# HInDoLA: A Unified Cloud-based Platform for Annotation, Visualization and Machine Learning-based Layout Analysis of Historical Manuscripts

Abhishek Trivedi, Ravi Kiran Sarvadevabhatla
Centre for Visual Information Technology (CVIT)
International Institute of Information Technology, Hyderabad (IIIT-H)
Gachibowli, Hyderabad 500032, INDIA.
{abhishek.trivedi@research.,ravi.kiran@}iiit.ac.in

***Abstract*—Palm-leaf manuscripts are one of the oldest medium of inscription in many Asian countries. Especially, manuscripts from the Indian subcontinent form an important part of the world's literary and cultural heritage. Despite their significance, large-scale datasets for layout parsing and targeted annotation systems do not exist. Addressing this, we propose a web-based layout annotation and analytics system. Our system, called HInDoLA, features an intuitive annotation GUI, a graphical analytics dashboard and interfaces with machine-learning based intelligent modules on the backend. HInDoLA has successfully helped us create the first ever large-scale dataset for layout parsing of Indic palm-leaf manuscripts. These manuscripts, in turn, have been used to train and deploy deep-learning based modules for fully automatic and semi-automatic instance-level layout parsing.**

## I. Introduction

Palm leaf manuscripts written in different Indian languages are scattered all over the country in monasteries, temples, libraries, museums, with individuals and in several private collections. In contrast with modern or recent era documents, such manuscripts are considerably more fragile, prone to degradation from elements of nature and tend to have a short shelf life. Moreover, the domain experts who can decipher such content are small in number and dwindling. Manuscripts give a lasting impression of the multicultural society that India and many South-east Asian countries are and its deep-rooted knowledge system that has been passed down from generations. Therefore, it is essential to access the content within these documents before it is lost forever. Many studies on Indic palm-leaf and paper-based manuscripts exist, but these are typically conducted on small and often, private collections of documents [1], [2]. No publicly available large-scale, annotated dataset of historical Indic manuscripts exists to the best of our knowledge.

In this paper, we take a significant step to address the issue by building an annotation and analytics ecosystem called Historical Intelligent Document Layout Analytics or HInDoLA for short. We are certainly not the first to develop such a system. Indeed, many useful annotation and analysis tools exist to facilitate progress in creation and analysis of historical document manuscripts [3]–[5]. The system we propose contains many of the features found in existing annotation systems. However, some of these systems are primarily oriented towards offline annotation, single user interaction and are unable to provide a unified management of annotation process and monitoring of annotation performance. In contrast, our web-based system addresses these aspects and provides additional capabilities. The additional features are tailored for annotation and examining annotation analytics for documents with dense and irregular layout elements, especially those found in Indic manuscripts. More generally, our system is an instance of the recent trend of collaborative, cloud/web based annotation systems and services [6], [7]. Our system is completely open-source and contains documentation for ease of installation and use.

The overall architecture of HInDoLA comprises of three major components - Annotation Tool, Dashboard Analytics and ML Engines. An overview of the architecture can be seen in Figure 1. Next, we shall look at the components of HInDoLA individually.

## II. Annotation Tool

The Annotation tool enables interactive annotation of manuscripts with user level involvement. The tool requires annotators to perform a one-time registration. This enables tracking their annotation performance at multiple levels of granularity. It is configured as a web-based service available at a pre-specified URL. However, the tool is designed such that it can operate offline as well. This is especially convenient for sandboxing and extending the functionality of the tool. The offline version of the tool is also convenient when the server-based components of HInDoLA cannot be installed due to access restrictions. The web-based version, however, is more flexible, allows multiple concurrent annotation sessions and enables addition of new features without the intervention of users. Most systems like [7], [8] have also released their tools in web based version.

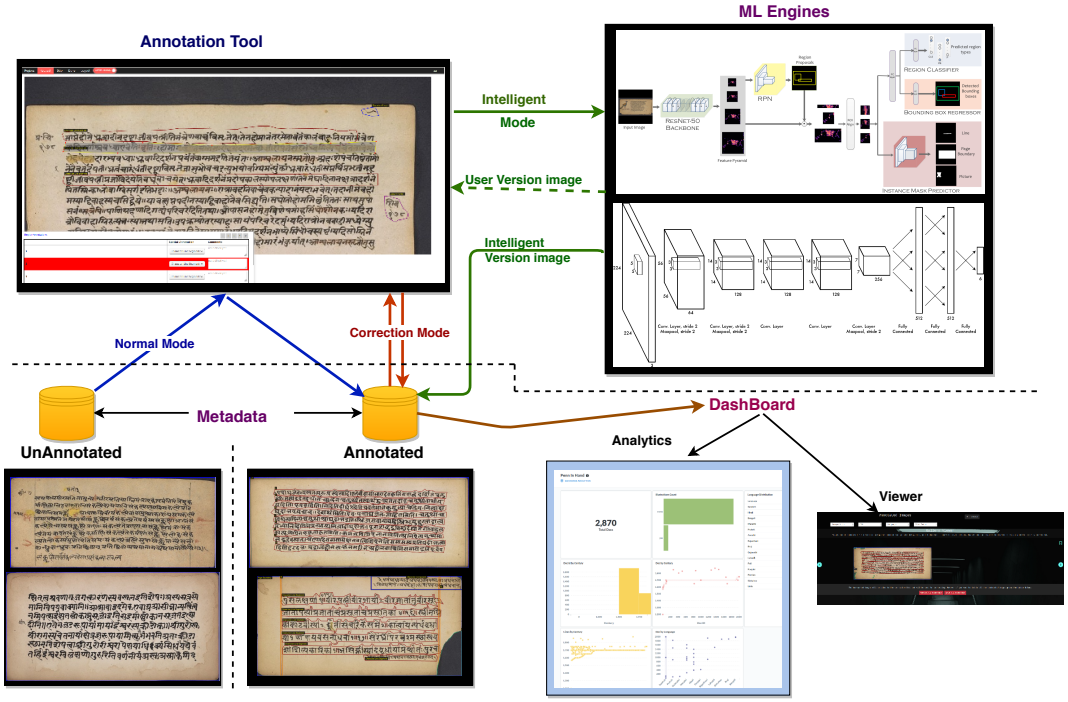


Fig. 1: The Architecture of HInDoLA

### A. Region Types & Spatial Annotations

For annotating components of the manuscript, we have introduced a free-hand drawing feature along with the usual polygon and rectangle draw feature. The free-hand region drawing feature enables rapidity in annotations of irregular, small regions and very large regions in the images thus providing maximum annotation flexibility. Just a single mouse click to start and smoothly moving the mouse along the component boundary creates a region. The polygon region is created by addition of vertices through mouse clicks. The ultra zoom and spanning feature enable the users to overcome the challenge of dense layout parsing and aid in pixel level accuracy of annotations. Once a region is drawn, it can be resized by dragging the vertices on the layout edge. Documents, even with irregular layouts, contain repetitive structural elements (e.g. lines). To exploit this repetitive structure, annotations of previous regions can be copied and shifted to yet unannotated regions of the image, thus saving time.

Once a region is annotated, an annotation window pop-up appears for label specification for the region. Keeping the Indic palm-leaf manuscripts in mind, we have introduced 9 different labels for regions on image - Character Line Segment, Character Component, Hole, Page Boundary, Library Marker, Decorator, Picture, Physical Degradation and Boundary Line. The region boundary of annotated components of the image are shown in different vibrant colors which helps in visual analysis of quality of annotations. Some annotated images and their intelligent mode predictions are shown in Figure 2. For enabling text annotations in future, we have also introduced a textbox along with label list.

All the annotations from the HInDoLA Annotation tool are extracted in JSON format. Along with the region coordinates and labels, the JSON file also stores time-stamp for each region annotated which serves as a metric to support training of intelligent models for automatic annotation (see Figure 3). The JSON schema is inspired from VIA [9] annotation scheme.

### B. Backend Annotation Manager

For handling HInDoLA annotation activities and sessions of different annotators, we have introduced an Annotation Manager which handles the distributed parallel sessions of registered annotators while managing and updating the database. The system is designed with a request queue which handles the sessions of multiple users working on the system simultaneously and loads un-annotated images in Normal mode and annotated images in Correction mode. The Correction mode toggle is useful for expert annotators to correct the annotations and thus improves the quality of annotations. The Skip button helps annotator to skip the served image if the image is too corrupted or if they are not sure about proper layout parsing of newly introduced components. The Done button saves the annotation JSON file at the backend folder. An interesting feature is the ability to save multiple copies of JSON files for the same document image which enables the admins to save the best version of the annotated image.

For building intelligent annotation systems, we configure HInDoLA to interface with Deep Network based machine-learning models. One of the models is fully automatic and provides a instance-level layout estimate for a given document. Another model is semi-automatic. The annotator provides partial information in the form of the region's bounding box.

| | 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 | 2455 | 521 | 3 | 262 | 33 | 60 | 95 | 35 | 404 |
| BHOOMI | 2476 | 202 | 563 | 314 | 134 | 6 | 2 | 2231 | 4 |
| Combined | 4931 | 723 | 566 | 576 | 167 | 66 | 97 | 2266 | 408 |

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

| | Result from Indiscapes (Instance semantic segmentation Model) Paper / Bounding Box Supervision Intelligent Mode | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | H | CLS | PD | PB | CC | P | D | LM | BL |
| PIH | $NA$ | 74.17/76.90 | $NA$ | 86.90/99.29 | 52.84/66.79 | 60.49/57.75 | 5.23/53.94 | 50.29/74.11 | 29.45/50.70 |
| Bhoomi | 79.29/76.78 | 29.07/72.53 | 8.72/70.06 | 91.09/91.28 | 32.50/65.85 | $NA$ | $NA$ | 38.25/76.65 | $NA$ |
| Combined | 79.29/76.78 | 57.77/74.97 | 8.72/70.06 | 88.47/96.79 | 45.87/66.49 | 60.49/57.75 | 5.23/53.94 | 42.93/75.56 | 29.45/50.70 |

TABLE II: Classwise average IoUs for Penn in Hand (PIH), Bhoomi and Combined datasets (reference Table I). The better of IoU scores is highlighted in red for easy comparison.

The model then predicts a tight region boundary estimate . Both the models predict masks for different regions. To store these estimates in the standard annotation point, control points along the boundary are estimated using the Ramer-Douglas-Peucker curve approximation method.

## III. HINDOLA & INDISCAPES : THE INDIC MANUSCRIPT DATASET

Using HInDoLA, we have created Indiscapes, the first ever detailed spatial layout annotated dataset for Indic manuscripts. The images in our dataset are obtained from two sources. The first source is publicly available Indic manuscript collection from University of Pennsylvania's Rare book and manuscript library [10], also referred to as Penn-in-Hand(PIH). From the 2,880 Indic manuscript book-sets, we annotated 204 manuscript images. The selection was aimed at maximizing the diversity in terms of various attributes such as script language, physical degradation and presence of non-textual elements (e.g. picturs, tables) and number of lines. The second source is Bhoomi, a collection of 322 images sourced from multiple Oriental research institutes and libraries across India. The latter set of images are characterized by relatively larger width, presence of multiple languages, close and irregular spaced text lines, binding holes and degradations. Overall, we annotated 526 annotated Indic manuscripts with HInDoLA. The statistics of the dataset are provided in Table I.

## IV. ML ENGINES

The ML Engines integration comprises of the intelligent region boundary predictions from deep learning models to enable fully-automatic or semi-automatic annotation system. The deep models can learn to automate annotations given initial annotated training data.

One of our models is the Instance Semantic Segmentation model which can isolate individual instances of each region in the manuscript image. We employ the Mask R-CNN [11] which has proven to be very effective at the task of object-instance segmentation in photos. The network architecture contains three stages - BackBone, Region Proposal Network and Multi-Task Branch networks. The network is initialized with weights obtained from a mask R-CNN trained on the MS-COCO [12] dataset with a ResNet-50 [13] backbone. Details about the model can be found in our work [14]. The workflow is as follows: A new image is served on Annotation tool upon request. If the user selects Intelligent Mode, the Instance segmentation prediction model becomes active. The model predicts the masks for different regions in the input image following which a sub-system generates control points on the boundaries of regions. These control points are overlaid on the input image and enable the annotator to modify the prediction as required to obtain a tight region boundary.

Another model for the intelligent system is the Bounding box supervision model which learns the edge features and generates the boundary around the region. Annotators select a bounding box around a region in a newly requested image which is automatically forwarded to the edge model at back-end. The model predicts edge features and edge-logits (pixel values ranging from 0-1) of the region through a Convolution Neural Network and generates a concave-hull boundary from the most prominent features. The boundary is then integrated with control points similar to Instance segmentation model boundary prediction which enables serving the region boundary annotation on the Annotation tool. Annotators may then adjust the control points to get a tight bounding box around the region. The details can be found in our technical report. Comparison of our current two intelligent models- Instance Semantic Segmentation and Bounding Box supervision model on basis of IoU (Intersection over Union) of predicted annotation boundary masks is detailed in Table II.

Currently, our system can also be configured with any deep learning based intelligent boundary prediction engines. Models just need to predict the masks and the system auto-generates control points (vertices) on the boundaries of masks to enable the serving of auto-annotated image on tool. This functionality saves the time of users and experts, otherwise spent on tedious task of annotating whole manuscript image from scratch.

## V. DASHBOARD

We introduce dashboard-style analytics keeping in mind the non-technical nature of domain experts and the unique

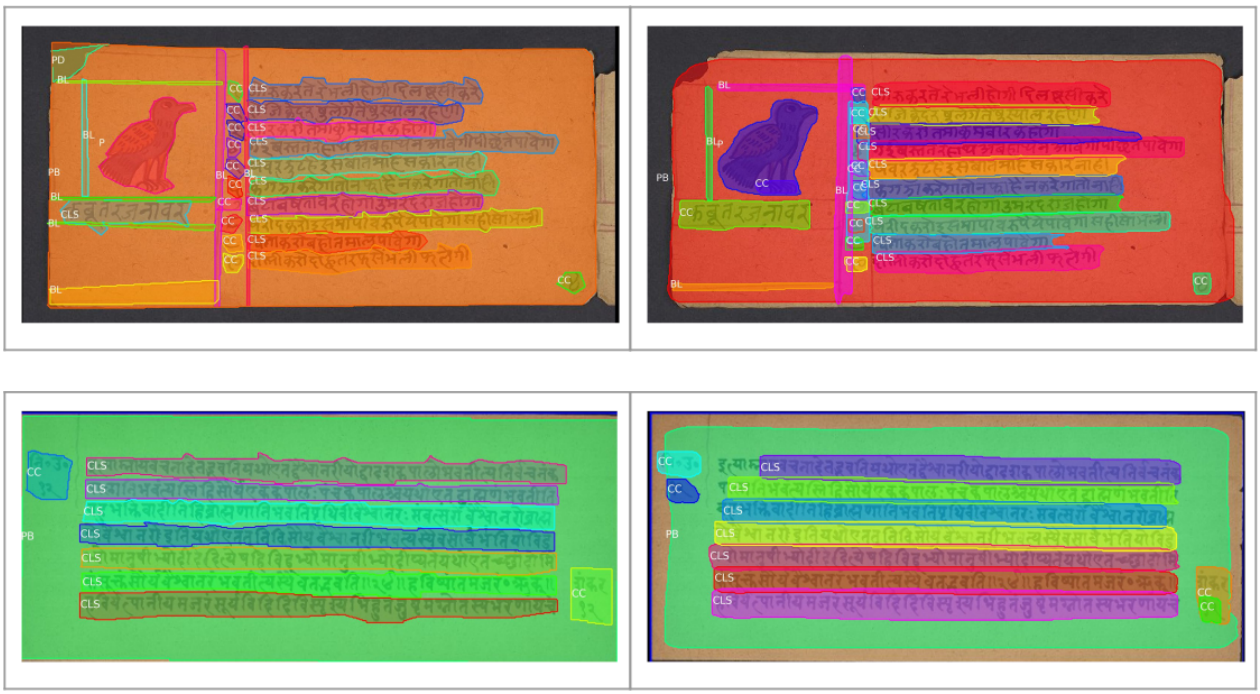

Fig. 2: Ground truth annotations (Normal Mode) (left) and predicted instance segmentations (right) for Intelligent Mode. Note that we use colored shading only to visualize individual region instances and not to color-code region types.

layout-level challenges of Indic Manuscripts. The dashboard features many annotation management activities such as displaying graphs for the annotation stats, progress report of the annotators. An additional feature is the database viewer. This component is helpful in monitoring the live annotation sessions and viewing annotation statistics of the documents in an interactive fashion.

### A. Annotation Analytics

This section of dashboard is responsible for displaying the annotation statistics. The manuscripts are grouped and displayed by annotated time, languages, number of regions etc. Various kind of histograms and pie-charts are shown on dashboard demonstrating number of experiments added, completed document annotations (snapshot in Figure 1). This functionality gives an edge over similar systems [6], [7], saving the time of researchers otherwise spent on book-keeping of annotation activities. The progress of annotators can be also be tracked via the score of annotated files done.

### B. Viewer

This section of dashboard monitors the quality of completed annotations. Annotated and un-annotated document images can be searched and displayed on the interface with the help of special search filters. Some of the basic filters are searching by users, date of annotation, language type and time taken for annotation. A special feature of 'Bookmark' is introduced to select priority files for annotation. Bookmarking of both un-annotated and annotated images is enabled, moving the files to special database tables with immediate serving on request into the tool from next annotation sessions.

## VI. The HInDoLA System Availability, Documentation and Hands-on Experience

The HInDoLA is designed for Palm-Leaf Manuscript document images. However, the system can be easily used for any other historical document collection. To enable such a possibility, we have setup Github repositories to access the components developed for the system along with documentation and usage guides.

- Annotation Tool - https://github.com/ihdia/hindola-backend
- Metadata - https://github.com/ihdia/hindola-dbs
- DashBoard- https://github.com/ihdia/hindola-dash
- Instance semantic segmentation prediction - https://github.com/ihdia/instance-segmentation-v1

Hands-on experience videos are recorded for each of these tools that clearly show how to setup and use/run all of their functionalities. The source code of these tools and their video links are fully available on their respective repositories. The tool can be deployed on a single machine with the help of scripts enabling incremental addition and modification of any historical document image collection and database files. The snapshots of the tools on run can be seen in Figure 1. Comparison of our Hindola system with other existing historical document image annotations and ground-truth creation systems is shown in Table III.

| | Online/Desktop Tool | Dashboard Analytics | Integration of ML Engines | Viewer(Bookmarking) | Annotations Export Format | Performance Analysis | Semi/Auto Supervised | Availability(as of July 2019) |
|---|---|---|---|---|---|---|---|---|
| HINDOLA | Online/Desktop | ✓✓ | ✓✓ | ✓✓ | VIA Json Schema | ✓✓ | Semi/Auto (Deep network) | Source-Code + Web Service |
| D.A.B [15] | Online | ✓✗ | ✓✗ | ✓✗ | XML | ✓✓ | N.A | Web Service |
| TRANSCRIPTORIUM PROJECT [16] | Online | ✓✗ | ✓✓ | ✓✗ | TEI | ✓✓ | Semi (Non-deep) | Web Service |
| DIVADIAWI [17] | Online (HTTP request) | ✓✗ | ✓✓ | ✓✗ | Json | ✓✓ | Semi/Auto (Non-deep) | Web Service (Restful API) |
| MIRADOR ANNOTATION TOOL | Online/Desktop | ✓✓ | ✓✗ | ✓✓ | Json (stringify) | ✓✗ | N.A | Source Code + Web Service |
| ALETHIA [5] | Desktop | ✓✗ | ✓✓ | ✓✓ | PAGE | ✓✓ | Semi/Auto (Non-deep) | Executable offline |

TABLE III: Comparison with existing Historical document image Annotation Tools.

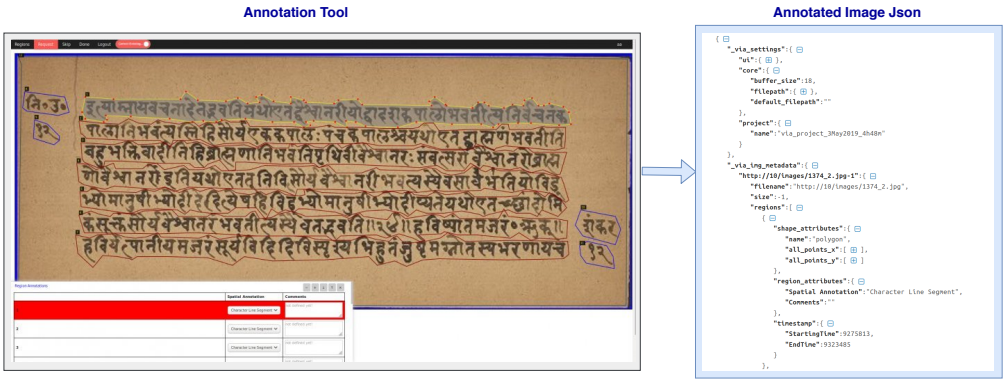


Fig. 3: Screenshots of our web-based Annotation Tool(left) and the Attributes of annotated region shown in annotated image Json (right)

## VII. CONCLUSION

The presented HInDoLA open-source cloud system is a collection of tools for large scale annotation and analysis of Indic palm-leaf manuscripts. The system not only contains state-of-art historical document layout analysis techniques (i.e Annotation Tool, Dashboard Analytics), but also the Semantic instance-level segmentation prediction. We believe that the availability of layout annotations will play a crucial role in reducing the overall complexity for subsequent OCR and similar tasks such as word-spotting, style-and-content based retrieval. All the utilities are made available for the research community with proper documentation and hands-on experience videos, to help use the tools with ease. See http://ihdia.iiit.ac.in/ for additional information.

## REFERENCES


[1] 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*. Springer, 2017, pp. 387–395.

[2] C. Clausner, A. Antonacopoulos, T. Derrick, and S. Pletschacher, "Icdar2017 competition on recognition of early indian printed documents-reid2017," in *ICDAR*, vol. 1. IEEE, 2017, pp. 1411–1416.

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

[4] 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*. IEEE, 2016, pp. 126–131.

[7] M. Wursch, R. Ingold, and M. Liwicki, "Divaservicesa restful web service for document image analysis methods," *Digital Scholarship in the Humanities*, vol. 32, no. 1, pp. 150–156, 2016.

[5] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia-an advanced document layout and text ground-truthing system for production environments," in *ICDAR*. IEEE, 2011, pp. 48–52.

[6] "Web aletheia." [Online]. Available: https://github.com/PRImA-Research-Lab/prima-gwt-lib

[8] S. Bukhari, A. Kadi, M. Ayman Jouneh, F. M. Mir, and A. Dengel, "anyocr: An open-source ocr system for historical archives," 11 2017, pp. 305–310.

[9] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/ vgg/software/via/, 2016.

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

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

[12] 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*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[14] A. Prusty, S. Aitha, A. Trivedi, and R. K. Sarvadevabhatla, "Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts," *ICDAR*, 2019.

[15] B. Lamiroy and D. Lopresti, "An open architecture for end-to-end document analysis benchmarking," in *2011 International Conference on Document Analysis and Recognition*, Sep. 2011, pp. 42–47.

[16] 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 *Proceedings of the 2013 27th Brazilian Symposium on Software Engineering*, ser. SBES '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 237–241. [Online]. Available: https://doi.org/10.1109/DAS.2014.23

[17] M. Wursch, R. Ingold, and M. Liwicki, "DivaServicesA RESTful web service for Document Image Analysis methods," *Digital Scholarship in the Humanities*, vol. 32, no. $\text{suppl}_1, pp.150 - -156, 112016$.