Automatic extraction of specimens from multi specimen herbaria

KENZO MILLEVILLE[*], Ghent University - imec, IDLab
KRISHNA KUMAR THIRUKOKARANAM CHANDRASEKAR[*], Ghent University - imec, IDLab
STEVEN VERSTOCKT, Ghent University - imec, IDLab

Since herbarium specimens are increasingly becoming digitised and accessible in online repositories, an important need has emerged to develop automated tools to process and enrich these collections to facilitate better access to the preserved archives. Particularly, automatic enrichment of multi specimen herbaria sheets pose unique challenges and problems that have not been adequately addressed. The complexity of localization of species in a page increases exponentially when multiple specimens are present in the same page. This already challenges the performance of models that work accurately with single specimens. Therefore in this work, we have performed experiments to identify the models that perform well for the plant specimen localization problem. The major bottleneck for performing such experiments was the lack of labelled data. We also address this problem, by proposing tools and algorithms to semi-automatically generate annotations for herbarium images. Based on our experiments, segmentation models perform much better than detection models for the task of plant localization. Our binary segmentation model can accurately extract specimens from the background and achieves an F1 score of 0.977. The ablation experiments for multi specimen instance segmentation show that our proposed augmentation method provides a 38% increase in performance (0.51 mAP@0.9 versus 0.37) on a dataset of 1500 plant instances.

CCS Concepts: • **Applied computing** → **Digital libraries and archives**; **Agriculture**; • **Computing methodologies** → **Image segmentation**; **Object detection**; **Object identification**.

Additional Key Words and Phrases: Herbarium, Data enrichment, Data augmentation

1 INTRODUCTION

Herbarium specimens have been collected globally for hundreds of years and serve as a rich source of crucial data for studying plant biodiversity and ecology. They provide users with documented occurrences of a plant in a specific location over time and therefore have been the basis of systematic botany for centuries. According to the latest data summary report released by Index Herbariorum, there are around 3400 active herbaria in the world comprising 397 million specimens that are protected across 182 countries [17]. Since the early 90s, libraries have conducted multiple digitization on a regular basis to ensure restoration and lasting preservation of herbarium collections. This protects them from further degradation caused by repetitive handling. Expanding access to specimen types across digitization is essential in maintaining specimens and making relevant knowledge readily available to researchers and the public. The sudden growth in high-quality image capturing devices induced by enormous amounts of collections (that are yet to be uncovered) has further led to a rising interest in large scale digitization initiatives across the world [2]. As herbarium specimens are increasingly becoming digitised

[*] Authors contributed equally to the paper

Authors' addresses: Kenzo Milleville, kenzo.milleville@ugent.be, Ghent University - imec, IDLab, Ghent, Belgium, 9000; Krishna Kumar Thirukokaranam Chandrasekar, krishnakumar.tc@ugent.be, Ghent University - imec, IDLab, Ghent, Belgium, 9000; Steven Verstockt, steven.verstockt@ugent.be, Ghent University - imec, IDLab, Ghent, Belgium, 9000.

XXXX-XXXX/2023/3-ART $15.00
https://doi.org/10.1145/3575862

and accessible on online repositories, there arises an important need to develop automated tools to process and enrich these collections in order to facilitate better access to the digitised archives.

This rising number of digitised herbarium sheets provides an opportunity to employ image processing techniques, such as deep learning, to automatically identify species and higher taxa [3–5] or to extract other useful information from the herbaria sheets, such as detecting handwritten text, colour bars, scales, and barcodes. The species identification task works well for herbarium sheets that have only one specimen on a page. However, there are many herbaria books that have multiple species on the same page (as shown in Figure 1) for which the complexity of the detection and localization of the plants increases tremendously. It also involves a great deal of time and effort if they are to be enriched manually. Therefore in this work, we propose a pipeline that can automatically localize, identify, and enrich plant specimens in multi-specimen herbaria as shown in Figure 2.

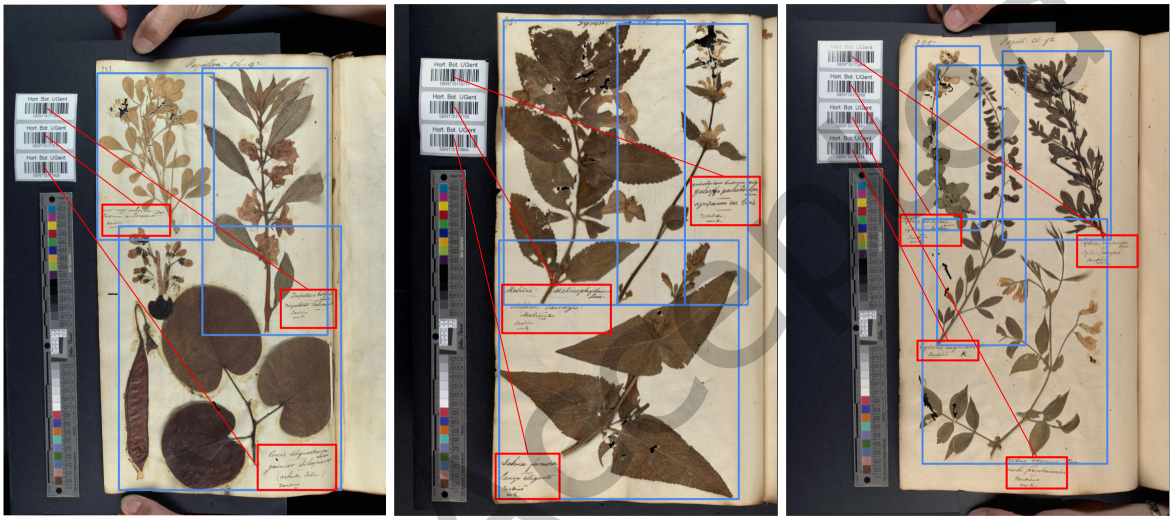


Fig. 1. Visual representation of multi specimen herbaria with three to four plant specimens on a single page. The text boxes around each plant provide more information about the plant. The number of bar codes denote the number of unique plant specimens in the image.

The proposed pipeline consists of three main steps; preprocessing of the images, extraction of the plants and associated labels, and linking the extracted plants to the plant database. The preprocessing step applies mainly to images coming from herbaria books, that are often warped due to the thickness of the book. The image may also contain additional elements besides the page of interest, such as bar codes and colour bars that need to be removed (see Figure 1). Next, the plants visible on the page need to be extracted separately. Finally, a text recognition model would be used to recognize the text on the page and link it to a database of taxa. The feasibility of the pipeline was experimentally validated. Based on the results of the feasibility study, three different methods were investigated for the localization of plant species, namely plant detection (one bounding box per plant), plant segmentation (one segmentation mask per page), and instance segmentation (one segmentation mask per plant). Each technique has its distinct advantages and drawbacks, which are discussed further in this paper. This work is part of an ongoing project and therefore the focus of this paper is limited to the extraction of the plant species.

In summary, the main contributions of this paper can be listed as follows:

(1) A modular pipeline is proposed that can automatically localize, detect and enrich plant specimens in multi specimen herbaria. The feasibility of the various blocks used in the pipeline is also investigated.

(2) Detection and segmentation models are discussed for the localization of plant species.
(3) A semi-automatic labelling algorithm is proposed to generate labels and masks that can be used to train detection and segmentation models. An alternate approach is also been discussed to compensate for the drawbacks of the algorithm.
(4) A new mosaic based augmentation technique is proposed for training segmentation models.
(5) Finally, we also provide the labelled data used for training and validating our models.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Subsequently, Section 3 presents the feasibility of the building blocks of the pipeline used for preprocessing, localization and identification of plant species. The dataset along with the algorithms used to generate it are presented in Section 4. Section 5 discusses experiments and their results while Section 6 concludes this paper and explains the future directions.

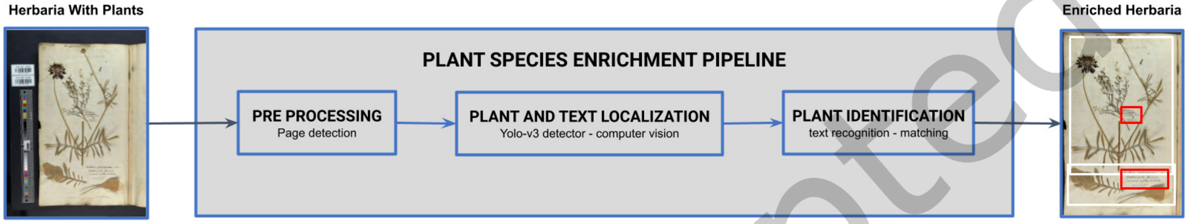

Herbaria With Plants
PLANT SPECIES ENRICHMENT PIPELINE
PRE PROCESSING
Page detection
PLANT AND TEXT LOCALIZATION
Yolo-v3 detector - computer vision
PLANT IDENTIFICATION
text recognition - matching
Enriched Herbaria

Fig. 2. Proposed pipeline for automatic enrichment of the herbaria books.

2 RELATED WORK

To automatically detect or segment herbarium objects, deep learning-based computer vision methods have become more and more popular over the last decade. Object detection models are one of the most straightforward methods to try, as these will output the location of the objects visible in the image. These locations are represented in the form of bounding boxes, which are (rotated) rectangles denoted by the 4 coordinate points of their vertices. To train such models for plant extraction, each plant on each page has to be annotated by its bounding box, which is a relatively straightforward process. A typical object detection network consists of object localisation and classification integrated into one convolutional network. There are two main types of meta-architectures available for this application: single-stage detectors like Single Shot Multibox Detectors (SSD) [12] and 'You only look once' (YOLO) [13] and two-stage, region-based CNN detectors, such as Faster R-CNN [14]. YOLO predicts bounding boxes of objects and their corresponding class probabilities in a single pass for the entire image. Faster R-CNN on the other hand is composed of three modules: 1) a deep CNN image feature extraction network, 2) a Region Proposal Network (RPN), used for detection of a predefined number of Regions of Interests (RoIs) where the object(s) of interest could reside within the image, followed by 3) Fast R-CNN [8] that computes a classification score along with class-specific bounding box regression for each of these regions. On the other hand, single-stage detectors use a single feed-forward network to predict object class probabilities along with bounding box coordinates on the image. Such popular pretrained object detection models generalize well to a new dataset on few amounts of labelled samples. In [21], the authors modified the YOLO architecture to identify plant specimens and other objects (rulers, colour bars, text blocks, and barcodes) within herbarium sheets. The results indicate that the proposed approach achieved good accuracy with mean average precision (mAP@0.5) of 0.932 compared to 0.901 of the original YOLO model. While object detection approaches require little resources and a low labelling effort, their outputs can be hard to interpret for irregularly shaped objects, such as plants.

Semantic segmentation approaches can be used to classify each image pixel as part of a specimen or the background. Such models take a binary image mask as the label for each object in the image and produce an output segmentation mask of the image. In this segmentation mask, each pixel is classified as an object class or as background. Traditional image processing techniques use colour clustering and contour based methods to differentiate between foreground and background objects. In this paper, we have proposed an algorithm based on such hand crafted features to generate specimen label masks. Although it works well for single specimen herbaria, generalizing them would be a hard task. Also since they are based on hand crafted features, the obtained masks are noisy and normally requires further cleaning. The U-net architecture [15] and its variants on the other hand are deep learning based approaches that have been widely used in the medical imaging domain and have become the standard architecture for (binary) segmentation problems in recent times. The U-net architecture consists of an encoder network that will downsample the image and extract features, which are then reconstructed in the symmetrical decoder network to produce segmentation masks. In [22], the authors retrained a U-net model on 400 images and masks of ferns. The model was able to successfully segment the ferns from the background, resulting in an F_1 score of 0.96, validated on 80 of the images. In [11], the authors employed a deep learning semantic segmentation approach based on the DeepLab-V3+ architecture and Full-Resolution Residual Networks (FRNN-A) to segment herbarium specimens from the background. They also achieved impressive segmentation results, with IoU scores of 0.992 and 0.981 for the FRNN-A and DeepLab models, respectively. While this semantic segmentation works really well in extracting specimens from the background, it does not differentiate between multiple specimens and will predict all the occurrences in a page as single specimen.

To both segment each specimen in the image and differentiate between them, an instance segmentation approach can be used. Instead of generating a single mask for all of the objects of the same class, these models generate a unique mask for each object detected. Mask R-CNN [9] is a widely-used state-of-the-art segmentation approach that extends the object detection system of Faster R-CNN. Apart from the 2 outputs of faster R-CNN (i.e. a class label and a bounding-box offset), there is an additional branch with a fully connected layer that outputs segmentation masks for each output proposal box. The design of Mask R-CNN consists of three main steps. First, the obtained feature maps are extracted from input images using the backbone network. The backbone is further expanded by using a Feature Pyramid Network (FPN) such that strong semantically correlated features are maintained at various resolution scales and orientations. The next step processes the feature maps using a fully convolutional network called Region Proposal Network (RPN). The final proposals of the RPN produce regions of interest (ROIs) from various pyramid feature levels. A Mask R-CNN approach was successfully used in [20] to segment and localize specimen leaves, colour bars, rulers, and text blocks. By segmenting the rulers on the images, which were each of a uniform size, they were able to automatically estimate morphological traits of the leaves. This approach resulted in an accurate estimation of these traits, achieving a relative error of 4.6%, and 5.7% for leaf lengths and widths, respectively.

3 FEASIBILITY TESTING OF THE PIPELINE

The idea of performing this feasibility experiment is to validate the proposed pipeline and check whether text and plant detections can be used for localizing and identifying plants in multi-specimen herbaria.

3.1 Data

The Botanical garden of Ghent University (Belgium) has made efforts to digitize herbarium books from three different and prominent Belgian botanists from the late 18th and early 19th century, Charles Van Hoorebeke, Aimé Mac Leod and Julius Mac Leod. There are 78 books of Charles Van Hoorebeke with 20 to 40 single-sided specimens per book. The books of Julius and Aime Mac Leod are a bit more complex with approximately 200 pages each and with multiple specimens per page (e.g. Figure 1). They are currently hosted within the virtual herbarium of Plantentuin Meise [1]. The goal of this paper is to develop methods that can automatically extract plant specimens from the books such as Julius and Aime Mac Leod that have multiple specimens in them.

3.1.1 Annotation

In order to perform feasibility experiments with multi specimen herbaria, a small random sample of 70 pages was chosen from our selected books. To annotate ground truth data, we used the VGG Image Annotator (VIA) [13]. Each object ranging from the plant specimen to colour bar, scale and text box were represented by a bounding box described by four coordinates: x, y, w, h. The coordinates (x, y) represent its left top corner while (w, h) represent width and height. Note that the bounding box is dedicated to annotating all objects within the input images including the plant specimen region.

3.2 Feasibility experiment: Preprocessing

Page detection is the process of finding pixels and regions in an image that constitute a page. Within the domain of historical digitization, page detection is predominantly applied for preprocessing of documents before handwritten text detection and recognition tasks, line and character detection, and segmentation of historical pictures.

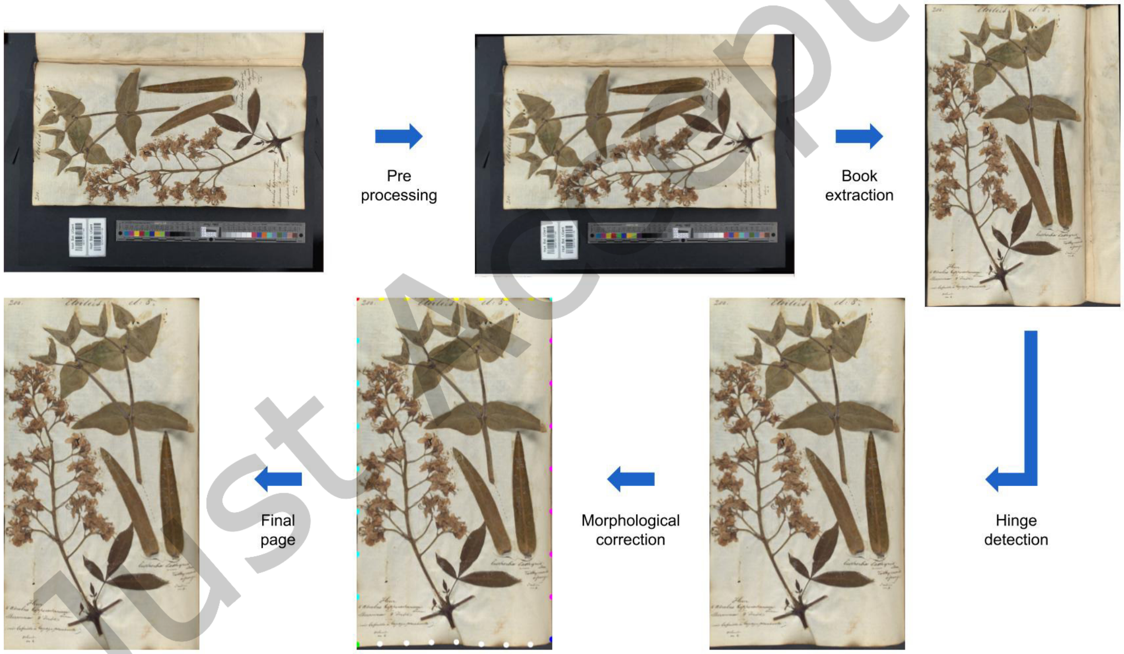


Fig. 3. The preprocessing of herbarium images

The page detection algorithm is an improved version of the algorithm proposed in [19]. The paper was addressed towards historical documents that used a hard threshold for extracting books from the background. Instead of using a hard threshold, we propose a colour clustering approach that can work with different types of books.

[1] Meise Virtual Herbarium : Mac Leod

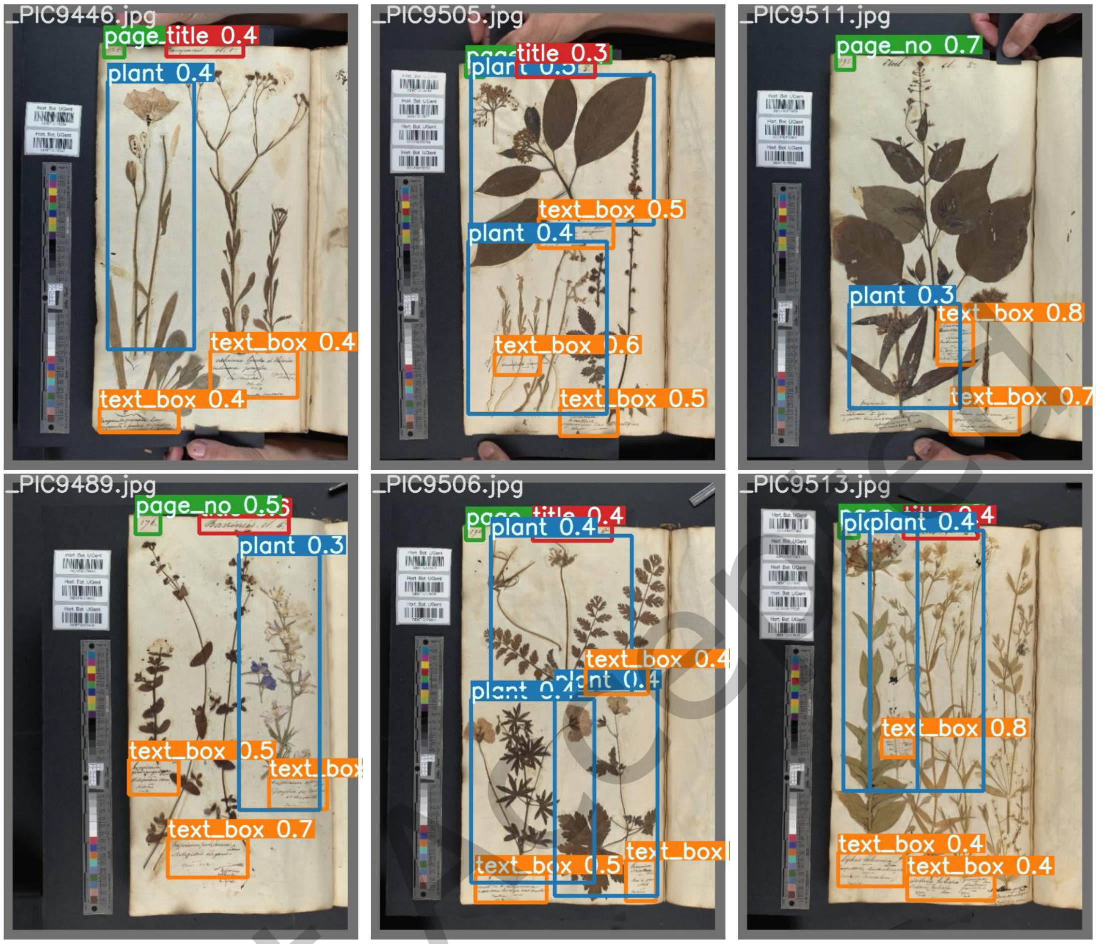


Fig. 4. Feasibility experiment: Results of the Yolov3 model trained on 50 labelled herbaria pages. The text regions are well detected whereas the plant boundaries require a lot more training examples to perform as expected.

As shown in Figure 3, the proposed pipeline for page detection begins with the preprocessing of images. The images are rotated and aligned such that the longest edge is maintained as its height. The core methodology of the proposed algorithm can be subdivided into two main steps namely book extraction and hinge region detection. The book extraction step filters background noise and extracts the main book region using k-means colour clustering. The k is chosen to be 3 and the idea behind it was primarily because the scans had three predominant clusters, namely (1) the black background on which the book is placed and imaged (2) the background of the page that is in white or brown and (3) the plant specimens that are in a different colour. The colour cluster pertaining to the largest contour area is chosen as the book. The hinge region detection step detects the hinges and extracts the main page region as explained in [19]. Finally, morphological transformations are performed by interpolating and transforming the page region such that the extracted page matches the expected page dimensions. This reduces deformations and occlusions caused by the thickness of the book. As explained in [19] the idea of detecting the pages is to perform morphological corrections in case of physical deformations. Therefore this step would not be necessary for individual herbaria sheets that does not have such physical deformations.

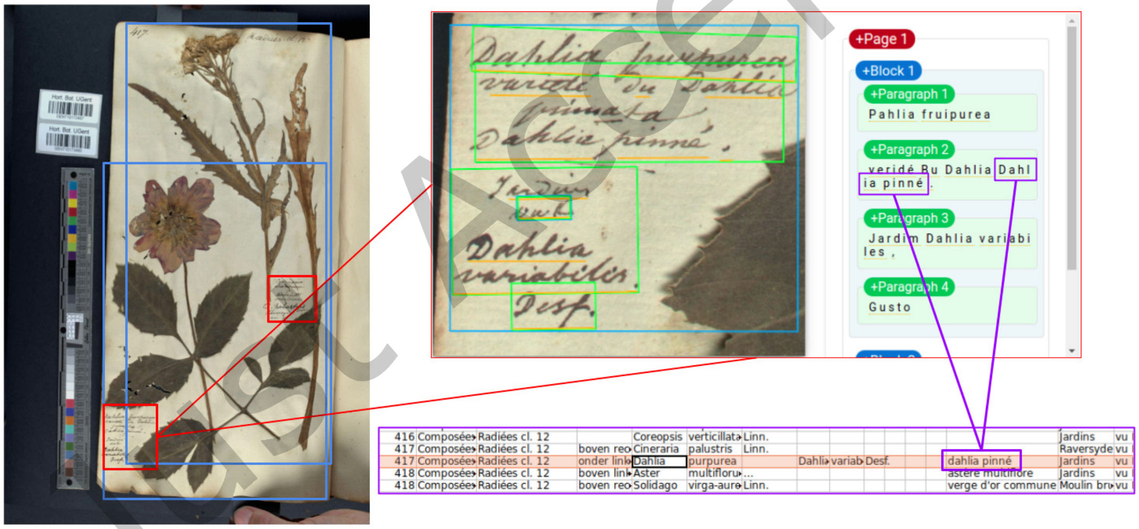
3.3 Feasibility experiment: Specimen localization

We performed an experiment with the small annotated dataset to see if it was feasible to annotate the entire book collection with bounding box ground truth labels. We trained a Yolov3 model for 100 epochs with 4 classes (Plant, Text, PageNumber and Title) with 50 images for training and 20 images for validating.

As seen in Figure 4, Text, PageNumber and Title classes had no problems and the model with the limited training data was already able to detect and localize them with decent accuracy. Plant specimens, however, had some trouble. Although the model was learning, with the amount of complexity, the model would require significantly more data to learn the plant boundaries correctly. As seen in Figure 1, the bounding box for the plant specimens overlap significantly, which would further confuse the model to learn the plant boundaries when augmentations are performed based on these annotations.

3.4 Feasibility experiment: Text recognition

An experiment was also performed to test the feasibility of text recognition to automatically recognise handwritten text from the detected text boxes. For this test, 50 random text boxes were cropped in the original resolution and processed with the Google Cloud Vision API[2] for text recognition. The results were subjectively evaluated to see if the botanical names of the plant specimens were recognised correctly.



Pahlia fruipurea

veridé Bu Dahlia Dahl ia pinné .

Jardim Dahlia variabi les ,

Gusto

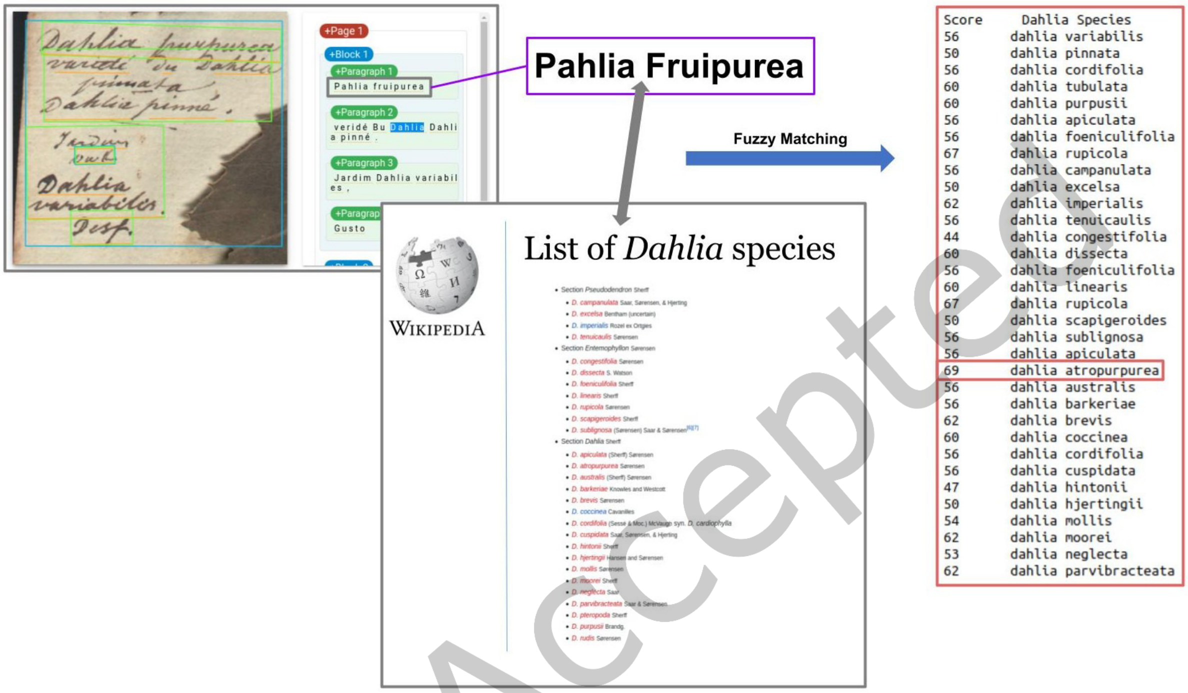
416	Composée▸	Radiées cl. 12		Coreopsis	verticillata▸	Linn.								Jardins	vu
417	Composée▸	Radiées cl. 12	boven reo▸	Cineraria	palustris	Linn.								Raversyde	vu
417	Composée▸	Radiées cl. 12	onder link▸	Dahlia	purpurea		Dahli▸	variab▸	Desf.				dahlia pinné	Jardins	vu
418	Composée▸	Radiées cl. 12	boven link▸	Aster	multifloru▸	...							astere multiflore	Jardins	vu
418	Composée▸	Radiées cl. 12	boven reo▸	Solidago	virga-aure▸	Linn.							verge d'or commune	Moulin bru▸	vu

Fig. 5. Sample text recognition results using Google Cloud Vision API. The plant specimen to the left belongs to the Dahlia flowering plant family. The text had *Dahlia variabilis* and *Dahlia Pinne* that was recognised by the API correctly.

An example of text recognition result is shown in Figure 5. As shown, the API could recognise some text correctly. There were also instances were most of the letters were predicted correctly (e.g Pahlia instead of Dahlia and Fruipurea instead of Purpurea). Based on our evaluation, about 24% of the words were recognised correctly while close to 36% had partial predictions that could be further improved. One option is to perform

[2] Google Vision API :Try it here

fuzzy matching on all the list of expected names from the dahlia family and find the one with the closest match as shown in Figure 6. There are still problems on how partial recognition would be chosen. However, this would not be further investigated in this work.



Pahlia fruipurea

veridé Bu Dahlia Dahlia pinné .

Jardim Dahlia variabiles ,

Gusto

List of *Dahlia* species

- Section *Pseudodendron* Sherff
  - *D. campanulata* Saar, Sørensen, & Hjerting
  - *D. excelsa* Bentham (uncertain)
  - *D. imperialis* Rozel ex Ortgies
  - *D. tenuicaulis* Sørensen
- Section *Entemophyllon* Sørensen
  - *D. congestifolia* Sørensen
  - *D. dissecta* S. Watson
  - *D. foeniculifolia* Sherff
  - *D. linearis* Sherff
  - *D. rupicola* Sørensen
  - *D. scapigeroides* Sherff
  - *D. sublignosa* (Sørensen) Saar & Sørensen[6][7]
- Section *Dahlia* Sherff
  - *D. apiculata* (Sherff) Sørensen
  - *D. atropurpurea* Sørensen
  - *D. australis* (Sherff) Sørensen
  - *D. barkeriae* Knowles and Westcott
  - *D. brevis* Sørensen
  - *D. coccinea* Cavanilles
  - *D. cordifolia* (Sessé & Moc.) McVaugh syn. *D. cardiophylla*
  - *D. cuspidata* Saar, Sørensen, & Hjerting
  - *D. hintonii* Sherff
  - *D. hjertingii* Hansen and Sørensen
  - *D. mollis* Sørensen
  - *D. moorei* Sherff
  - *D. neglecta* Saar
  - *D. parvibracteata* Saar & Sørensen
  - *D. pteropoda* Sherff
  - *D. purpusii* Brandg.
  - *D. rudis* Sørensen

```
Score     Dahlia Species
56       dahlia variabilis
50       dahlia pinnata
56       dahlia cordifolia
60       dahlia tubulata
60       dahlia purpusii
56       dahlia apiculata
56       dahlia foeniculifolia
67       dahlia rupicola
56       dahlia campanulata
50       dahlia excelsa
62       dahlia imperialis
56       dahlia tenuicaulis
44       dahlia congestifolia
60       dahlia dissecta
56       dahlia foeniculifolia
60       dahlia linearis
67       dahlia rupicola
50       dahlia scapigeroides
56       dahlia sublignosa
56       dahlia apiculata
69       dahlia atropurpurea
56       dahlia australis
56       dahlia barkeriae
62       dahlia brevis
60       dahlia coccinea
56       dahlia cordifolia
56       dahlia cuspidata
47       dahlia hintonii
50       dahlia hjertingii
54       dahlia mollis
62       dahlia moorei
53       dahlia neglecta
62       dahlia parvibracteata
```

Fig. 6. Sample fuzzy matching results using the text recognition result obtained from Google Cloud Vision API and list of species in *Dahlia* family obtained from wikipedia. As seen, although *Dahlia purpurea* was detected as *Pahlia Fruipurea*, fuzzy matching on all the species in dahlia family showed that the highest score was obtained by *Dahlia atropurpurea* which was the correct species.

3.5 Discussion

Based on the experiments it is evident that, it is possible to identify the plants based on the specimen and textual descriptions provided in the herbaria. However, for the proper functioning of the pipeline it is important for the current state of the art models to correctly localize specimens in a multi specimen herbaria. As identified in Section 3.3, the major bottlenecks for improving the localization of specimens include the lack of labelled data and selecting the right model. The rest of the paper is structured to address these problems.

4 DATASET CREATION, LABELLING AND AUGMENTATION

Detection models expect bounding box coordinates of the object as explained in Section 3.1.1. This is a lot easier to obtain owing to its rectangular structure. The segmentation model on the other hand requires polygons or masks of the objects to be learned. Since the plant specimens have an irregular structure it would be an extremely inaccurate and time-consuming job to manually label the plant specimen regions. The difficulty of labelling increases exponentially due to overlapping specimens in multi specimen herbaria. Additionally, we would also need a good number of labelled plant specimens for the model to learn the plant features effectively. Therefore we followed an alternative approach to generate plant specimen data.

The idea here is to use single specimen herbaria which is relatively easier to label in combination with augmentations that can introduce complexity. The single specimen herbarium scans provided by the New York Botanical Garden (NYBG) for the Kaggle Herbarium 2020 FGVC7 challenge was used for as the source of plant specimen data. Although the specimen labels in the herbaria sheets in the dataset were blurred for the contest, the plant specimens in the sheets were unaltered and therefore was suitable for our purpose. The test split of the challenge contains 100K images representing over 32,000 plant species from which we used a small random sample of 10k images.

To semi-automatically label the plant specimens in the scans, we propose an unsupervised algorithm to generate plant specimen masks. As detailed in Algorithm 1, the images were transformed to HSV colour space and were clustered based on the colours using k-means clustering with k=3. The largest contour at the centre was chosen and was processed to remove noises. This contour represented the mask of the plant specimen. Since, the algorithm also generated noisy masks, the generated masks were manually verified before choosing them for training the model. After the manual verification, masks and bounding box labels were obtained for 1875 images which were later used for training and evaluation.

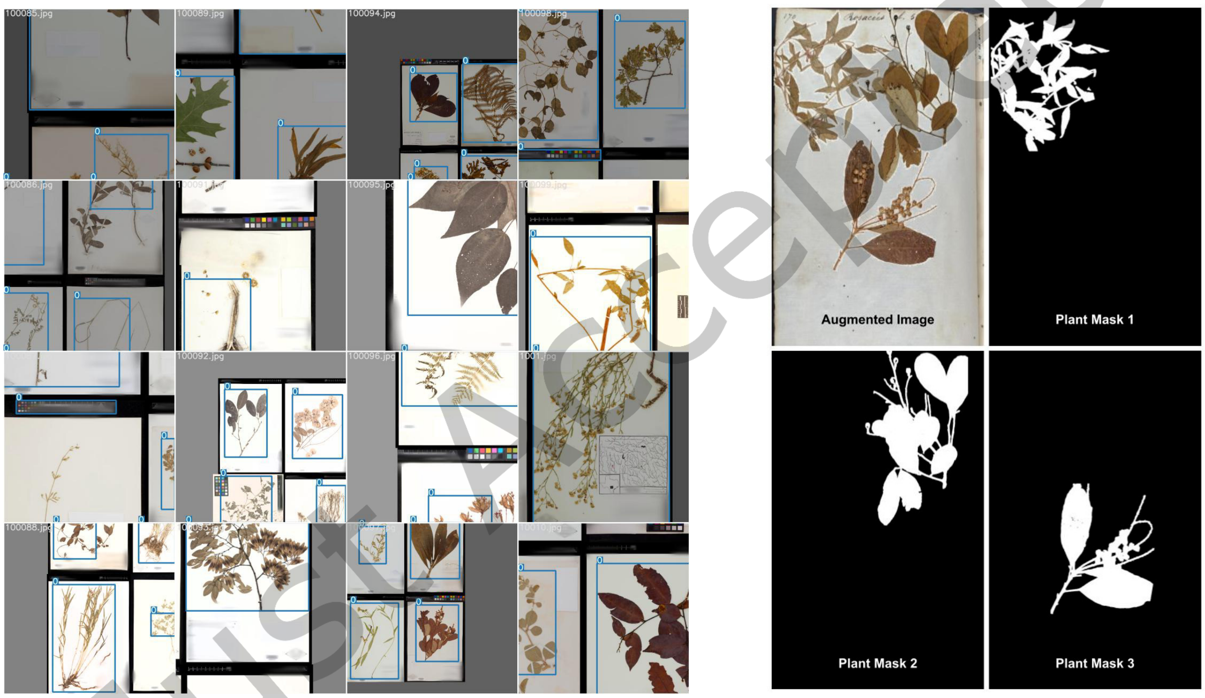
Algorithm 1 Mask generation and labelling

```
Require: Preprocessed image scan with single herbarium specimen
  Convert RGB to HSV
  Generate masks using k-means colour clustering.
  Specimen mask ← cluster that has the specimen
  Mask ← image mask with all zeros
  Improved specimen mask ← Dilation followed by erosion (filtering noises)
  Contour ← Find rectangle enclosing largest contour from Improved specimen mask
  Generate clean Mask using largest Contour
  Visualize Mask over input image
  if Mask is acceptable then
    USER → Press 1
    Generate best fitting rectangle bbox coordinates using largest Contour
    Write mask and bbox coordinates to separate files
  else
    USER → Press 0
  end if
```

During training and testing, the images were downscaled based on the task to reduce the computation time. The selected data set is divided into two sets: 80% was used as the training set, while 20% was used for the validation set. This is a normal data split used in most training methodologies. After completing the training, 375 images were used for validating the trained model's reliability. Note that the generated bounding masks were then used to determine the model's efficiency by measuring the cross entropy loss. Additionally, the trained model's consistency for the instance segmentation was validated by comparing the annotated mask images with the predicted mask results.

4.1 Augmentation

The single specimen sheets were primarily chosen to automatically label the plant specimens. However, for the models to identify plant specimens from complex scenarios, it requires to be trained accordingly. Therefore, the datasets were subjected to augmentations such that the models were trained on complex scenarios. For improving the training speed and complexity of the dataset, the mosaic augmentation [1] was used to train the object detection model. The basic idea behind mosaic augmentation is to combine 4 images for every input in a random scale and rotation. Example of an augmented image used for training the object detection model with a batch size of 16 is shown in Figure 7a. On the other hand, a modified form of mosaic augmentation inspired from [7] was used to train the instance segmentation models. Three plant specimens from three different images were randomly rotated and placed on the fourth image. The fourth image was one of the chosen blank pages from the herbarium database. With such an augmentation, 500 images with 1500 plant instances were created to train the instance segmentation model. An example of the modified mosaic augmentation is shown in Figure 7b.



(a) (b)

Fig. 7. The different types of data augmentation methods used during the training of the models. (a) Mosaic Augmentation used for training object detection model (batch size 16). (b) Modified version of mosaic augmentation used for training segmentation models.

5 EXPERIMENTS

The first experiment was carried out in an attempt to automate the labelling and creation of datasets for further study. Experiments were also performed to compare the performance of detection and segmentation models on the task of species localization. All the experiments were performed on a local Linux Intel(R) Core(TM) i5-7440HQ

CPU system with a RAM capacity of 64GB; the GPU was NVidia GeForce 980 with 12GB memory, and the operating system was Ubuntu version 18.04. The entire pipeline was implemented in python 3.6 and Pytorch deep learning library.

5.1 Evaluation metrics

For estimating the bounding box accuracy of the detector, mean Average Precision (mAP) and Intersection over union (IoU) metrics were used. IoU measures the percentage overlap between 2 bounding boxes. We use that to measure how much our predicted boundary overlaps with the ground truth (the labelled object boundary). The general definition for the Average Precision (AP) is finding the area under the precision-recall curve. The mean Average Precision or mAP score is calculated by taking the mean AP over all classes and/or overall IoU thresholds, depending on different detection challenges that exist. Our experiments focus on single class detection (plant specimen) and therefore mAP and AP are the same. We use two thresholds mAP groups namely mAP@[0.05:0.5] (written as mAP@0.5) and mAP@[0.5:0.90] (written as mAP@0.90) respectively, where mAP@0.90 means average mAP over different IoU thresholds, from 0.5 to 0.90 with a step of 0.05 (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9).

5.2 Automatic mask generation: Pixel-wise segmentation

As discussed in section 4, the automatic generation of the plant segmentation masks still required a quick manual verification step. This is mainly due to the colour clustering sometimes failing to segment the full specimen if part of the specimen was discoloured. Therefore, an image segmentation model was trained on the generated plant masks, to improve this validation step and generate a usable segmentation mask for images that failed. Two different U-net models were trained, one having a Resnet50 [10] backbone and the other an EfficientNetB0 [16] backbone. Both of the backbones were pretrained on Imagenet, allowing for a faster generalization to our dataset compared to training from scratch. The models were trained on a total of 1500 images and validated on 375 images.

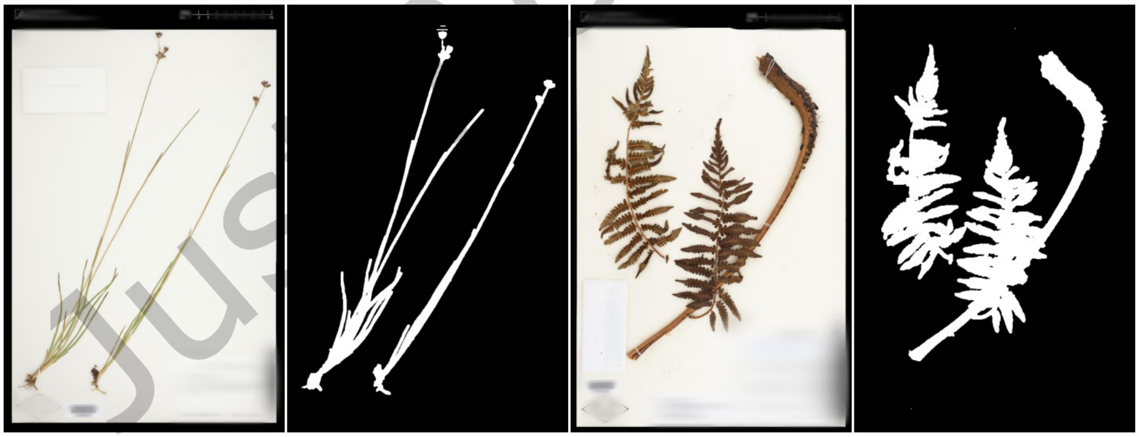


Fig. 8. Some sample outputs of the EfficientNetB0 segmentation model on images from the validation set. The predicted segmentation masks were thresholded with a value of 0.5.

Because most of the high-resolution images were too large to train on effectively, the models were trained on random crops of 256x256 pixels of the images and labels. However, there were a number of images that had a very small specimen region in them. In those cases, the random selection of crops had a higher chance

Model	IoU	F_1
Resnet50	0.951	0.974
EfficientNetB0	**0.956**	**0.977**

Table 1. Results of both image segmentation models on the validation set.

of selecting crops without any specimen (positive) pixel in them. Therefore the random selection was slightly modified to discard crops that only contained background (negative) pixels. Additionally, these image crops were augmented by randomly slightly modifying their colour values, as this generally improves the model performance and reduces overfitting [6]. At the end of each epoch, the validation loss was however calculated on the full validation images, which were resized to 800x606. This approach provides an alternative way to train the models on high-resolution images as compared to [22], where they resized the entire input images to 256x256 pixels. The models were trained for a maximum of 50 epochs, using a combination of the binary cross entropy loss function and Sørensen dice loss (F_1) function, where the model with the lowest validation loss was saved.

Table 1 lists the performance of both models on the validation set containing 375 images. The Resnet and EfficientNet models achieved F_1 scores of 0.974 and 0.977, respectively. For the EfficientNetB0 model, both the training time per epoch, epochs until convergence and best validation loss were lower than that of the Resnet50 model. This makes the EfficientNetB0 architecture the better candidate for a segmentation model backbone as it is more accurate with a faster training time. Figure 8 shows some example outputs from the EfficientNet model on the validation set. The model can correctly segment the specimens from the background. However, it does not make any distinction between individual specimens on a single herbarium image. If these multiple specimens are clearly separated (as is the case in Figure 8), they could be separately extracted with an additional postprocessing step. However, many herbarium images contain overlapping specimens, which are not trivial to separate in postprocessing.

5.3 Automatic specimen localization: Detection vs segmentation

To detect and localize plant specimens in multi specimen herbaria, we trained and compared both detection and segmentation models. Yolov3 was trained to detect bounding boxes of plant specimens in the herbaria. The main reason for choosing Yolov3 over the other discussed models was because it had a really good prediction accuracy and was much faster than the other mentioned models. The image size used for training was 416 X 416. The weights of the model (yolov3-416) pretrained on coco was initialized and used for training. The models were trained on a total of 1500 images and validated on 375 images.

On the other hand, Mask R-CNN was the model trained to generate segmentation masks of the plant specimens in the herbaria. The model from detectron2 - model zoo (mask_rcnn_R_50_FPN_400ep) pretrained on coco was initialized and used for training. The usage of pretrained weights speeds up the training process [22]. The models were trained on a total of 1500 instances and validated on 375 instances.

Type	Model	mAP@0.5	mAP@0.90
Detection	Yolov3	0.61	0.17
Segmentation	Mask R-CNN	**0.89**	**0.53**

Table 2. Results of object detection and segmentation on the validation set

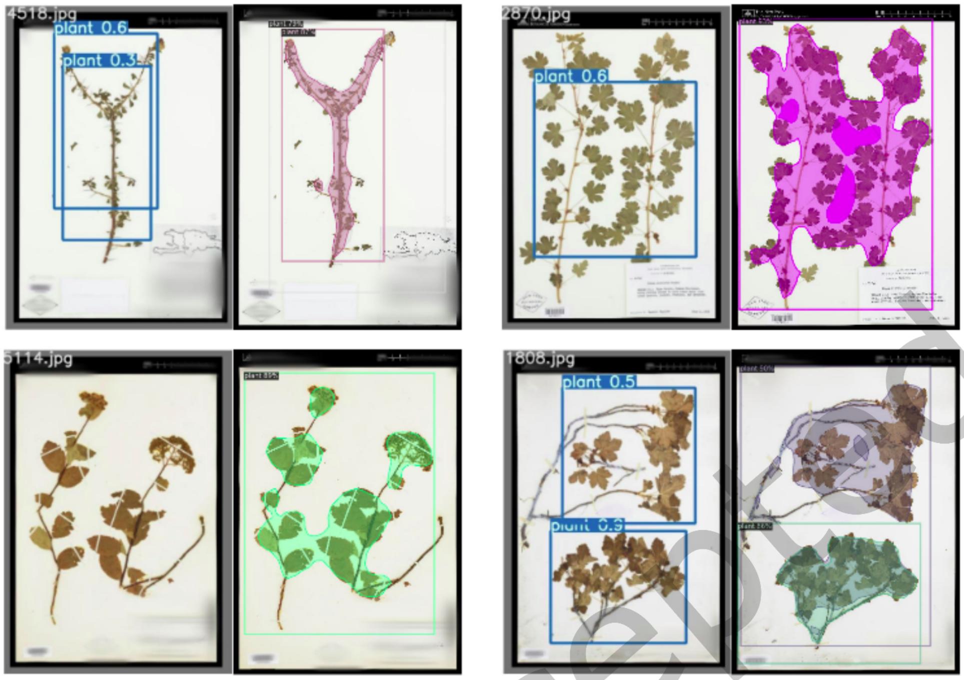


Fig. 9. Sample results for Yolov3 detection and Mask R-CNN segmentation model. Left images are results of the detection model while the right images are results of the segmentation model.

Table 1 lists performance of both models on the validation set containing 375 images for predicting the bounding box. The Yolov3 had a mAP@0.5 0f 0.61 but performed significantly worse for higher thresholds. The results are shown in Figure 9. It could be seen that the model fails to detect plants even in simple single specimen herbaria. Mask R-CNN on the other hand performed better with a mAP@0.5 of 0.89 and a mAP@0.9 of 0.53.

5.3.1 *Ablation experiment: Augmentation Mask R-CNN*

To study the effects of our proposed augmentation, we created a training experiment with and without augmentation. To have a fair comparison, the number of epochs were kept to 500 and the number of plant instances was limited to 1500.

Model	Augmentation	mAP@0.5	mAP@0.90
Mask R-CNN	No Augmentation	0.74	0.37
	Modified Mosaic	**0.85**	**0.51**

Table 3. Results of ablation experiment performed with and without data augmentation on Mask R-CNN model.The training was limited to 500 epochs.

Sample results are shown in Figure 10. As expected, the training with the augmented images learned the plant boundaries much faster than the non augmented images. After 500 epochs, the mAP@0.90 of augmented images

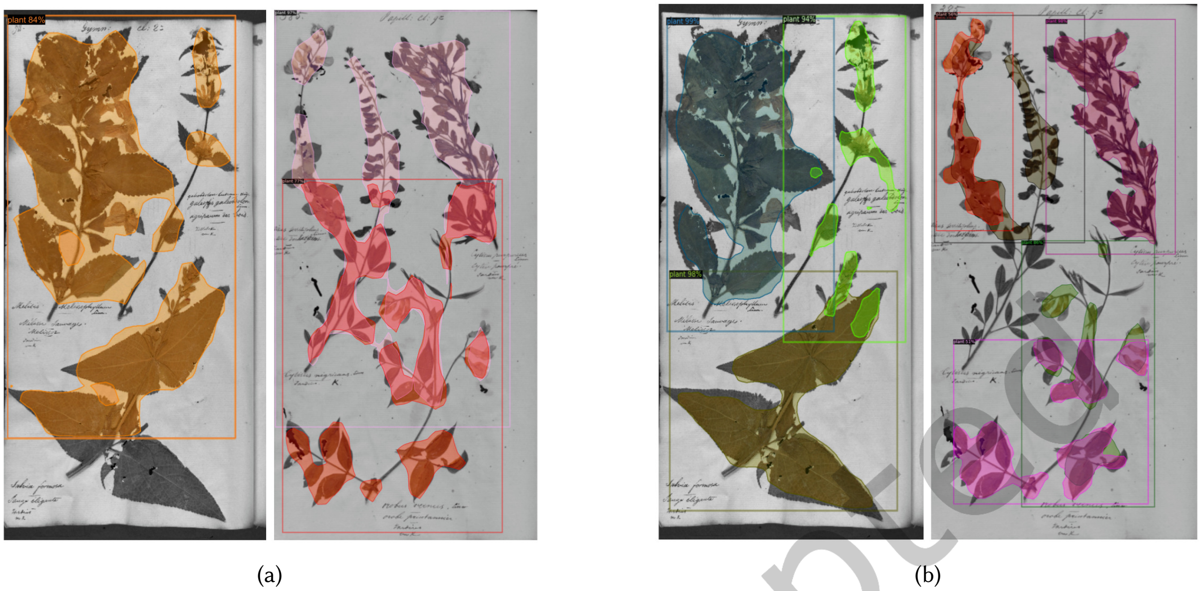


Fig. 10. Sample results of ablation experiments (a) Segmentation results for the model trained using non augmented images (b) Segmentation results for the same images for the model trained using augmented images.

was 0.51 which was 38% higher than the non augmented images, which had a mAP of 0.37 as shown in Table 3. With more augmented data and resources, the performance of the segmentation model could further be improved.

6 CONCLUSION AND FUTURE WORK

The paper has established the idea of having an automatic enrichment pipeline for multi specimen herbaria. The feasibility of different blocks has been tested and major bottlenecks have been identified. The scarcity of labelled data for the detection and segmentation of plant specimens has been addressed by proposing a semi-automatic labelling algorithm. The algorithm was further improved by using pixel-wise segmentation models such as U-net that further refined the specimen masks as explained in Section 5.2. The proposed augmentation technique has proved to be extremely beneficial for plant segmentation. It has contributed to a 38% increase in the overall performance of the segmentation models. Finally, it was observed that the instance segmentation models were better suited for the task of plant localization in complex multi specimen scenarios since they performed considerably better than the detection models.

Despite having promising results, we believe that this work merely marks the beginning of this direction. Accurate plant specimen masks are required for performing automatic trait extraction and therefore we would further work on fine-tuning the models to obtain better instance masks. Besides segmenting each entire specimen, different parts of each specimen, such as the leaves, flowers, and fruits could be extracted. However, such an approach will require labelled data with this additional information, which will involve more manual annotation. Apart from extending the framework to other collections, we want to extend the model towards extracting rulers, colour bars, and other objects around the plant specimen that would further improve the data enrichment quality. Finally, the methods explained in this paper would be further adapted to explore and automatically identify plant illustrations in books and paintings that can give rise to interesting cross domain applications such as [18].

ACKNOWLEDGMENTS

The research activities described in this paper were funded by The Department of Culture, Youth & Media, Flanders (Belgium) for the *Flore de Gand* project and the Research foundation-Flanders research infrastructure under the grant number FWO I001721N.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:cs.CV/2004.10934

[2] Gwenaël Le Bras, Marc Pignal, Marc L. Jeanson, Serge Muller, Cécile Aupic, Benoît Carré, Grégoire Flament, Myriam Gaudeul, Claudia Gonçalves, Vanessa R. Invernón, and et al. 2017. The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. *Scientific Data* 4, 1 (2017). https://doi.org/10.1038/sdata.2017.16

[3] Gwenaël Le Bras, Marc Pignal, Marc L. Jeanson, Serge Muller, Cécile Aupic, Benoît Carré, Grégoire Flament, Myriam Gaudeul, Claudia Gonçalves, Vanessa R. Invernón, and et al. 2017. The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. *Scientific Data* 4, 1 (2017). https://doi.org/10.1038/sdata.2017.16

[4] Jose Carranza-Rojas, Herve Goeau, Pierre Bonnet, Erick Mata-Montero, and Alexis Joly. 2017. Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology* 17, 1 (2017). https://doi.org/10.1186/s12862-017-1014-z

[5] Jose Carranza-Rojas, Alexis Joly, Hervé Goëau, Erick Mata-Montero, and Pierre Bonnet. 2018. Automated Identification of Herbarium Specimens at Different Taxonomic Levels. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics* (2018), 151–167. https://doi.org/10.1007/978-3-319-76445-0_9

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, 1597–1607. https://proceedings.mlr.press/v119/chen20j.html

[7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. 2021. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. arXiv:cs.CV/2012.07177

[8] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[11] Burhan Rashid Hussein, Owais Ahmed Malik, Wee-Hong Ong, and Johan Willem Frederik Slik. 2020. Semantic segmentation of herbarium specimens using deep learning techniques. In *Computational Science and Technology*. Springer, 321–330.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

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

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[16] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[17] Barbara M. Thiers. [n.d.]. A Summary Report Based on Data from Index Herbariorum. http://sweetgum.nybg.org/science/wp-content/uploads/2021/01/The_World_Herbaria_2020_7_Jan_2021.pdf

[18] Thirukokaranam Chandrasekar, Krishna Kumar and Deman, Emile and Verstockt, Steven. 2021. Cross-collection linking of botanical imagery in Ghent altarpiece to learn more about van Eyck's masterpiece and to explore a region's plant richness and diversity over time. *ACM JOURNAL ON COMPUTING AND CULTURAL HERITAGE* 14, 3, Article 40 (2021). http://dx.doi.org/10.1145/3457184

[19] Thirukokaranam Chandrasekar, Krishna Kumar and Verstockt, Steven. 2020. Page boundary extraction of bound historical herbaria. In *ICAART: PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON AGENTS AND ARTIFICIAL INTELLIGENCE, VOL 1*, Rocha, A. P. and Steels, L. and Van Den Herik, J. (Ed.). 476–483. http://dx.doi.org/10.5220/0009154104760483

[20] Abdelaziz Triki, Bassem Bouaziz, Jitendra Gaikwad, and Walid Mahdi. 2021. Deep leaf: Mask R-CNN based leaf detection and segmentation from digitized herbarium specimen images. *Pattern Recognition Letters* 150 (2021), 76–83.

[21] Abdelaziz Triki, Bassem Bouaziz, Walid Mahdi, and Jitendra Gaikwad. 2020. Objects Detection from Digitized Herbarium Specimen based on Improved YOLO V3.. In *VISIGRAPP (4: VISAPP)*. 523–529.

[22] Alexander E White, Rebecca B Dikow, Makinnon Baugh, Abigail Jenkins, and Paul B Frandsen. 2020. Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. *Applications in Plant Sciences* 8, 6 (2020), e11352.