manuscript pages) from existing data. Incidentally, variants of copy-paste strategy have been shown to be an effective augmentation strategy for instance segmentation [34].

The formulation and implementation of the synthetic data generation procedure using copy-paste augmentation strategy is discussed in section (Section 4.2.2). The resulting synthetic data serves to supplement our dataset and provide a larger quantity of data for training data-hungry deep networks.

4.2 IMMI Dataset

4.2.1 Composition of the Dataset

Non-Synthetic Images: To source images, we scraped those available on the Internet. We also sourced images from private collections available to us. Another major source of images was the recently introduced Indiscapes2 dataset [78]. We selected images to maximize the diversity in terms of manuscript count per document. Sample non-synthetic document images can be viewed in Figure 4.1. As the first step, we use HInDoLA [88], a web-based multi-feature annotation framework, for annotating the page boundaries of manuscripts in document images.

Synthetic Images: The statistics of manuscript counts in non-synthetic images can be viewed in Table 4.1. Clearly, the frequency distribution of images in terms of manuscript count per document is highly imbalanced. To address this imbalance, we also create synthetic images (described in the next section). Sample images can be viewed in Figure 4.2.

From Table 4.1, note that the data is split into training, validation and test in the ratio of 65 : 15 : 20 respectively. The data is split such that training and validation are done on both synthetic and non-synthetic data. However, the overall performance is reported on non-synthetic data. The dataset is also balanced throughout the splits (train, validation, test) by ensuring all of them have images spanning the manuscript count range.

4.2.2 Synthetic Image Creation

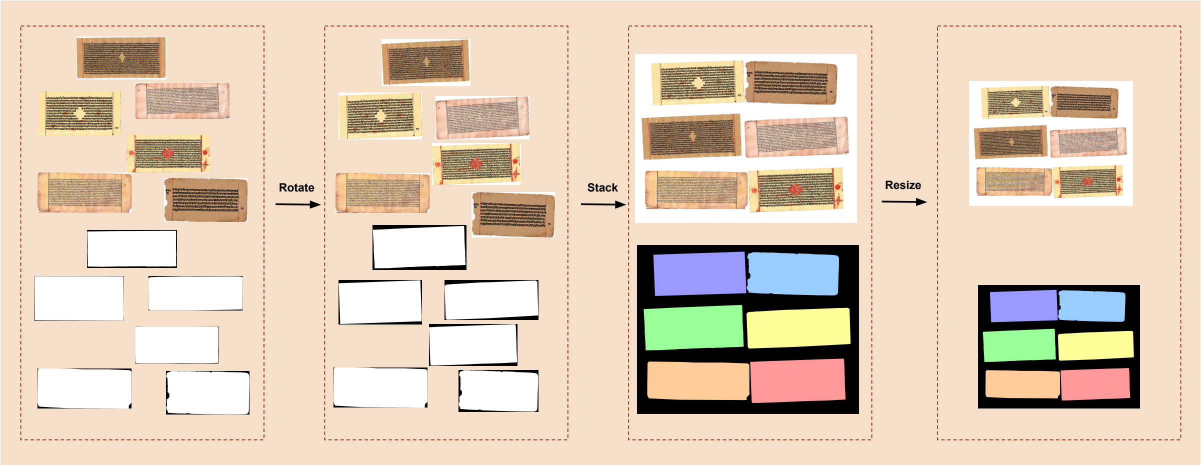


Figure 4.3: Step-by-step sequence for generation of synthetic images - refer to Sec. 4.2.2 for details.

1. The first step is to determine the number of manuscripts to be stacked. To ensure consistent background for synthetic images, non-synthetic manuscripts with same background colour are selected for stacking. The background colour is customized exclusively for manuscripts with no

	Non-synthetic				Synthetic		
# manuscripts per image	Train	Validation	Test	**Total**	**Total**	Train	Validation
1	45	14	11	**70**	**0**	0	0
2	29	6	12	**47**	**14**	11	3
3	22	8	12	**42**	**20**	19	1
4	20	4	20	**44**	**47**	38	9
5	17	0	13	**30**	**29**	20	9
6	4	1	12	**17**	**46**	38	8
7	2	0	13	**15**	**45**	37	8
8	0	0	3	**3**	**64**	54	10
9	0	0	9	**9**	**52**	43	9
10	0	0	1	**1**	**67**	58	9
18	0	0	3	**3**	**0**	0	0
Total	139	33	109	**281**	**384**	318	66

Table 4.1: Distribution statistics of real (non-synthetic) and synthetic data in IMMI dataset.

background but for images with pre-existing background colours (e.g. white), the same colour is used. The width and height of the synthetic image canvas are also controlled by the chosen non-synthetic manuscripts. In case of vertical stacking, the height of the canvas is decided by the sum of total image heights in the subset whereas the widest manuscript sub-image determines the width. In horizontal stacking, height of the canvas is decided by the sub-image with the maximum height, and the width of the canvas is equal to the sum of the widths of all the images being stacked. An extra 100 pixels are added to all sides of the canvas for padding. On account of each manuscript, 55 pixels are added to the canvas's height (vertical stack) or width (horizontal stack) to accommodate for the space between the scripts.

2. Once the canvas is ready, the selected manuscripts are positioned vertically, horizontally or a combination of both. The non-synthetic images with the largest widths are stacked only vertically to ensure a practical range for synthetic image width. The ground truth for these manuscripts is also stacked in parallel, with the appropriate offset values to match the location.

3. To mimic unstructured layouts, non-synthetic images are randomly rotated between -2 and 2 degrees and stored separately. The extra area created during rotation operation is filled with

same colour as the background. The rotated ground truth label map is generated by multiplying the original image's ground truth coordinates with the rotation matrix corresponding to the randomly chosen angle. Manuscripts are subsequently chosen from these rotated versions for stacking within the synthetic image canvas.

4. The synthetic images produced are unusually large in terms of width and height since the images are stacked in their original resolution. Hence they are scaled using a set scale factor s defined as $s = min(\frac{W}{w}, \frac{H}{h})$ where W, H are the maximum height and width seen for non-synthetic images and h, w are corresponding dimensions for a given synthetic image. To determine the new dimensions, the scale factor is multiplied to the original dimensions of the synthetic image while retaining the aspect ratio. The original ground truth label map coordinates are also scaled to match the resized synthetic image.

Our synthetic data generation procedure ensures that the resulting images resemble non-synthetic counterparts. Apart from diversity in appearance, we also ensure a more balanced manuscript count per image across the collection. This can also be seen from the statistics in Table 4.1. Refer to Figures 4.1 and 4.2 for illustrative examples of synthetic and non-synthetic images and their annotations.

Chapter 5

Deep Networks for Manuscript Region Instance Segmentation

A number of research groups have invested significant efforts in the creation and maintenance of annotated, publicly available historical manuscript image datasets [39, 62, 68, 71, 75, 82, 83]. Other collections contain character-level and word-level spatial annotations for South-East Asian palm-leaf manuscripts [40, 86, 89]. In these latter set of works, annotations for lines are obtained by considering the polygonal region formed by union of character bounding boxes as a line. While studies on Indic palm-leaf and paper-based manuscripts exist, these are typically conducted on small and often, private collections of documents [5, 19, 61, 76, 77, 80, 85]. No publicly available large-scale, annotated dataset of historical Indic manuscripts exists to the best of our knowledge. In contrast with existing collections, our proposed dataset contains a much larger diversity in terms of document type (palm-leaf and early paper), scripts and annotated layout elements (see Tables 3.1,3.3). An additional level of complexity arises from the presence of multiple manuscript pages within a single image (see Fig. 3.1). A number of contributions can also be found for the task of historical document layout parsing [16, 17, 94, 95]. Wei et al. [94] explore the effect of using a hybrid feature selection method while using autoencoders for semantic segmentation in five historical English and Medieval European manuscript datasets. Chen et al. [17] explore the use of Fully Convolutional Networks (FCN) for the same datasets. Barakat et al. [12] propose a FCN for segmenting closely spaced, arbitrarily oriented text lines from an Arabic manuscript dataset. The mentioned approaches, coupled with efforts to conduct competitions on various aspects of historical document layout analysis have aided progress in this area [3, 4, 42]. A variety of layout parsing approaches, including those employing the modern paradigm of deep learning, have been proposed for Indic [73, 76, 80, 85] and South-East Asian [16, 43, 63, 86, 91] palm-leaf and paper manuscript images. However, existing approaches typically employ brittle hand-crafted features or demonstrate performance on datasets which are limited in terms of layout diversity. Similar to many recent works, we employ Fully Convolutional Networks in our approach. However, a crucial distinction lies in our formulation of layout parsing as an *instance* segmentation problem, rather than just a *semantic* segmentation problem. This avoids the problem of closely spaced layout regions (e.g. lines) being perceived as contiguous blobs.

Layout analysis is an actively studied problem in the document image analysis community [14, 23, 48]. For an overview of approaches employed for historical and modern document layout analysis, refer to the work of Prusty et al. [65] and Liang et al. [50]. In recent times, large-scale datasets such as PubLayNet [98] and DocBank [49] have been introduced for document image layout analysis. These datasets focus on layout segmentation of modern language printed magazines and scientific documents.

Among recent approaches for historical documents, Ma et al. [54] introduce a unified deep learning approach for layout parsing and recognition of Chinese characters in a historical document collection. Alaasam et al. [8] use a Siamese Network to segment challenging historical Arabic manuscripts into main text, side text and background. Alberti et al. [11] use a multi-stage hybrid approach for segmenting text lines in medieval manuscripts. Monnier et al. [58] introduce docExtractor, an off-the-shelf pipeline for historical document element extraction from 9 different kinds of documents utilizing a modified U-Net [72]. dhSegment [60] is a similar work utilizing a modified U-Net for document segmentation of medieval era documents. Unlike our instance segmentation formulation (i.e. a pixel can simultaneously have two distinct region labels), existing works (except dhSegment) adopt the classical segmentation formulation (i.e. each pixel has a single region label). Also, our end-to-end approach produces page and region boundaries in a single stage end-to-end manner without any post-processing.

Approaches for palm-leaf manuscript analysis have been mostly confined to South-East Asian scripts [64, 92] and tend to focus on the problem of segmented character recognition [41, 56, 66]. The first large-scale dataset for palm leaf manuscripts was introduced by Prusty et al. [65], which we build upon to create an even larger and more diverse dataset.

Among deep-learning based works in document understanding, using deformable convolutions [26] to enable better processing of distorted layouts is a popular choice. However, existing works have focused only on tabular regions [6, 81]. We adopt deformable convolutions, but for the more general problem of multi-category region segmentation.

5.1 Mask R-CNN

5.1.1 Network Architecture

The Mask R-CNN architecture contains three stages as described below (see Figure 5.1).

Backbone: The first stage, referred to as the backbone, is used to extract features from the input image. It consists of a convolutional network combined with a feature-pyramid network [51], thereby enabling multi-scale features to be extracted. We use the first four blocks of ResNet-50 [37] as the convolutional network.

Region Proposal Network (RPN): This is a convolutional network which scans the pyramid feature map generated by the backbone network and generates rectangular regions commonly called 'object proposals' which are likely to contain objects of interest. For each level of the feature pyramid and for each spatial location at a given level, a set of level-specific bounding boxes called anchors are generated.

The anchors typically span a range of aspect ratios (e.g. $1:2, 1:1, 2:1$) for flexibility in detection. For each anchor, the RPN network predicts (i) the probability of an object being present ('objectness score') (ii) offset coordinates of a bounding box relative to location of the anchor. The generated bounding boxes are first filtered according to the 'objectness score'. From boxes which survive the filtering, those that overlap with the underlying object above a certain threshold are chosen. After applying non-maximal suppression (NMS) to remove overlapping boxes with relatively smaller objectness scores, the final set of boxes which remain are termed 'object proposals' or Regions-of-Interest (RoI).

Multi-Task Branch Networks: The RoIs obtained from RPN are warped into fixed dimensions and overlaid on feature maps extracted from the backbone to obtain RoI-specific features. These features are fed to three parallel task sub-networks. The first sub-network maps these features to region labels (e.g. *Hole*,*Character-Line-Segment*) while the second sub-network maps the RoI features to bounding boxes. The third sub-network is fully convolutional and maps the features to the pixel mask of the underlying region. Note that the ability of the architecture to predict masks independently for each RoI plays a crucial role in obtaining instance segmentations. Another advantage is that it naturally addresses situations where annotations or predictions overlap.

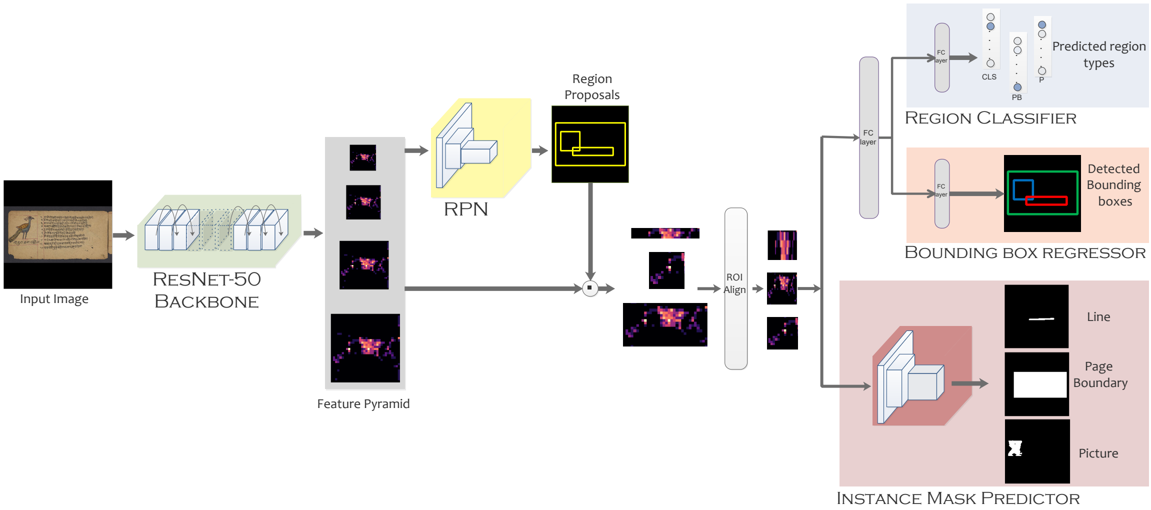


Figure 5.1: The architecture adopted for Indic Manuscript Layout Parsing. Refer to Section 5.1.1 for details.

5.1.2 Implementation Details

The dataset splits used for training, validation and test phases can be seen in Table 3.2. All manuscript images are adaptively resized to ensure the width does not exceed 1024 pixels. The images are padded with zeros such that the input to the deep network has spatial dimensions of 1024×1024. The ground truth region masks are initially subjected to a similar resizing procedure. Subsequently, they are downsized to 28×28 in order to match output dimensions of the mask sub-network.

5.1.2.1 Training

The network is initialized with weights obtained from a Mask R-CNN trained on the MS-COCO [52] dataset with a ResNet-50 backbone. We found that this results in faster convergence and stabler training compared to using weights from a Mask R-CNN trained on ImageNet [27] or training from scratch. Within the RPN network, we use custom-designed anchors of 5 different scales and with 3 different aspect ratios. Specifically, we use the following aspect ratios – 1:1,1:3,1:10 – keeping in mind the typical spatial extents of the various region classes. We also limit the number of RoIs ('object proposals') to 512. We use categorical cross entropy loss $\mathcal{L}_{RPN}$ for RPN classification network. Within the task branches, we use categorical cross entropy loss $\mathcal{L}_r$ for region classification branch, smooth L1 loss [69] ($\mathcal{L}_{bb}$) for final bounding box prediction and per-pixel binary cross entropy loss $\mathcal{L}_{mask}$ for mask prediction. The total loss is a convex combination of these losses, i.e. $\mathcal{L} = \lambda_{RPN}\mathcal{L}_{RPN} + \lambda_r\mathcal{L}_r + \lambda_{bb}\mathcal{L}_{bb} + \lambda_{mask}\mathcal{L}_{mask}$. The weighting factors (λs) are set to 1. However, to ensure priority for our task of interest namely mask prediction, we set $\lambda_{mask} = 2$. For optimization, we use Stochastic Gradient Descent (SGD) optimizer with a gradient norm clipping value of 0.5. The batch size, momentum and weight decay are set to 1, 0.9 and 10^{-3} respectively. Given the relatively smaller size of our manuscript dataset compared to the photo dataset (MS-COCO) used to originally train the base Mask R-CNN, we adopt a multi-stage training strategy. For the first stage (30 epochs), we train only the task branch sub-networks using a learning rate of 10^{-3} while freezing weights in the rest of the overall network. This ensures that the task branches are fine-tuned for the types of regions contained in manuscript images. For the second stage (20 epochs), we additionally train stage-4 and up of the backbone ResNet-50. This enables extraction of appropriate semantic features from manuscript images. The omission of the initial 3 stages in the backbone for training is due to the fact that they provide generic, re-usable low-level features. To ensure priority coverage of hard-to-localize regions, we use focal loss [53] for mask generation. For the final stage (15 epochs), we train the entire network using a learning rate of 10^{-4}.

5.1.2.2 Inference

During inference, the images are rescaled and processed using the procedure described at the beginning of the subsection5.1.2. The number of RoIs retained after non-maximal suppression (NMS) from the RPN is set to 1000. From these, we choose the top 100 region detections based on their classification branch score and feed only the corresponding RoIs to the mask branch sub-network for mask generation. It is important to note that this strategy is different from the parallel generation of outputs and use of the task sub-networks during training. The generated masks are then binarized using an empirically chosen threshold of 0.4 and rescaled to their original size using bilinear interpolation.

5.1.3 Evaluation

For quantitative evaluation, we compute Average Precision (AP) for a particular Intersection-over-Union (IoU) threshold, a measure widely reported in instance segmentation literature [25, 52]. We specifically report AP_{50} and AP_{75}, corresponding to AP at IoU thresholds 50 and 75 respectively [36]. In addition, we report an overall score by averaging AP at different IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05.

The AP measure characterizes performance at document level. To characterize performance for each region type, we report two additional measures [17] – average class-wise IoU (cwIoU) and average class-wise per-pixel accuracy (cwAcc). Consider a fixed test document k. Let n_{ij} be the number of pixels of class i predicted as class j. Let $t_i = \sum_j n_{ij}$ be the total number of pixels whose ground-truth label is i. The class-wise IoU for class i and document k is computed as $dIOU_i^k = \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$. Let N_i be the total number of instances of class i in our test set. We define the average IoU for class i as $cwIOU_i = \frac{1}{N_i} \sum_k dIOU_i^k$. We define class-wise pixel accuracy for class i and document k as $dAcc_i^k = \frac{n_{ii}}{t_i}$ and average class-wise per pixel accuracy as $cwAcc_i = \frac{1}{N_i} \sum_k dAcc_i^k$. For an input image, the mask branch of the network outputs many instance masks. Note that in order to compute cwIOU and cwAcc for a class i, only the instance mask predictions from the same class (i) are considered.

5.2 Our Layout Parsing Network (PALMIRA)

In their work, Prusty et al. [65] utilize a modified Mask R-CNN [36] framework for the problem of localizing document region instances. Although the introduced framework is reasonably effective, the fundamental convolution operation throughout the Mask R-CNN deep network pipeline operates on a fixed, rigid spatial grid. This rigidity of receptive fields tends to act as a bottleneck in obtaining precise boundary estimates of manuscript images containing highly deformed regions. To address this shortcoming, we modify two crucial stages of the Mask R-CNN pipeline in a novel fashion to obtain our proposed architecture (see Fig. 5.2). To begin with, we briefly summarize the Mask R-CNN approach adopted by Prusty et al. We shall refer to this model as the Vanilla Mask R-CNN model. Subsequently, we shall describe our novel modifications to the pipeline.

5.2.1 Vanilla Mask R-CNN

Mask R-CNN [36] is a three stage deep network for object instance segmentation. The three stages are often referred to as Backbone, Region Proposal Network (RPN) and Multi-task Branch Networks. One of the Branch Networks, referred to as the Mask Head, outputs individual object instances. The pipeline components of Mask R-CNN are modified to better suit the manuscript image domain by Prusty et al [65]. Specifically, the ResNet-50 used in Backbone is initialized from a Mask R-CNN trained on the MS-COCO dataset. Within the RPN module, the anchor aspect ratios of 1:1,1:3,1:10 were chosen keeping the peculiar aspect ratios of manuscript images in mind and the number of proposals from RPN

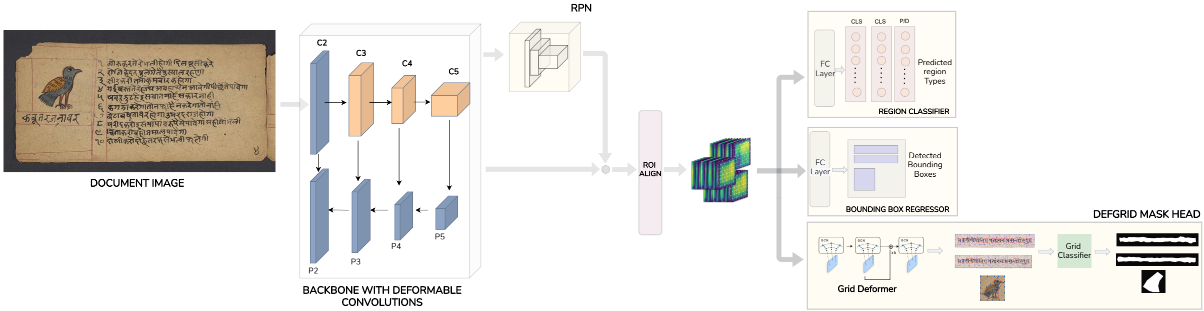


Figure 5.2: A diagram illustrating PALMIRA's architecture (Sec. 5.2). The orange blocks in the backbone are deformable convolutions (Sec. 5.2.2). Refer to Fig. 6.2b, Sec. 5.2.3 for additional details on Deformable Grid Mask Head which outputs region instance masks.

were reduced to 512. The various thresholds involved in other stages (objectness and NMS) were also modified suitably. Some unique modifications were included as well – the weightage for loss associated with the Mask head was set to twice of that for the other losses and focal-loss [53] was used for robust labelling.

We use the modified pipeline described above as the starting point and incorporate two novel modifications to Mask R-CNN. We describe these modifications in the sections that follow.

5.2.2 Modification-1: Deformable Convolutions in Backbone

Before examining the more general setting, let us temporarily consider 2D input feature maps $\boldsymbol{x}$. Denote the 2D filter operating on this feature map as $\boldsymbol{w}$ and the convolution grid operating on the feature map as $\mathcal{R}$. As an example, for a 3×3 filter, we have:

$$\mathcal{R} = \begin{Bmatrix} (-1,-1) & (-1,0) & (-1,1) \\ (0,-1) & (0,0) & (0,1) \\ (1,-1) & (1,0) & (1,1) \end{Bmatrix} \tag{5.1}$$

Let the output feature map resulting from the convolution be $\boldsymbol{y}$. For each pixel location $\boldsymbol{p}_0$, we have:

$$\boldsymbol{y}(\boldsymbol{p}_0) = \sum_{\boldsymbol{p}_n \in \mathcal{R}} \boldsymbol{w}(\boldsymbol{p}_n) \cdot \boldsymbol{x}(\boldsymbol{p}_0 + \boldsymbol{p}_n) \tag{5.2}$$

where n indexes the spatial grid locations associated with $\mathcal{R}$. The default convolution operation in Mask R-CNN operates via a fixed 2D spatial integer grid as described above. However, this setup does not enable the grid to deform based on the input feature map, reducing the ability to better model the high inter/intra-region deformations and the features they induce.

As an alternative, Deformable Convolutions [26] provide a way to determine suitable local 2D offsets for the default spatial sampling locations (see Fig. 6.2a). Importantly, these offsets $\{\Delta\boldsymbol{p}_n; n = 1, 2 \ldots\}$

are adaptively computed as a function of the input features for each reference location $\boldsymbol{p}_0$. Equation 5.2 becomes:

$$\boldsymbol{y}(\boldsymbol{p}_0) = \sum_{\boldsymbol{p}_n \in \mathcal{R}} \boldsymbol{w}(\boldsymbol{p}_n) \cdot \boldsymbol{x}(\boldsymbol{p}_0 + \boldsymbol{p}_n + \Delta \boldsymbol{p}_n) \tag{5.3}$$

Since the offsets $\Delta \boldsymbol{p}_n$ may be fractional, the sampled values for these locations are generated using bilinear interpolation. This also preserves the differentiability of the filters because the offset gradients are learnt via backpropagation through the bilinear transform. Due to the adaptive sampling of spatial locations, broader and generalized receptive fields are induced in the network. Note that the overall optimization involves jointly learning both the regular filter weights and weights for a small additional set of filters which operate on input to generate the offsets for input feature locations (Fig. 6.2a).

5.2.3 Modification-2: Deforming the Spatial Grid in Mask Head

The ‘Mask Head’ in Vanilla Mask R-CNN takes aligned feature maps for each plausible region instance as input and outputs a binary mask corresponding to the predicted document region. In this process, the output is obtained relative to a 28×28 regular spatial grid representing the entire document image. The output is upsampled to the original document image dimensions to obtain the final region mask. As with the convolution operation discussed in the previous section, performing upsampling relative to a uniform (integer) grid leads to poorly estimated spatial boundaries for document regions, especially for our challenging manuscript scenario.

5.2.4 Implementation Details

Architecture: The Backbone in PALMIRA consists of a ResNet-50 initialized from a Mask R-CNN network trained on the MS-COCO dataset. Deformable convolutions (Sec. 5.2.2) are introduced as a drop-in replacement for the deeper layers C3-C5 of the Feature Pyramid Network present in the Backbone (see Fig. 6.2a). Empirically, we found this choice to provide better results compared to using deformable layers throughout the Backbone. We use $0.5, 1, 2$ as aspect ratios with anchor sizes of $32, 64, 128, 256, 512$ within the Region Proposal Network. While the Region Classifier and Bounding Box heads are the same as one in Vanilla Mask R-CNN (Sec. 5.2), the conventional Mask Head is replaced with the Deformable Grid Mask Head as described in Sec. 5.2.3.

Optimization: All input images are resized such that the smallest side is 800 pixels. The mini-batch size is 4. During training, a horizontal flip augmentation is randomly performed for images in the mini-batch. To address the imbalance in the distribution of region categories (Table 3.4), we use repeat factor sampling [35] and oversample images containing tail categories. We perform data-parallel optimization distributed across 4 GeForce RTX 2080 Ti GPUs for a total of 15000 iterations. A multi-step learning scheduler with warmup phase is used to reach an initial learning rate of 0.02 after a linear warm-up over 1000 iterations. The learning rate is decayed by a factor of 10 at 8000 and 12000 iterations. The optimizer used is stochastic gradient descent with gamma 0.1 and momentum 0.9.

Except for the Deformable Grid Mask Head, other output heads (Classifier, Bounding Box) are optimized based on choices made for Vanilla Mask R-CNN [65]. The optimization within the Deformable Grid Mask Head involves multiple loss functions. Briefly, these loss functions are formulated to (i) minimize the variance of features per grid cell (ii) minimize distortion of input features during differentiable reconstruction (iii) avoid self-intersections by encouraging grid cells to have similar area (iv) encourage neighbor vertices in region localized by a reference central vertex to move in same spatial direction as the central vertex. Please refer to Gao et al. [29] for details.

5.3 Experimental Setup

5.3.1 Baselines

Towards fair evaluation, we consider three strong baseline approaches.

Boundary Preserving Mask R-CNN [18], proposed as an improvement over Mask R-CNN, focuses on improving the mask boundary along with the task of pixel wise segmentation. To this end, it contains a boundary mask head wherein the mask and boundary are mutually learned by employing feature fusion blocks.

CondInst [87] is a simple and effective instance segmentation framework which eliminates the need for resizing and RoI-based feature alignment operation present in Mask R-CNN. Also, the filters in CondInst Mask Head are dynamically produced and conditioned on the region instances which enables efficient inference.

In recent years, a number of instance segmentation methods have been proposed as an alternative to Mask R-CNN's proposal-based approach. As a representative example, we use PointRend [45] - a proposal-free approach. PointRend considers image segmentation as a rendering problem. Instead of predicting labels for each image pixel, PointRend identifies a subset of salient points and extracts features corresponding to these points. It maps these salient point features to the final segmentation label map.

5.3.2 Evaluation Setup

We partition Indiscapes2 dataset into training, validation and test sets (see Table 3.4) for training and evaluation of all models, including PALMIRA. Following standard protocols, we utilize the validation set to determine the best model hyperparameters. For the final evaluation, we merge training and validation set and re-train following the validation-based hyperparameters. A one-time evaluation of the model is performed on the test set.

Table 5.1: Document-level scores summarized at collection level for various performance measures.

Collection name	# of test images	HD ↓	HD_{95} ↓	Avg. HD ↓	IoU ↑	AP ↑	AP_{50} ↑	AP_{75} ↑
PIH	94	66.23	46.51	11.16	76.78	37.57	59.68	37.63
BHOOMI	96	220.38	175.52	46.75	69.83	30.40	50.53	29.03
ASR	14	629.30	562.19	169.03	67.80	51.02	73.09	64.27
JAIN	54	215.14	159.88	38.91	76.59	48.25	70.15	50.34
OVERALL	258	184.50	145.27	38.24	73.67	42.44	69.57	42.93

5.3.3 Evaluation Measures

Intersection-over-Union (IoU) and Average Precision (AP) are two commonly used evaluation measures for instance segmentation. IoU and AP are area-centric measures which depend on intersection area between ground-truth and predicted masks. To complement these metrics, we also compute boundary-centric measures. Specifically, we use Hausdorff distance (HD) [46] as a measure of boundary precision. For a given region, let us denote the ground-truth annotation polygon by a 2D point set $\mathcal{X}$. Let the prediction counterpart be $\mathcal{Y}$. The Hausdorff Distance between these point sets is given by:

$$\text{HD} = d_H(\mathcal{X}, \mathcal{Y}) = max\left\{\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} d(x,y), \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} d(x,y)\right\} \tag{5.4}$$

where $d(x, y)$ denotes the Euclidean distance between points $x \in \mathcal{X}$, $y \in \mathcal{Y}$. The Hausdorff Distance is sensitive to outliers. To mitigate this effect, the Average Hausdorff Distance is used which measures deviation in terms of a symmetric average across point-pair distances:

$$\text{Avg. HD} = d_{AH}(\mathcal{X}, \mathcal{Y}) = \left(\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} d(x,y) + \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} d(x,y)\right)/2 \tag{5.5}$$

Note that the two sets may contain unequal number of points ($|\mathcal{X}|, |\mathcal{Y}|$). Additionally, we also compute the 95^{th} percentile of Hausdorff Distance (HD_{95}) to suppress the effect of outlier distances.

For each region in the test set documents, we compute HD, HD_{95}, IoU, AP at IoU thresholds of 50 (AP_{50}) and 75 (AP_{75}). We also compute overall AP by averaging the AP values at various threshold values ranging from 0.5 to 0.95 in steps of 0.05. We evaluate performance at two levels - document-level and region-level. For a reference measure (e.g. HD), we average its values across all regions of a document. The resulting numbers are averaged across all the test documents to obtain document-level score. To obtain region-level scores, the measure values are averaged across all instances which share the same region label. We use document-level scores to compare the overall performance of models. To examine the performance of our model for various region categories, we use region-level scores.

Table 5.2: PALMIRA's overall and region-wise scores for various performance measures. The HD-based measures (smaller the better) are separated from the usual measures (IoU, AP etc.) by a separator line.

Metric	**Overall**	Character Line Segment (CLS)	Character Component (CC)	Hole (Virtual) (Hv)	Hole (Physical) (Hp)	Page Boundary (PB)	Library Marker (LM)	Decorator/ Picture (D/P)	Physical Degradation (PD)	Boundary Line (BL)
HD_{95}	**171.44**	34.03	347.94	70.79	88.33	52.01	289.81	593.99	851.02	111.97
AVG HD	**45.88**	8.43	103.98	18.80	16.82	13.19	73.23	135.46	255.86	17.95
IOU (%)	**72.21**	78.01	54.95	74.85	77.21	92.97	67.24	50.57	27.68	61.54
AP	**42.44**	58.64	28.76	45.57	56.13	90.08	27.75	32.20	03.09	39.72
AP_{50}	**69.57**	92.73	64.55	81.20	90.53	93.99	55.18	54.23	12.47	81.24
AP_{75}	**42.93**	92.74	64.55	81.20	90.52	93.99	55.18	54.24	12.47	81.24

Table 5.3: Document-level scores for various performance measures. The baseline models are above the upper separator line while ablative variants are below the line. PALMIRA's results are at the table bottom.

Model	Add-On	HD ↓	HD_{95} ↓	Avg. HD ↓	IoU ↑	AP ↑	AP_{50} ↑	AP_{75} ↑
PointRend [45]	-	252.16	211.10	56.51	69.63	41.51	66.49	43.49
CondInst [87]	-	267.73	215.33	54.92	69.49	42.39	62.18	43.03
Boundary Preserving MaskRCNN [18]	-	261.54	218.42	54.77	69.99	42.65	68.23	**44.92**
Vanilla MaskRCNN [65]	-	270.52	228.19	56.11	68.97	41.46	68.63	34.75
Vanilla MaskRCNN	Deformable Convolutions	229.50	202.37	51.04	65.61	41.65	65.97	44.90
Vanilla MaskRCNN	Deformable Grid Mask Head	**179.84**	153.77	45.09	71.65	42.35	69.49	43.16
PALMIRA : Vanilla MaskRCNN	Deformable Conv., Deformable Grid Mask Head	184.50	**145.27**	**38.24**	**73.67**	**42.44**	**69.57**	42.93

To understand the results at collection level, we summarize the document-level scores of PALMIRA in Table 5.1. While the results across collections are mostly consistent with overall average, the scores for ASR are suboptimal. This is due to the unsually closely spaced lines and the level of degradation encountered for these documents. It is easy to see that reporting scores in this manner is useful for identifying collections to focus on,for improvement in future.

5.4 Results

The performance scores for our approach (PALMIRA) and baseline models can be viewed in Table 5.3. Our approach clearly outperforms the baselines across the reported measures. Note that the improvement is especially apparent for the boundary-centric measures (HD, HD_{95}, Avg. HD). As an ablation study, we also evaluated variants of PALMIRA wherein the introduced modifications were re-

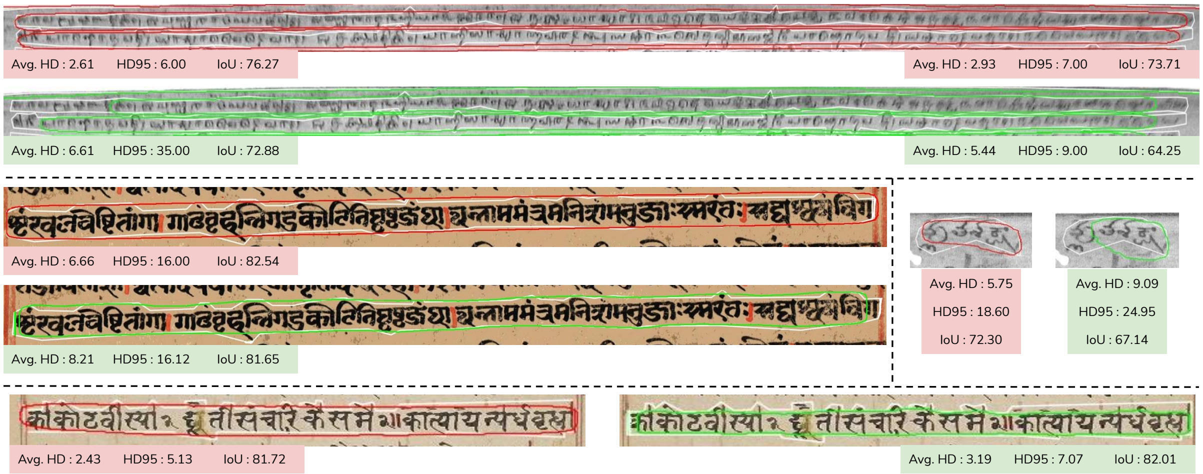

Avg. HD : 2.61 HD95 : 6.00 IoU : 76.27
Avg. HD : 2.93 HD95 : 7.00 IoU : 73.71
Avg. HD : 6.61 HD95 : 35.00 IoU : 72.88
Avg. HD : 5.44 HD95 : 9.00 IoU : 64.25
Avg. HD : 6.66 HD95 : 16.00 IoU : 82.54
Avg. HD : 8.21 HD95 : 16.12 IoU : 81.65
Avg. HD : 5.75 HD95 : 18.60 IoU : 72.30
Avg. HD : 9.09 HD95 : 24.95 IoU : 67.14
Avg. HD : 2.43 HD95 : 5.13 IoU : 81.72
Avg. HD : 3.19 HD95 : 7.07 IoU : 82.01

Figure 5.3: A comparative illustration of region-level performance. PALMIRA's predictions are in red. Predictions from the best model among baselines (Boundary-Preserving Mask-RCNN) are in green. Ground-truth boundary is depicted in white.

moved separately. The corresponding results in Table 5.3 demonstrate the collective importance of our novel modifications over the Vanilla Mask-RCNN model.

We also report the performance measures for PALMIRA, but now at a per-region level, in Table 5.2. In terms of the boundary-centric measures (HD_{95}, Avg. HD), the best performance is seen for the most important and dominant region category - Character Line Segment. The seemingly large scores for some categories ('Picture/Decorator', 'Physical Degradation') are due to the drastically small number of region instances for these categories. Note that the scores for other categories are reasonably good in terms of boundary-centric measures as well as the regular ones (IoU, AP).

A qualitative perspective on the results can be obtained from Figure 5.4. Despite the challenges in the dataset, the results show that PALMIRA outputs good quality region predictions across a variety of document types. A comparative illustration of region-level performance can be viewed in Figure 5.3. In general, it can be seen that PALMIRA's predictions are closer to ground-truth. Figure 5.5 shows PALMIRA's output for sample South-East Asian, Arabic and Hebrew historical manuscripts. It is important to note that the languages and aspect ratio (portrait) of these documents is starkly different from the typical landscape-like aspect ratio of manuscripts used for training our model. Clearly, the results demonstrate that PALMIRA readily generalizes to out of dataset manuscripts without requiring additional training.

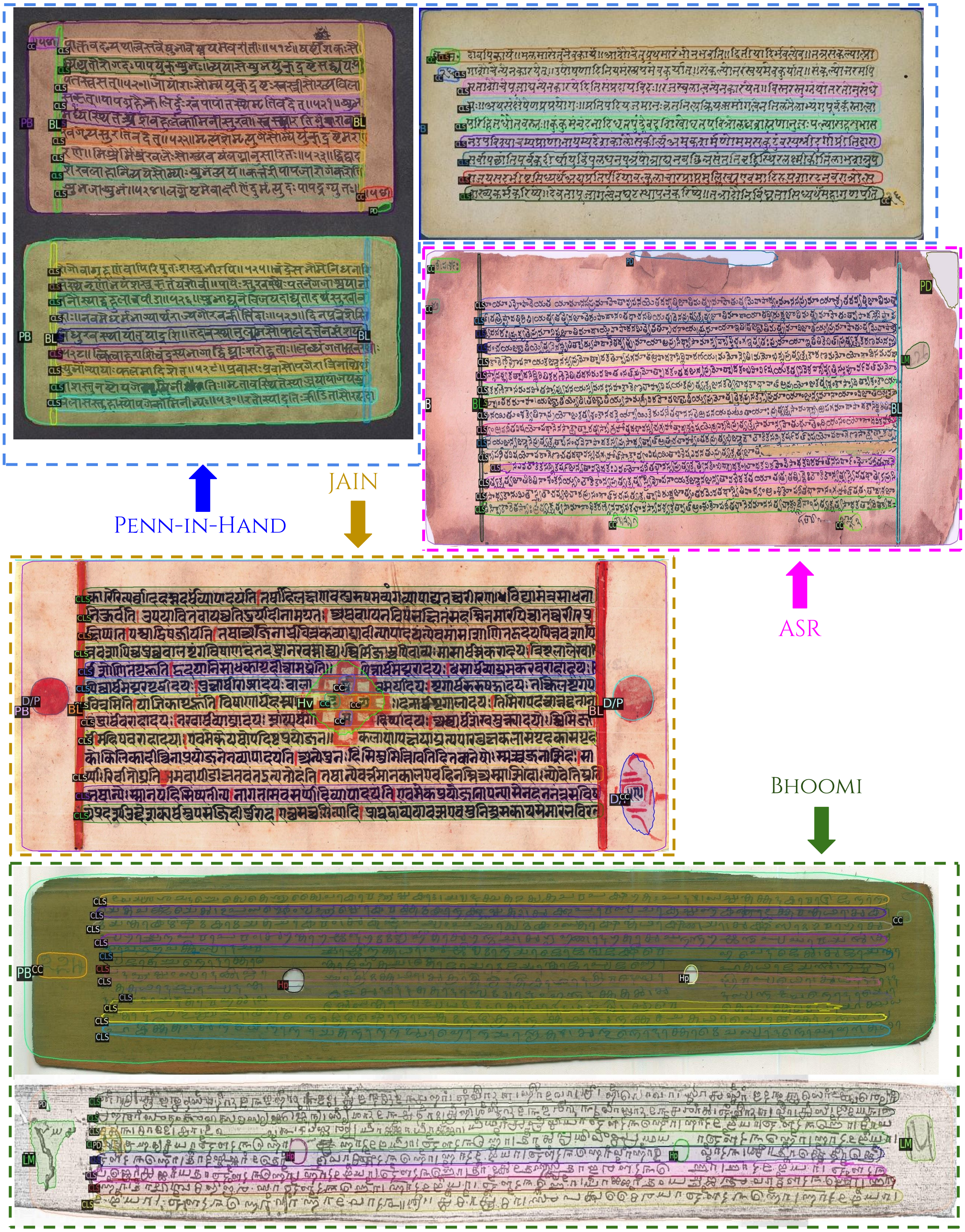


Figure 5.4: Layout predictions by PALMIRA on representative test set documents from Indiscapes2 dataset. Note that the colors are used to distinguish region instances. The region category abbreviations are present at corners of the regions.

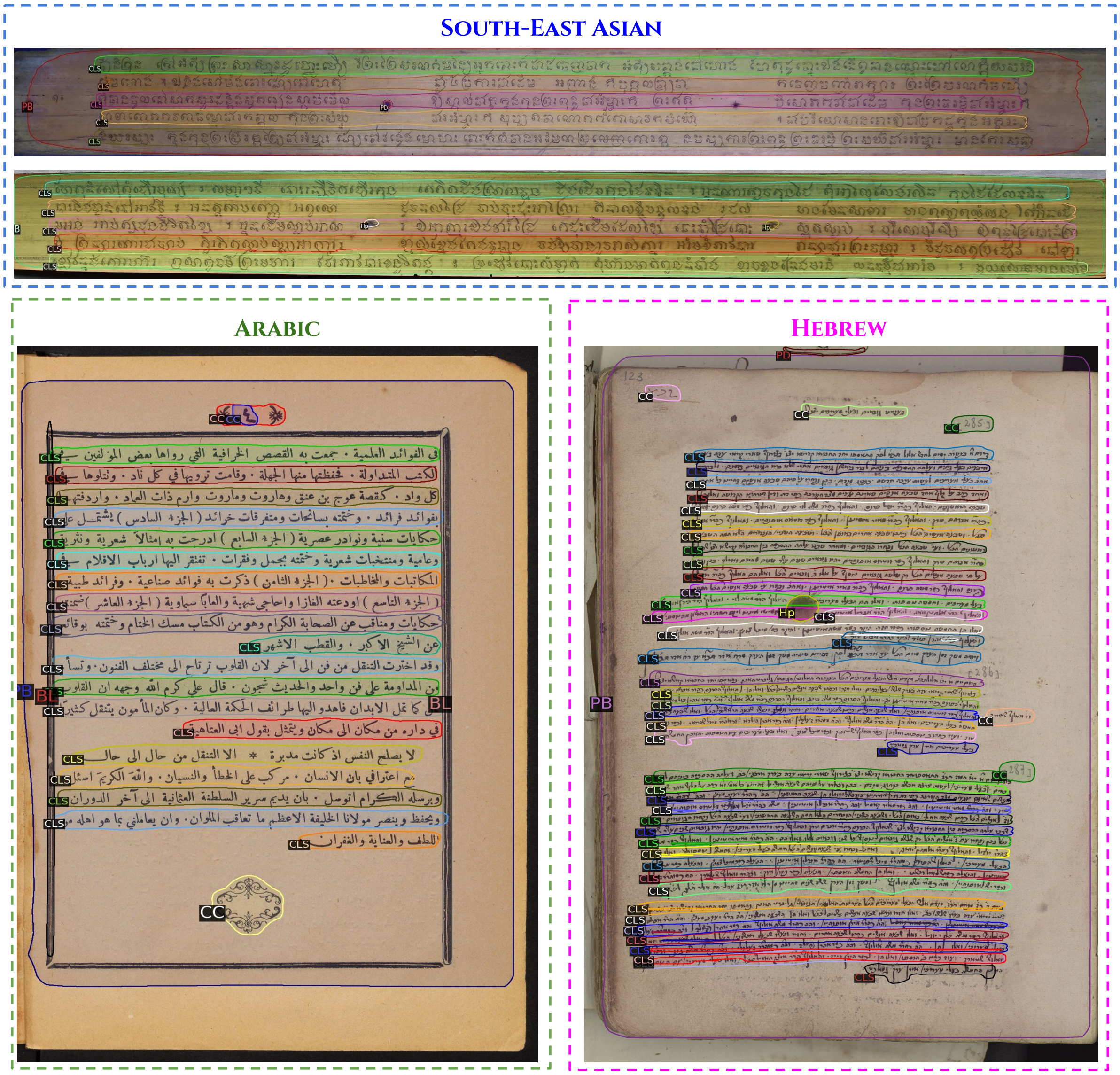


Figure 5.5: Layout predictions by PALMIRA on out-of-dataset handwritten manuscripts.

Chapter 6

Deep Networks for Multi-Manuscript Document Image Segmentation

As we realize the fact that digitizing via scanning the physical artifact often forms the first primary step in preserving historical handwritten manuscripts, multiple manuscripts are usually scanned together into a single image to maximally utilize scanner surface area and minimize manual labor. Hence, ensuring that each individual manuscript within a scanned image can be isolated (segmented) on a per-instance basis is the first crucial step towards understanding the content of a manuscript.

Existing deep network based approaches for manuscript layout understanding implicitly assume a single or two manuscripts per image. Since this assumption may be routinely violated, there is a need for a precursor system which extracts individual manuscripts before downstream processing. Therefore, We conduct experiments using modified versions of existing document instance segmentation frameworks. Our new networks can segment a document image containing multiple possibly overlapping manuscripts into individual instances. Our experiments and results (Section 6.3) demonstrate the benefit of employing our approach as the first step in processing digitized handwritten manuscript collections for the task.

The problem of document image layout analysis is a popular area [13, 20–22, 55, 78]. In particular, the problem of understanding the structural and semantic content of historical manuscripts is being pursued with works on topics such as layout parsing and text recognition [43, 65, 79, 82, 90, 97]. Semantic segmentation of documents into constituent regions is an active research area. Some works focus on layout parsing of entire image [13, 20–22, 55, 78]. Specifically for historical documents, layout analysis of challenging complex Arabic Handwritten Manuscripts utilising characteristics derived from connected components and Siamese network was done by Bukhari et al. [15] and Alaasam et al. [9]. Monnier et al. [59] also present an off the shelf historical document extractor for extracting visual elements.

Other works focus on segmenting a single category such as tables or images. Michele et al. [10] utilize deep-learning based pre-classification and segmentation algorithms to extract text lines from medieval manuscripts. Madhav et al. [7] introduce an end-to-end trainable network using deformable convolutions in their dual backbone to identify tables in documents. Renton et al. [70] propose a learning based method using fully convolutional network to detect handwritten text lines.

In existing approaches, the problem is cast as semantic segmentation. We cast our manuscript page segmentation as a semantic *instance* segmentation problem. This enables us to effectively tackle potential overlap between manuscript pages and isolate individual page instances.

In recent times, state of the art approaches have been proposed for segmenting manuscripts [65, 78]. However, these approaches focus on segmenting individual regions *within* manuscripts such as lines, pictures, binding holes. Although the approaches provide page boundary segmentation, they implicitly assume that the input image contain at most two well separated manuscripts. Therefore, in practice, such approaches do not perform well on the task of segmenting individual manuscripts.

Now, we first briefly summarize existing architectures for document region instance segmentation. As we shall demonstrate experimentally later, these approaches are too specialized for within document *region* instance segmentation and perform poorly for our task, i.e. *page* instance segmentation. Therefore, we propose modifications to the existing architectures in terms of task setup and architectural components which lead to good performance.

6.1 Vanilla Mask R-CNN

The work of Prusty et al. [65] was the first approach employing a custom modified Mask R-CNN [36] architecture to segment individual region instances (e.g. line segments, pictures) in a handwritten manuscript. Mask R-CNN [36] is a three-stage deep network developed to tackle the problem of image instance segmentation for photographic scenes. The three stages include Backbone, Region Proposal Network (RPN), and Multi-branch Networks, which in turn comprise of a Region of Interest (ROI) classifier, Bounding Box regressor and Mask predictor. Prusty et al. [65] modified these stages for better adaptation to regions within manuscript images.

6.2 Palmira

Prusty et al.'s [65] approach uses rigid axis-oriented receptive fields which are not suited to tackle deformations and alignment issues in manuscripts. More recently, Sharan et al. presented an improved version of the earlier approach via an architecture called Palmira [78] which handles region deformations better. For better understanding, we next describe two key architectural components introduced in Palmira.

6.2.1 Modification #1: Deformable Convolutions in the Backbone

Deformable convolutions [26] are a better replacement to traditional convolutions because of the latter's limitations in modeling geometric tranformations. This is due to fixed and rigid geometric structures in traditional convolution. The main idea behind using deformable convolutions is to introduce additional offsets to the regular grid spatial sampling locations in the standard convolution and learn

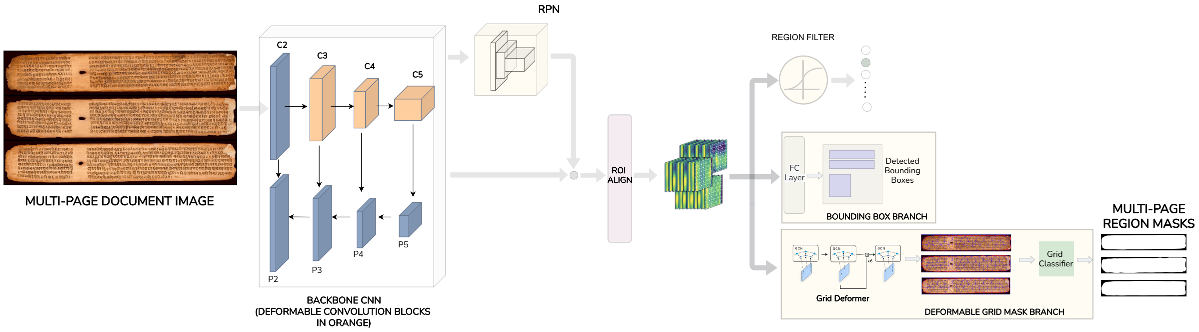


Figure 6.1: Architectural diagram of modified Palmira network for multi-page instance segmentation. Refer to Section 5.2.

these offsets from the target tasks without any extra supervision. Additional convolutional layers are used to learn the offsets from the previous feature maps (see Figure 6.2a). Suppose a 2D input feature map is denoted as *x* and a weight filter operating on the feature map be *w*. Let the convolution grid on the feature map be denoted *R*. Assuming the traditional convolution operates on the regular grid *R* with N locations, the deformable convolution operates on the same grid with additional offsets $\{\Delta p_n | n = 1..N\}$. This encourages capture of deformations in the images. For a sample grid *R* with 3 x 3 kernel, we have:

$$\mathcal{R} = \begin{Bmatrix} (-1,-1) & (-1,0) & (-1,1) \\ (0,-1) & (0,0) & (0,1) \\ (1,-1) & (1,0) & (1,1) \end{Bmatrix} \tag{6.1}$$

The output of the regular convolution is given as:

$$\boldsymbol{y}(\boldsymbol{p}_0) = \sum_{\boldsymbol{p}_n \in \mathcal{R}} \boldsymbol{w}(\boldsymbol{p}_n) \cdot \boldsymbol{x}(\boldsymbol{p}_0 + \boldsymbol{p}_n) \tag{6.2}$$

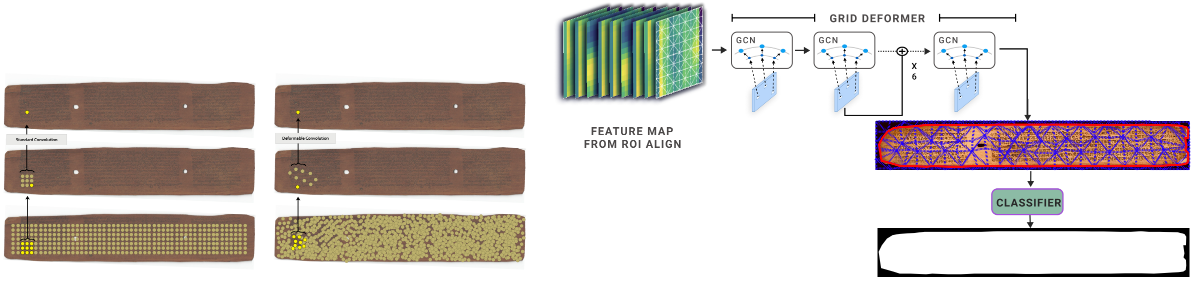
where y at *p0* corresponds to convolutional output at each point. The corresponding deformable convolution variant is given by:

$$\boldsymbol{y}(\boldsymbol{p}_0) = \sum_{\boldsymbol{p}_n \in \mathcal{R}} \boldsymbol{w}(\boldsymbol{p}_n) \cdot \boldsymbol{x}(\boldsymbol{p}_0 + \boldsymbol{p}_n + \Delta \boldsymbol{p}_n) \tag{6.3}$$

Typically, the offsets Δp_n are fractional. Therefore, bilinear interpolation is employed to get feature values at the desired locations.

6.2.2 Modification #2: Deforming the Spatial Grid in Mask Branch

Since it is crucial to capture the boundary deformation, i.e. at the output mask end, the deformable grid setup is used in the Mask prediction. In the baseline setup, the mask branch processes the input



(a) Deformable convolution showing learnt offsets and comparison with regular convolution.

(b) Deformable Spatial Grid Mask Branch

Figure 6.2: Modifications included as part of Palmira framework.

feature map corresponding to each instance obtained from the ROI-Align module and produces corresponding predictions in the form of a binary map. This resulting 28 × 28 region mask output is upsampled to the actual image size. Up sampling with respect to a uniform (integer) grid leads to poor predictions using regular convolution operations. Therefore, we adopt deformable grid processing for increased accuracy in predicting region boundaries.

The first step in deformable grid processing is to create a grid of 2D triangles whose vertices are present at regular spatial grid's 2D locations. A neural network module [30] then learns to predict location offsets of the triangle vertices such that the edges and vertices of the deformed grid align with image boundaries. Please view Figure 6.2b and refer to the work of Dai et al. [30] for details.

6.2.3 Our Architectural and Optimization Modifications

In our setting, we consider the task as segmenting individual manuscript instances. For this purpose, we propose three architectural configurations.

Model	AP↑	HD↓	HD95↓	Avg HD↓	IOU↑	Accuracy↑
FT-Palmira-AS	**83.30**	**163.73**	**87.84**	**18.38**	**91.24**	**95.29**
FT-Palmira	80.73	187.03	103.58	21.60	90.77	94.30
FT-Vanilla Mask R-CNN	82.20	210.40	111.56	21.00	87.34	93.48
Vanilla Mask R-CNN	30.59	620.17	546.77	208.61	75.28	77.31
Palmira	8.63	629.39	552.70	211.76	73.02	76.80

Table 6.1: Quantitative results of our newly proposed and other baseline models on IMMI (Section 5.3). ↑ indicates larger is better, ↓ indicates smaller is better.

6.2.4 FT-Vanilla-MRCNN

In this configuration, the starting point is the architecture of Vanilla Mask R-CNN of Prusty et al. [65]. We initialize this architecture with pre-trained MS-COCO weights [52] within the backbone component. We modify the region classifier branch to predict a single class. Although predicting a single class seems redundant, this design choice ensures that we can use the thresholded prediction score as an effective strategy for eliminating low-quality predictions in practice.

6.2.5 FT-Palmira

For this configuration, we follow a procedure similar to that followed for FT-Vanilla-MRCNN (i.e. initialization with pre-trained MS-COCO weights, single class region classifier branch), but apply these modifications to the more recent Palmira architecture – see Figure 6.1.

6.2.6 FT-Palmira-AS

This configuration is the same as *FT-Palmira* mentioned above. However, the aspect ratios and sizes of the anchor boxes in the Region Proposal Network (RPN) stage are customised keeping in mind the typical aspect ratios of manuscripts within document images of IMMI dataset. Specifically, we use aspect ratios of $1, 3, 10$ and sizes $[64, 128, 256, 512, 1024]$ to generate anchors that can accommodate the range of manuscript dimensions. Given the relatively smaller instance count of manuscripts (compared to region instances for which Palmira was originally trained), the number of proposals generated by RPN is reduced to 512. In addition, modifications included changing the network hyperparameters such as increasing the weight of the mask prediction loss in the final loss formulation and introducing focal loss [53].

6.3 Results

Table 6.1 contains performance of various models on our IMMI dataset in terms of IOU, Accuracy, AP, HD and its variants. The baselines (models optimized for manuscript *region* instance segmentation) – Palmira [78] and Vanilla Mask R-CNN [65] – were trained on images with not more than two manuscripts per image. Clearly, they perform very poorly at segmenting manuscript boundaries in IMMI where the number of manuscripts per image can be much greater. This is also evident from the qualitative results in Figure 6.4. The benefits of fine-tuning the baseline models on IMMI dataset is also evident from Table 6.1. The substantial improvement is very obvious in all the metrics reported. The slight but useful improvement in performance when the baseline hyperparameters and architectural components are customized for manuscript aspect ratios (*FT-Palmira-AS*) is also visible. The qualitative results from Figure 6.3 also highlight the effectiveness of proposed networks for challenging settings such as contiguous or overlapping images and deformations.

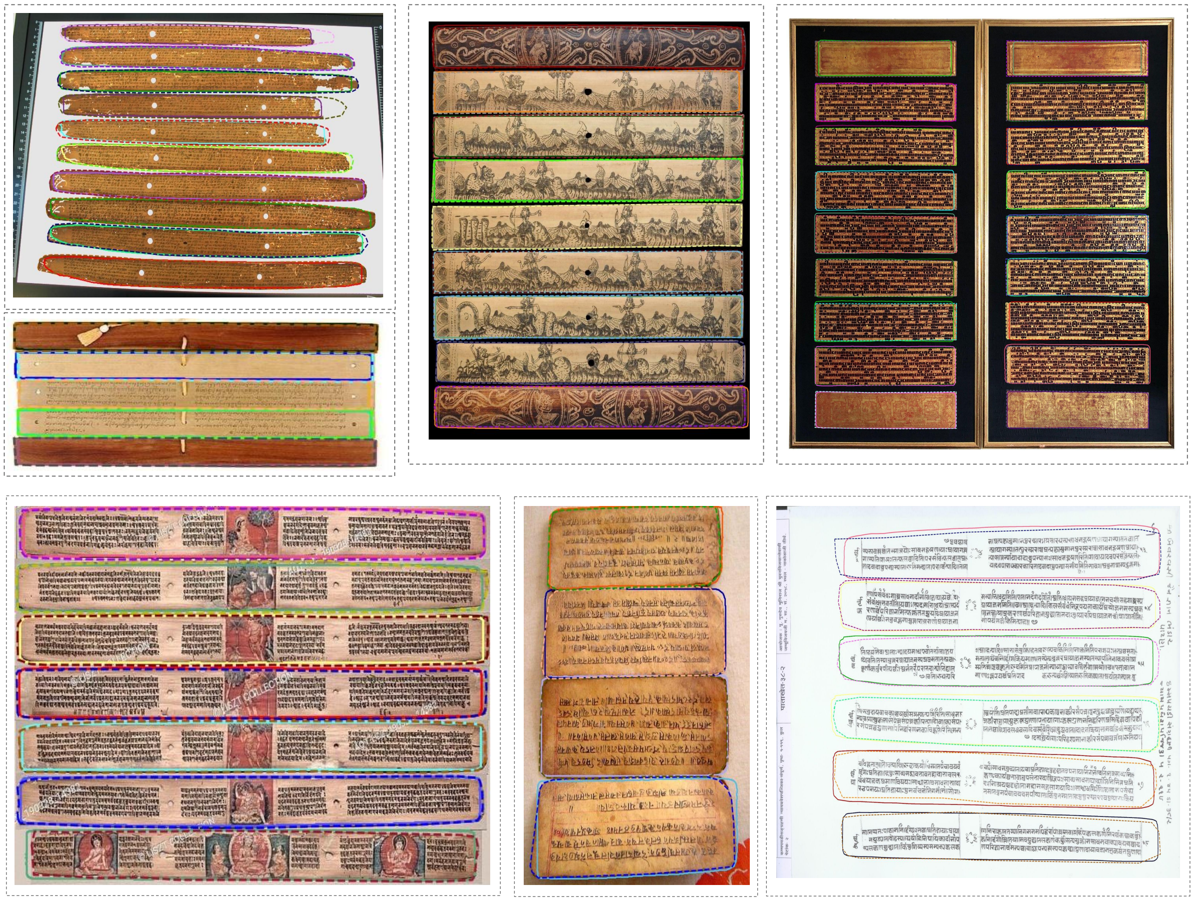


Figure 6.3: Test set images superimposed with ground truth and our prediction output represented by dotted and solid lines respectively. Please zoom in for better detail.

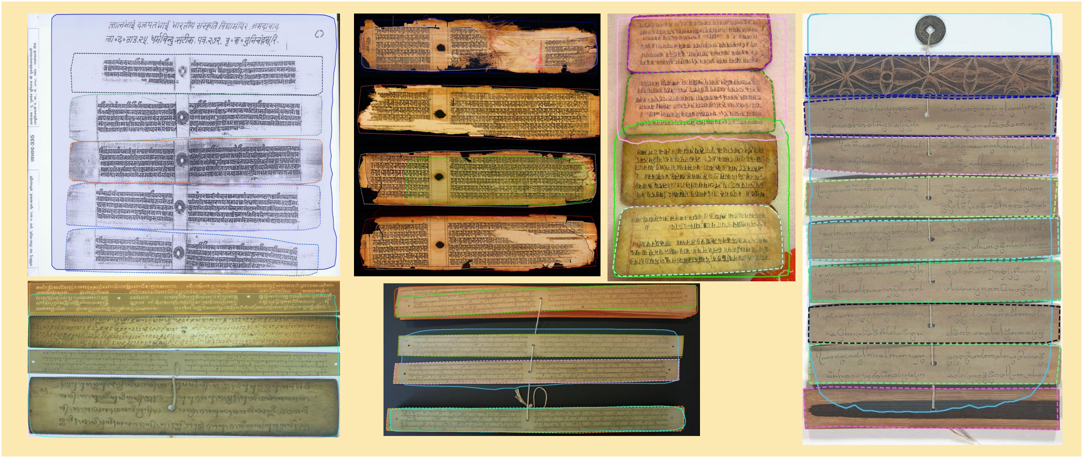


Figure 6.4: Qualitative results of pretrained models. Note that these are models trained for manuscript *region* instance segmentation. The inability of these models to handle multiple manuscripts is evident.

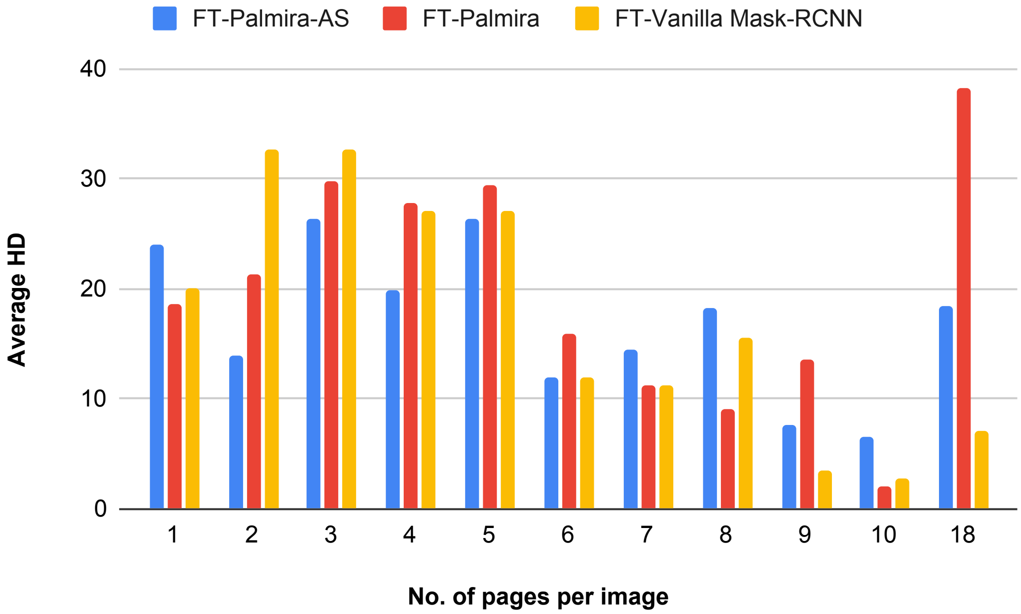


Figure 6.5: Average HD vs No. of Manuscripts per image.

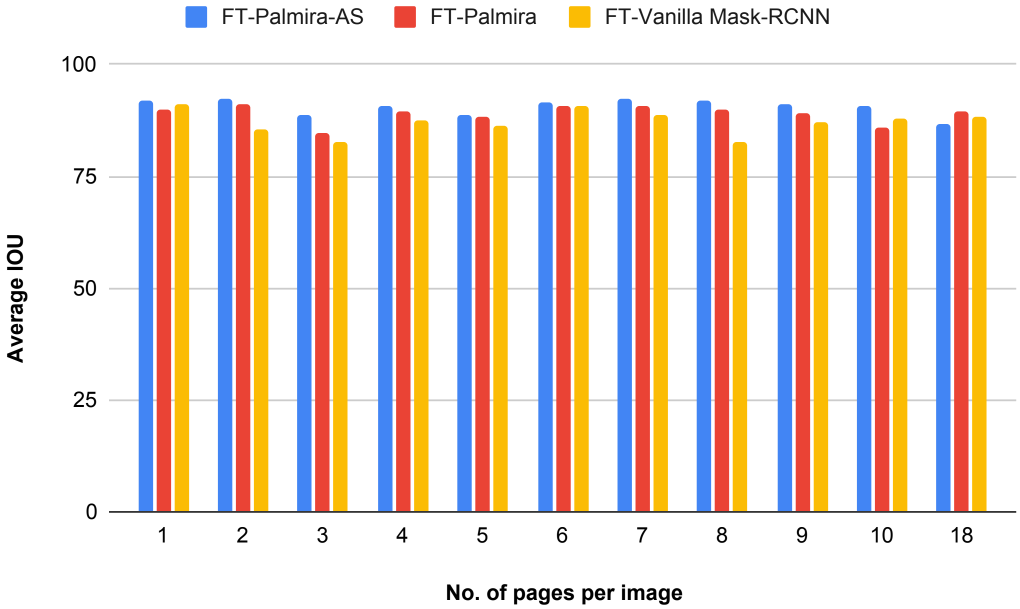


Figure 6.6: Average IOU vs No. of Manuscripts per image.

In addition to the above, Average HD and Average IOU are evaluated based on the number of manuscripts per image. The metrics corresponding to the images containing same number of manuscripts are averaged and reported in Figures 6.5 6.6 for all the fine-tuned models. *FT-Palmira-AS* dominates the other two models in terms of Average IoU for all the cases. Although relative performance of the models can be seen in the Average IOU plot (Figure 6.6), the differences are not very evident. In contrast, the Average HD plot (Figure 6.5) clearly depicts the relative differences, especially the large difference in performance for the most extreme case (18 manuscripts per document). More broadly, the results highlight the necessity of analyzing relative performance using multiple metrics. From the graphs shown in Figures 6.5,6.6, *FT-Palmira-AS* can be seen to obtain best results on average. Although the models were exposed to images containing not more than 10 manuscripts,they generalize so well that they predict instances even on the images containing 18 manuscripts.Qualitative results regarding the same can be seen in 6.3. The importance of specialized subcomponents designed to handle image deformations (deformable convolutions and deformable grid processing in mask branch) can also be seen by comparing the performance of Palmira variants (which contain these subcomponents) and *Vanilla Mask R-CNN* (see Table 6.1).

Chapter 7

Conclusions and Future Directions

This thesis primarily focuses on identifying the limitations and attempting to propose solutions in automatically deciphering the content from historical document images. Firstly, we started out by introducing Indiscapes,the first ever dataset with layout annotations for historical Indic manuscripts with the intent of expanding it later in terms of diversity such as layout, script and language. Alongside,we also adapted a deep-network based instance segmentation framework (Mask R-CNN) custom modified for fully automatic layout parsing. We achieved decent results using this model but it failed to generalise in few of the cases as it was limited to documents from only 2 sources and few challenges.

However, in a later stage, we expanded the dataset as intended before to Indiscapes2 (Sec. 3.2.1) by enhancing the size, diversity and robustness. Furthermore, we developed a novel deep learning based framework for fully automatic region-level instance segmentation of handwritten documents containing dense and uneven layouts, called Palm leaf Manuscript Region Annotator or PALMIRA in short (Sec. 5.2). Our experiments were able to demonstrate that Palmira performs qualitatively and quantitatively better than the earlier methods, strong baselines, and their ablative versions (Sec. 6.3). As part of our evaluation strategy for characterising the performance of document region boundary prediction, we also proposed boundary-aware measures such as Hausdorff distance and its variants in addition to the traditional area-centric measures such as Intersection-over-Union (IoU) and mean Average Precision (AP). We also demonstrated PALMIRA's out-of-dataset generalization ability via predictions on South-East Asian, Arabic and Hebrew manuscripts. Consequently, we discovered another challenge where multiple manuscripts are scanned together into a single scanned image for variety of reasons during the digitization process. However, our current models can handle one or two manuscripts per image. We thereby introduced IMMI (Indic Multi Manuscript Images), an annotated dataset with variability in per-image manuscript count, to compensate for the shortage of datasets in this field.We also proposed a synthetic data generation procedure and used the resultant data to make up for the imbalance and data deficit caused by the real-world distribution of multi-page document images. Finally,we also incorporated deep networks to handle semantic region deformations for instance segmentation of document images. Despite overlap and variation in the number of manuscripts per image, our experimental results showed the efficacy of the proposed deep networks for the task of manuscript instance segmentation.

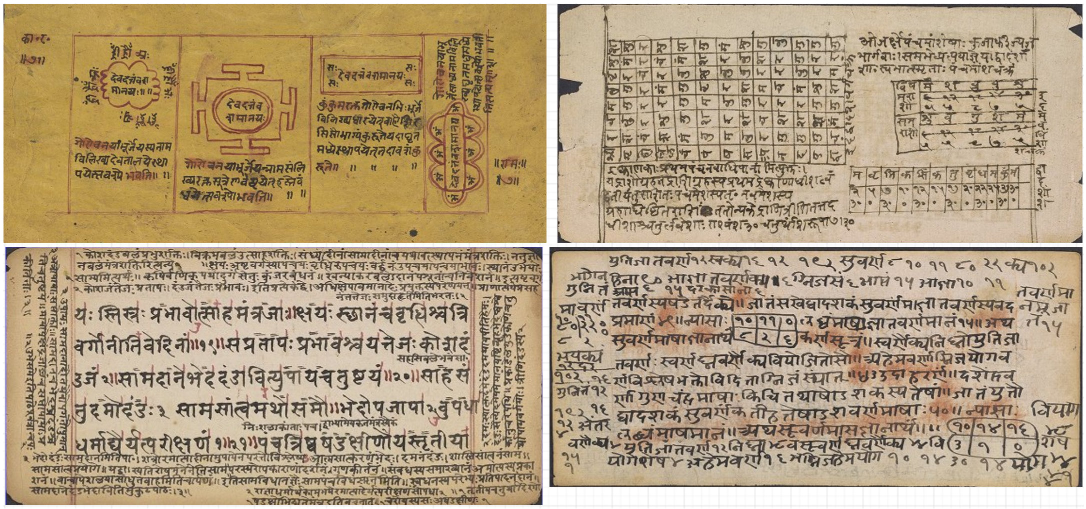


Figure 7.1: Images containing more challenges such as vertical text,more dense text lines and tables/Grids

Overall, our contributions made it possible to reliably recover individual historical manuscript pages for subsequent processing and tasks such as region-level instance segmentation and optical character recognition.

Moving forward, there are numerous significant challenges in this area. Future research could go into any of three directions.

1. Historical Indic manuscripts pose an ample assortment of challenges. There are still a great deal of layout elements such as tables or grid such as elements, oriented text, mixed graphic,(refer fig 7.1) that haven't been addressed or touched in terms of data till now. The datasets can be made even more robust and generic by adding more diversity in terms of size, layout, script and language

2. As we continue to believe that our contributions so far are good enough to aid in advancing robust layout estimates for handwritten documents, it can be planned to add downstream processing modules such as OCR for an end-to-end optimization.

3. Although we have good layout segmentation results, there is always room to improve the performance using new methodologies and challenges.

Related Publications

1. Prusty Abhishek, Sowmya Aitha, Abhishek Trivedi, and Ravi Kiran Sarvadevabhatla. “Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts.” In 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 999-1006. IEEE, 2019.

2. Sharan S. P. , Sowmya Aitha, Amandeep Kumar, Abhishek Trivedi, Aaron Augustine, and Ravi Kiran Sarvadevabhatla. “Palmira: a deep deformable network for instance segmentation of dense and uneven layouts in handwritten manuscripts.” In International Conference on Document Analysis and Recognition, pp. 477-491. Springer, Cham, 2021..

3. Aitha Sowmya, Sindhu Bollampalli, and Ravi Kiran Sarvadevabhatla.“Deformable deep networks for instance segmentation of overlapping multi page handwritten documents.” In Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1-9. 2021.

Bibliography

[1] Penn in hand: Selected manuscripts. http://dla.library.upenn.edu/dla/medren/search.html?fq=collection_facet:"IndicManuscripts". 12

[2] Web aletheia. 4, 6

[3] *Proc. 3rd Intl. Wksp on Historical Document Imaging and Processing, HIP@ICDAR 2015*. ACM, 2015. 3, 22

[4] *Proc. 4th Intl. Workshop on Historical Document Imaging and Processing, Kyoto, Japan, November 10-11, 2017*. ACM, 2017. 3, 22

[5] A. Abeysinghe and A. Abeysinghe. Use of neural networks in archaeology: preservation of assamese manuscripts. International Seminar on Assamese Culture & Heritage, 2018. 3, 22

[6] M. Agarwal, A. Mondal, and C. Jawahar. Cdec-net: Composite deformable cascade network for table detection in document images. *ICPR*, 2020. 23

[7] M. Agarwal, A. Mondal, and C. V. Jawahar. Cdec-net: Composite deformable cascade network for table detection in document images. *CoRR*, abs/2008.10831, 2020. 35

[8] R. Alaasam, B. Kurar, and J. El-Sana. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 738–742. IEEE, 2019. 23

[9] R. Alaasam, B. Kurar, and J. El-Sana. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 738–742, 2019. 35

[10] M. Alberti, L. Voegtlin, V. Pondenkandath, M. Seuret, R. Ingold, and M. Liwicki. Labeling, Cutting, Grouping: an Efficient Text Line Segmentation Method for Medieval Manuscripts. In *2019 15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, sep 2019. 35

[11] M. Alberti, L. Vögtlin, V. Pondenkandath, M. Seuret, R. Ingold, and M. Liwicki. Labeling, cutting, grouping: An efficient text line segmentation method for medieval manuscripts. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1200–1206. IEEE, 2019. 23

[12] B. Barakat, A. Droby, M. Kassis, and J. El-Sana. Text line segmentation for challenging handwritten document images using fully convolutional network. In *ICFHR*, pages 374–379. IEEE, 2018. 3, 22

[13] R. Barman, M. Ehrmann, S. Clematide, S. Oliveira, and F. Kaplan. Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining & Digital Humanities*, HistoInformatics, 01 2021. 35

[14] R. Barman, M. Ehrmann, S. Clematide, S. A. Oliveira, and F. Kaplan. Combining visual and textual features for semantic segmentation of historical newspapers. *arXiv preprint arXiv:2002.06144*, 2020. 23

[15] S. Bukhari, T. Breuel, A. Asi, and J. El-Sana. Layout analysis for arabic historical document images using machine learning. pages 639–644, 09 2012. 35

[16] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana. Layout analysis for arabic historical document images using machine learning. In *ICFHR 2012*, pages 639–644. IEEE, 2012. 3, 22

[17] K. Chen, M. Seuret, J. Hennebert, and R. Ingold. Convolutional neural networks for page segmentation of historical document images. In *ICDAR*, volume 1, pages 965–970. IEEE, 2017. 3, 22, 26

[18] T. Cheng, X. Wang, L. Huang, and W. Liu. Boundary-preserving mask r-cnn. In *European Conference on Computer Vision*, pages 660–676. Springer, 2020. 29, 31

[19] C. Clausner, A. Antonacopoulos, T. Derrick, and S. Pletschacher. Icdar2017 competition on recognition of early indian printed documents-reid2017. In *ICDAR*, volume 1, pages 1411–1416. IEEE, 2017. 3, 22

[20] C. Clausner, A. Antonacopoulos, T. Derrick, and S. Pletschacher. Icdar2019 competition on recognition of early indian printed documents – reid2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1527–1532, 2019. 35

[21] C. Clausner, A. Antonacopoulos, N. Mcgregor, and D. Wilson-Nunn. Icfhr 2018 competition on recognition of historical arabic scientific manuscripts – rasm2018. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476, 2018. 17, 35

[22] C. Clausner, A. Antonacopoulos, and S. Pletschacher. Icdar2019 competition on recognition of documents with complex layouts - rdcl2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1521–1526, 2019. 35

[23] C. Clausner, A. Antonacopoulos, and S. Pletschacher. Icdar2019 competition on recognition of documents with complex layouts-rdcl2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1521–1526. IEEE, 2019. 23

[24] C. Clausner, S. Pletschacher, and A. Antonacopoulos. Aletheia-an advanced document layout and text ground-truthing system for production environments. In *ICDAR*, pages 48–52. IEEE, 2011. 3, 6

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. 26

[26] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 23, 27, 36

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 25

[28] D. Doermann, E. Zotkina, and H. Li. GEDI-a groundtruthing environment for document images. In *Ninth IAPR Intl. Workshop on Document Analysis Systems*, 2010. 3, 6

[29] J. Gao, Z. Wang, J. Xuan, and S. Fidler. Beyond fixed grid: Learning geometric image representation with a deformable grid. In *European Conference on Computer Vision*, pages 108–125. Springer, 2020. 29

[30] J. Gao, Z. Wang, J. Xuan, and S. Fidler. Beyond fixed grid: Learning geometric image representation with a deformable grid. In *ECCV*, 2020. 38

[31] A. Garai, S. Biswas, S. Mandal, and B. Chaudhuri. A method to generate synthetically warped document image. *ArXiv*, abs/1910.06621, 2019. 18

[32] A. Garz, M. Seuret, F. Simistira, A. Fischer, and R. Ingold. Creating ground truth for historical manuscripts with document graphs and scribbling interaction. In *DAS*, pages 126–131. IEEE, 2016. 3, 6

[33] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J. A. Sánchez, A. H. Toselli, and E. Vidal. Ground-truth production in the transcriptorium project. In *DAS*, pages 237–241. IEEE, 2014. 4, 6

[34] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928. 18

[35] A. Gupta, P. Dollár, and R. B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. *CoRR*, abs/1908.03195, 2019. 28

[36] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *ICCV*, pages 2980–2988, 2017. 26, 36

[37] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 23

[38] R. Karpinski and A. Belaïd. Semi-synthetic data augmentation of scanned historical documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 268–273, 2019. 18

[39] M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam, and J. El-Sana. Vml-hd: The historical arabic documents dataset for recognition systems. In *1st Intl. Workshop on Arabic Script Analysis and Recognition*. IEEE, 2017. 3, 22

[40] M. W. A. Kesiman, J.-C. Burie, G. N. M. A. Wibawantara, I. M. G. Sunarya, and J.-M. Ogier. Amadi_lontarset: The first handwritten balinese palm leaf manuscripts dataset. In *ICFHR*, pages 168–173. IEEE, 2016. 3, 22

[41] M. W. A. Kesiman, G. A. Pradnyana, and I. M. D. Maysanjaya. Balinese glyph recognition with gabor filters. *Journal of Physics: Conference Series*, 1516:012029, apr 2020. 23

[42] M. W. A. Kesiman, D. Valy, J. Burie, E. Paulus, M. Suryani, S. Hadi, M. Verleysen, S. Chhun, and J. Ogier. ICFHR 2018 competition on document image analysis tasks for southeast asian palm leaf manuscripts. In *ICFHR*, pages 483–488, 2018. 3, 22

[43] M. W. A. Kesiman, D. Valy, J.-C. Burie, E. Paulus, M. Suryani, S. Hadi, M. Verleysen, S. Chhun, and J.-M. Ogier. Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia. *Journal of Imaging*, 4:43, 02 2018. 1, 3, 17, 22, 35

[44] C. Kieu, N. Journet, M. Visani, J.-P. Domenger, and R. Mullot. Semi-synthetic document image generation using texture mapping on scanned 3d document shapes. 08 2013. 18

[45] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 29, 31

[46] R. Klette and A. Rosenfeld, editors. *Digital Geometry*. The Morgan Kaufmann Series in Computer Graphics. Morgan Kaufmann, San Francisco, 2004. 30

[47] D. U. Kumar, G. Sreekumar, and U. Athvankar. Traditional writing system in southern india—palm leaf manuscripts. *Design Thoughts*, 9, 2009. 1, 2

[48] J. Lee, H. Hayashi, W. Ohyama, and S. Uchida. Page segmentation using a convolutional neural network with trainable co-occurrence features. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1023–1028. IEEE, 2019. 23

[49] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou. Docbank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, 2020. 23

[50] J. Liang, Q. Hu, P. Zhu, and W. Wang. Efficient multi-modal geometric mean metric learning. *Pattern Recognition*, 75:188–198, 2018. 23

[51] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017. 23

[52] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 25, 26, 39

[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 25, 27, 39

[54] W. Ma, H. Zhang, L. Jin, S. Wu, J. Wang, and Y. Wang. Joint layout analysis, character detection and recognition for historical document digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–36. IEEE, 2020. 23

[55] W. Ma, H. Zhang, L. Jin, S. Wu, J. Wang, and Y. Wang. Joint layout analysis, character detection and recognition for historical document digitization. *CoRR*, abs/2007.06890, 2020. 35

[56] D. Made Sri Arsa, G. Agung Ayu Putri, R. Zen, and S. Bressan. Isolated handwritten balinese character recognition from palm leaf manuscripts with residual convolutional neural networks. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pages 224–229, 2020. 23

[57] J. Martínek, L. Lenc, and P. Král. Building an efficient ocr system for historical documents with little training data. *Neural Computing and Applications*, 32, 12 2020. 17

[58] T. Monnier and M. Aubry. docExtractor: An off-the-shelf historical document element extraction. In *ICFHR*, 2020. 23

[59] T. Monnier and M. Aubry. docextractor: An off-the-shelf historical document element extraction. *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 91–96, 2020. 35

[60] S. A. Oliveira, B. Seguin, and F. Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE, 2018. 23

[61] N. S. Panyam, V. L. T.R., R. Krishnan, and K. R. N.V. Modeling of palm leaf character recognition system using transform based techniques. *Pattern Recogn. Lett.*, 84(C), Dec. 2016. 3, 22

[62] A. Pappo-Toledano, F. Chen, G. Latif, and L. Alzubaidi. Adoptive thresholding and geometric features based physical layout analysis of scanned arabic books. *2018 IEEE 2nd Intl. Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 171–176, 2018. 3, 22

[63] E. Paulus, M. Suryani, and S. Hadi. Improved line segmentation framework for sundanese old manuscripts. *Journal of Physics: Conference Series*, 978:012001, mar 2018. 3, 22

[64] E. Paulus, M. Suryani, and S. Hadi. Improved line segmentation framework for sundanese old manuscripts. In *Journal of Physics: Conference Series*, volume 978, page 012001. IOP Publishing, 2018. 23

[65] A. Prusty, S. Aitha, A. Trivedi, and R. K. Sarvadevabhatla. Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 999–1006. IEEE, 2019. 13, 14, 17, 23, 26, 29, 31, 35, 36, 39

[66] W. Puarungroj, N. Boonsirisumpun, P. Kulna, T. Soontarawirat, and N. Puarungroj. Using deep learning to recognize handwritten thai noi characters in ancient palm leaf manuscripts. In *International Conference on Asian Digital Libraries*, pages 232–239. Springer, 2020. 23

[67] Y. B. Rachman. Palm leaf manuscripts from royal surakarta, indonesia: Deterioration phenomena and care practices. *Intl. Journal for the Preservation of Library and Archival Material*, 39(4):235–247, 2018. 1

[68] T. M. Rath and R. Manmatha. Word spotting for historical documents. *IJDAR*, 9(2-4):139–152, 2007. 3, 22

[69] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 25

[70] G. Renton, C. Chatelain, S. Adam, C. Kermorvant, and T. Paquet. Handwritten text line segmentation using fully convolutional network. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 05, pages 5–9, 2017. 35

[71] C. Reul, M. Dittrich, and M. Gruner. Case study of a highly automated layout analysis and ocr of an incunabulum:'der heiligen leben'(1488). In *Proc. 2nd Intl. Conf. on Digital Access to Textual Cultural Heritage*, pages 155–160. ACM, 2017. 3, 22

[72] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 23

[73] R. S. Sabeenian, M. E. Paramasivam, P. M. Dinesh, R. Adarsh, and G. R. Kumar. Classification of handwritten tamil characters in palm leaf manuscripts using svm based smart zoning strategies. In *ICBIP*. ACM, 2017. 3, 22

[74] J. Sahoo. A selective review of scholarly communications on palm leaf manuscripts. *Library Philosophy and Practice (e-journal)*, 2016. 1, 2

[75] J. A. Sánchez, V. Bosch, V. Romero, K. Depuydt, and J. De Does. Handwritten text recognition for historical documents in the transcriptorium project. In *Proc. of the First Intl. Conf. on Digital Access to Textual Cultural Heritage*, pages 111–117. ACM, 2014. 3, 22

[76] P. N. Sastry, T. V. Lakshmi, N. K. Rao, and K. RamaKrishnan. A 3d approach for palm leaf character recognition using histogram computation and distance profile features. In *Proc. 5th Intl. Conf. on Frontiers in Intelligent Computing: Theory and Applications*, pages 387–395. Springer, 2017. 3, 22

[77] C. K. Savitha and P. J. Antony. Machine learning approaches for recognition of offline tulu handwritten scripts. *Journal of Physics: Conference Series*, 1142:012005, nov 2018. 3, 22

[78] S. P. Sharan, S. Aitha, K. Amandeep, A. Trivedi, A. Augustine, and R. K. Sarvadevabhatla. Palmira: A deep deformable network for instance segmentation of dense and uneven layouts in handwritten manuscripts. In *International Conference on Document Analysis and Recognition, ICDAR 2021*, 2021. 17, 19, 35, 36, 39

[79] Z. Shen, K. Zhang, and M. Dell. A large dataset of historical japanese documents with complex layouts. pages 2336–2343, 06 2020. 35

[80] Z. Shi, S. Setlur, and V. Govindaraju. Digital enhancement of palm leaf manuscript images using normalization techniques. In *5th Intl. Conf. On Knowledge Based Computer Systems*, pages 19–22, 2004. 3, 22

[81] S. Siddiqui, M. Malik, S. Agne, A. Dengel, and S. Ahmed. Decnt: Deep deformable cnn for table detection. *IEEE Access*, 6:74151–74161, 2018. 23

[82] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold. Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. In *ICFHR*, pages 471–476. IEEE, 2016. 3, 22, 35

[83] U. Springmann and A. Luedeling. Ocr of historical printings with an application to building diachronic corpora: A case study using the ridges herbal corpus. *Digital Humanities Quarterly*, (2), 2017. 3, 22

[84] F. Stahlberg and S. Vogel. Qatip – an optical character recognition system for arabic heritage collections in libraries. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 168–173, 2016. 17

[85] D. Sudarsan, P. Vijayakumar, S. Biju, S. Sanu, and S. K. Shivadas. Digitalization of malayalam palmleaf manuscripts based on contrast-based adaptive binarization and convolutional neural networks. In *Intl. Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2018. 3, 22

[86] M. Suryani, E. Paulus, S. Hadi, U. A. Darsa, and J.-C. Burie. The handwritten sundanese palm leaf manuscript dataset from 15th century. In *ICDAR*, pages 796–800. IEEE, 2017. 3, 22

[87] Z. Tian, C. Shen, and H. Chen. Conditional convolutions for instance segmentation. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 282–298, Cham, 2020. Springer International Publishing. 29, 31

[88] A. Trivedi and R. K. Sarvadevabhatla. Hindola: A unified cloud-based platform for annotation, visualization and machine learning-based layout analysis of historical manuscripts. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 31–35. IEEE, 2019. 6, 14, 19

[89] D. Valy, M. Verleysen, S. Chhun, and J.-C. Burie. A new khmer palm leaf manuscript dataset for document analysis and recognition: Sleukrith set. In *Proc. of the 4th Intl. Workshop on Historical Document Imaging and Processing*, pages 1–6. ACM, 2017. 2, 3, 22

[90] D. Valy, M. Verleysen, S. Chhun, and J.-C. Burie. A new khmer palm leaf manuscript dataset for document analysis and recognition: Sleukrith set. pages 1–6, 11 2017. 35

[91] D. Valy, M. Verleysen, S. Chhun, and J.-C. Burie. Character and text recognition of khmer historical palm leaf manuscripts. In *ICFHR*, pages 13–18, 08 2018. 3, 22

[92] D. Valy, M. Verleysen, S. Chhun, and J.-C. Burie. Character and text recognition of khmer historical palm leaf manuscripts. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 13–18. IEEE, 2018. 23

[93] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. Perantonis. A complete optical character recognition methodology for historical documents. *Document Analysis Systems, IAPR International Workshop on*, 0:525–532, 09 2008. 17

[94] H. Wei, M. Seuret, K. Chen, A. Fischer, M. Liwicki, and R. Ingold. Selecting autoencoder features for layout analysis of historical documents. In *Proc. 3rd Intl. Workshop on Historical Document Imaging and Processing*, HIP '15, pages 55–62. ACM, 2015. 3, 22

[95] C. Wick and F. Puppe. Fully convolutional neural networks for page segmentation of historical document images. In *DAS*, pages 287–292. IEEE, 2018. 3, 22

[96] M. Würsch, R. Ingold, and M. Liwicki. Divaservices—a restful web service for document image analysis methods. *Digital Scholarship in the Humanities*, 32(1):i150–i156, 2016. 4, 6

[97] X. Zhong, J. Tang, and A. Jimeno-Yepes. Publaynet: largest dataset ever for document layout analysis. *CoRR*, abs/1908.07836, 2019. 35

[98] X. Zhong, J. Tang, and A. J. Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep. 2019. 23