FACULTY OF ENGINEERING
AND ARCHITECTURE

Meta Data Enrichment for Improving the Quality and Usability of Botanical Collections

Krishna Kumar Thirukokaranam Chandrasekar

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Computer Science Engineering

Supervisors
Prof. Steven Verstockt, PhD - Prof. Sofie Van Hoecke, PhD
Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University

July 2022

GHENT
UNIVERSITY

ISBN 978-94-6355-614-9
NUR 984
Wettelijk depot: D/2022/10.500/55

Members of the Examination Board

Chair

Prof. Filip De Turck, PhD, Ghent University

Other members entitled to vote

Prof. Dieter De Witte, PhD, Ghent University
Quentin Groom, PhD, Plantentuin Meise
Prof. Hiep Luong, PhD, Ghent University
Prof. Aleksandra Pizurica, PhD, Ghent University
Prof. Andreas Weber, PhD, Universiteit Twente, the Netherlands

Supervisors

Prof. Steven Verstockt, PhD, Ghent University
Prof. Sofie Van Hoecke, PhD, Ghent University

Acknowledgement

"தெய்வத்தான் ஆகா தெனினும் முயற்சிதன்
மெய்வருத்தக் கூலி தரும்"

This Thirukural from Tamil translates to *"Even if fate stands against you, with hard work, you can rewrite it"*. I believe, this is very true in my case and this dissertation stands as the reward for the hard work of so many people who have literally carried me up to this point. I had immense support and guidance from my supervisor prof. Steven Verstockt without whom this dissertation would not be possible. I am forever indebted not only to his input in terms of ideas and suggestions but also in terms of emotional support that has made me the person I am today. He made sure I bring out the best version of myself in what ever task I was carrying out. I would also like to thank my co-promoter Prof. Sofie van Hoecke and her team members Olivier, Tamir, Sander and Dieter with whom I have had weekly brainstorming and team building activities that made me settle in with the new environment in the beginning. Since the start of my PhD, I have been fortunate to share the office with the most kind and friendly colleagues - Florian, Samnang, Jelle (x2), Kenzo, Dilawar, Alec, Maarten, Dieter and Lore. I want to thank you all for making me feel part of the group and introducing me to domains I would not have explored otherwise. To specifically mention a few, Samnang made sure I got a roof for the first two weeks of my stay in Ghent. Thank you so much for the incredible gesture. Jelle V helped me with most of the dutch in the book while the discussion with Maarten was very helpful to fine tune my presentation for internal defence. Apart from the professional front, the constant banter and fun moments kept the atmosphere at the office lively. I would like to specifically thank other colleagues of IDlab, Aza, Hammami, Hannes, Pieter Colpaert, Pieter Heyvaert, Mattias, Davi, Niels, Glenn, Geoffroy, Julian, Anup, Van Der Donckt brothers and former colleagues Baptist, Ruben, Jasper, Vasileios, Johan and Brecht with whom I have had numerous eye opening lunch talks and snippet conversations regarding various concepts. I am thankful to the technical and administrative staff at IDLab: Kristof, Laura, Martine, Davinia, Bernadetta, Joeri and Mike to name a few. Their assistance has been crucial both to carry out my research as well as to complete administrative formalities during my stay in Ghent. Writing a multidisciplinary PhD is one part; defending it is quite another. Thank you prof. Aleksandra Pizurica, prof. Hiep Luong, prof. Dieter de Witte, prof. Andreas Weber and dr. Quentin Groom for their effort spent on grasping all the concepts covered in this work and for your excellent feedback and opportunity to further improve this text.

This acknowledgement is incomplete without mentioning the family of friends

who made my stay in Ghent feel like home. Ashok anna and Sharanya ka truly deserve the mention for taking care of us like their own. They have guided us through the most difficult times which makes me emotional when I think about it. Kartik and myself have had countless discussions on various topics that have been inspirational to my work in many ways. The late night dinners, fries walks during corona and marvel marathons relieved off our stress mutually. Arun na, Jacob na, Shiva na, Gowri na, Srinath na, Senthil na, Pothi na, Navneeth bro, Feni ka, Ilanagai ka, Devi ka and Dhanya ka made our life in Ghent so easy with frequent trips and get-togethers. How could I miss my fun banter moments with the kids Saindhu, Aura and others which are ever memorable. Dhanraj, Naveen, Gopal, Nithin, Joseph and all the other cricket friends made sure we had cricket sessions every week during summer.

My turning point in life was the months in KTH when I met Thaya, Deepa, Sharan and Bala. We had so many inspiring discussions on machine learning, that ignited the spark in me to do a PhD. My family-like-friends from India, Sathya, Aishu, Sai, Santy, Vijay, Adi, Sai Prasad and Sivagguru have constantly supported and motivated me during my tough times.

Of course, I have a long list of family members - Thatha pattis, mama mamis, peripa perimas, athai athimbers, chitti chitappas and cousins Sanju Aarthi, Guru Shivu, Rashu, Nidhu, Raju, Ramji Amrutha and Raghu Chandri who are there for me forever. I know that I was not always very communicative and sometimes even dismissive, but in the end, I could always rely on your unconditional support in times when it went good and less good. I could not miss to thank the friends who became close family - Kumar pa, Raji ma, Gowri and Ramya. A special mention to my sister Roshni, Ganapathy and my maruman “Thik Thik Thik” (How my daughter calls him). They are the ones who have missed me the most during this journey of mine. Hope I make it up to you guys somehow. Last but not least (the most important people always come last) - *Amma, Appa, Janani* and *Raksha* - you have done a remarkable job tolerating me. Without your support, sacrifice and encouragement, this dissertation will just be a dream. Maha periyava sharanam.

Ghent, July 2022
Krishna Kumar Thirukokaranam Chandrasekar

Summary

"There were 5 exabytes of information created between the dawn of civilisation through 2003, but that much information is now created every two days."

–Eric Schmidt, Executive Chairman at Google

This quote sums up the sheer volume of data that's available nowadays. Over the next years up to 2025, global data creation is projected to grow to more than 180 zettabytes. In 2020, the amount of data created and replicated reached a new high since more people worked and learned from home and used home entertainment options more often. However, are we able to use all off these data with the best of its ability to build a better world for tomorrow?

Prior to the launch of the United Nations Summit on Biodiversity, the insurance group Swiss Re revealed that *"over half (55%) of global GDP, equal to USD 41.7 trillion, is dependent on high-functioning biodiversity and ecosystem services"* . In other words, half of the world's wealth will be affected in one way or another if degradation of biodiversity continues. This is also in correlation with the alarming message the Global Assessment Report on Biodiversity and Ecosystem Services published by Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) carried. They call for a transformative change that every sector of the economy has to engage so that unprecedented degradation of biodiversity can be reduced. However, in order for the change to be successful, it is important to increase public appreciation, education and awareness of the value and importance of biodiversity and increase the public involvement in its conservation and sustainable use. This requires biodiversity data to be easily and publicly available.

Over the past twenty years, a remarkable global effort has been made to archive raw biodiversity data and make them accessible. A great number of open data platforms have emerged as a result. The portal of the Global Biodiversity Information Facility (GBIF) has undertaken to bring together the data resulting from these various initiatives. GBIF provides access to 1.6 billion occurrence records covering 1.6 million species names from 54,000 datasets worldwide. Its aim is to provide open access to data on all types of organisms on earth to everyone, everywhere. The raw biodiversity data published in the database can be freely searched, selected, and downloaded. It brings together multiple sources, ranging from museum specimens from the 18th and 19th centuries to geo-referenced photographs shared by

amateur naturalists using their smartphones.

Collection of plants is an ancient practice with records dating back to as far as 5000 years ago. Herbaria are collections of preserved plant samples and their associated data for scientific purposes. Herbaria documents the world's flora and provides a constant and permanent record of botanical diversity. The information on the specimens can serve as a basis for diverse scientific disciplines. This role is increasingly important since the rate of habitat destruction is constantly increasing and climate change precipitates rapid change in species' range and all aspects of their ecology. Distribution information about species can be evaluated along past periods. Herbarium data can give proof of the point of time when an invasive species occurred for the first time in a certain area. Since the early 1990s, libraries and botanical gardens have conducted multiple digitisation initiatives with herbaria books and scientific resources on a regular basis to ensure restoration and lasting preservation of such collections. However, with the advent of frameworks like GBIF to host a large amount of data and exponential growth of high quality image capturing devices induced by the enormous amount of rich herbaria collections (that are yet to be uncovered) has further led to a rising interest in herbaria processing in recent times.

Therefore an important need has also emerged to develop automated tools to process and enrich these collections to facilitate better access to these preserved archives. However, the digitisation of herbaria books introduces significant noise to the page such as presence of unimportant objects, imaged slightly rotated or under different viewing angle or with a differential lighting conditions that causes shadows. Also, the herbaria books are very thick, which induces deformation and warping of the page while digitising. Warping has an adverse effect on the text labels and shape of the specimens. This directly affects historical plant phenotyping and experiments involving learning the evolution of shape of leaves for which these digitised herbaria could be greatly useful. This dissertation addresses these problems by proposing a novel page detection algorithm that extracts the exact boundaries of the page. The obtained page polygon is then morphologically corrected such that deformations and warping are greatly reduced.

This rising number of digitised herbarium sheets also provide an opportunity to employ computer-based image processing techniques, such as deep learning, to automatically identify species and higher taxa or to extract other useful information from the herbaria sheets, such as detecting handwritten text and barcodes. Herbarium also contains metadata in the form of colour bars and scales that provide more insights to the size and colour profile of the digitised specimens. Although the process of detecting them is pretty straightforward, each consortium has their own scales and colour bars making it difficult to have a generic model. Therefore in this work we have addressed this problem by building a diverse dataset of scales and colour bars and trained a detection model on it. There has been a lot of research for the segmentation of herbarium specimens that has resulted in systems that work

well for herbarium sheets that have only one species in a page. However, there are numerous old herbaria books that contain multiple specimens in a single page and the problem is exponentially difficult when there are multiple specimens present in the same page. The lack of labelled data makes the problem even tougher. This work addresses this problem by proposing an augmentation technique that can automatically generate labelled multi specimen data from single specimen herbaria. The idea here is to utilise the state of models to generate specimen masks and create a complex augmented herbaria sheet with multiple single specimen masks. Both the labelled data and generated augmented data is made publicly available for future research.

Apart from these historical collections, global consortium such as inaturalist and plantnet have crowd sourced the collection and annotation of plant species on a large scale. This has not only exploded with enormous amounts of data, but has paved for training specific deep learning models that can identify plants. The state of the art models have learnt to recognise over hundred thousand types of plants. Technological advancement, in addition to the pandemic, has given rise to an explosive increase in the consumption and creation of multimedia content worldwide. This has motivated people to enrich and publish their content in a way that enhances the experience of the user. Such content also contains plants in them and are often found across different viewpoints. Such multi dimensional data of the same plant makes it interesting for plant phenotyping to better train and understand the plant structure. In order to identify and label such plants, this work also proposes a context based structure mining pipeline that can also re-identify plants based on the context similarity. It is proved that the pipeline can not only identify plants but can also be extended to re-identify other objects making it more suited for today's world.

Making the specimen and data online available and merging data from different institutions allows cross domain research and data analysis. It also paves way for more interesting use cases that improves the usability of enriched biodiversity data. Apart from researchers, common people are not motivated enough to learn out of enriched herbaria. The collections are often backed up and listed as a long ever ending list that almost makes the learning and searching process extremely tedious. A similar trait is also witnessed in art where most of the informative details contained in artwork is often overlooked. As an example, the 75 different plants that can be found in the Ghent Altarpiece is something not a lot of people are aware of. There are so many such artworks that have plants in them. Enriching such paintings with plant information opens up the possibility of linking plant collections with artworks making it possible to simultaneously explore both domains. In this work a cross collection linking pipeline is proposed and demonstrated using The Ghent Altarpiece painting. The demo was presented at the De Krook library in Ghent, wherein the public were allowed to interact with the painting and learn more about the plants in them. Apart from learning, I have also demonstrated a preferences based dynamic routing application that lets the users create their own

personal routes to witness the plants in the painting in their neighbourhood. This not only improves the usability of plant collections but also motivates people to learn more about their environment.

As someone once said, '***The journey of a thousand miles begins with one step***.' This work merely marks that first step towards that journey of creating biodiversity awareness and making it accessible for everyone.

Samenvatting
– Summary in Dutch –

Tegen 2025 zal de wereldwijde data creatie naar verwachting groeien tot meer dan 180 zettabyte. In 2020 bereikte de hoeveelheid gecreëerde en gerepliceerde gegevens zelfs een nieuw hoogtepunt omdat meer mensen thuis werkten en leerden, en home entertainment-opties aan populariteit wonnen.

Voorafgaand aan de lancering van de VN-top over biodiversiteit onthulde de verzekeringsgroep Swiss Re dat "meer dan de helft (55%) van het wereldwijde BBP, gelijk aan 41,7 biljoen dollar, afhankelijk is van goed functionerende biodiversiteit en ecosysteemdiensten". Met andere woorden, de helft van de welvaart in de wereld zal op de een of andere manier worden aangetast als de achteruitgang van de biodiversiteit voortduurt. Dit stemt ook overeen met de alarmerende boodschap die het Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) publiceerde in hun Global Assessment Report on Biodiversity and Ecosystem Services. Daarin roepen ze op tot een transformatie binnen elke sector van de economie om de ongekende achteruitgang van de biodiversiteit terug te dringen. Hiervoor is het echter belangrijk om de publieke bewustwording van de waarde en het belang van biodiversiteit te vergroten, en de publieke betrokkenheid bij het behoud en het duurzame gebruik ervan te verhogen. Dit vereist dat biodiversiteit gegevens gemakkelijk en openbaar beschikbaar zijn.

De afgelopen twintig jaar is er wereldwijd een grote inspanning geleverd om ruwe biodiversiteit gegevens te archiveren en toegankelijk te maken. Als gevolg hiervan is een groot aantal open data platformen ontstaan. Het portaal van de Global Biodiversity Information Facility (GBIF) heeft zich ertoe verbonden de gegevens van deze verschillende initiatieven samen te brengen. GBIF biedt toegang tot 1,6 miljard waarnemingen van 1,6 miljoen soorten uit 54 000 datasets wereldwijd. Hun doel is om voor iedereen en overal open toegang te bieden tot gegevens over alle soorten organismen op aarde. De ruwe biodiversiteit gegevens die in de database zijn gepubliceerd, kunnen vrij worden doorzocht, geselecteerd en gedownload. Het brengt meerdere bronnen samen, variërend van museum exemplaren uit de 18e en 19e eeuw tot recent geogerefereerde foto's die door amateurnatuuronderzoekers zijn gedeeld met hun smartphones.

Het verzamelen van planten is een oude praktijk met gegevens die teruggaan

tot 5000 jaar geleden. Herbaria zijn verzamelingen van geconserveerde plantenmonsters en de bijbehorende gegevens voor wetenschappelijke doeleinden. Ze documenteren de flora van de wereld en zorgen voor een constant en permanent verslag van botanische diversiteit. De informatie op de specimens kan als basis dienen voor diverse wetenschappelijke disciplines. Deze rol wordt steeds belangrijker naarmate meer habitats vernietigd worden en de klimaatverandering een grotere impact heeft op het leefgebied van soorten en alle aspecten van hun ecologie. Gegevens uit herbaria kunnen gebruikt worden bij het bestuderen van de verspreiding van soorten over een bepaalde periode, alsook het moment aantonen waarop een invasieve soort voor het eerst in een bepaald gebied voorkwam.

Sinds het begin van de jaren negentig hebben bibliotheken en botanische tuinen regelmatig digitalisering initiatieven uitgevoerd met herbaria boeken en wetenschappelijke bronnen om te zorgen voor restauratie en duurzaam behoud van dergelijke collecties. De komst van frameworks zoals GBIF, voor het toegankelijk maken van grote hoeveelheden gegevens, en de exponentiële groei van camerasystemen met hoge kwaliteit, veroorzaakt door de enorme hoeveelheid rijke herbaria collecties (die nog moeten worden ontdekt), heeft verder geleid tot de toenemende interesse in het verwerken van herbaria.

Door de toenemende digitalisering is er ook een belangrijke behoefte ontstaan voor geautomatiseerde tools om de collecties te verwerken en te verrijken met als doel de toegang tot deze archieven te vergemakkelijken. De digitalisering van herbaria boeken introduceert echter aanzienlijke uitdagingen door ruis (zoals onbelangrijke objecten), lichte rotaties, verschillende kijkhoeken of vermenging met een scène achtergrond. Ook zijn de herbaria boeken erg dik, wat leidt tot vervorming en kromtrekken van de pagina tijdens het digitaliseren. Het kromtrekken heeft een nadelig effect op de teksten en vorm van de preparaten. Dit heeft op zijn beurt een directe invloed op de historische fenotypering van planten, waarbij correct gedigitaliseerde herbaria nodig zijn om de evolutie van de vorm van bladeren te kunnen bestuderen. Dit proefschrift pakt deze problemen aan door een nieuw pagina detectie-algoritme voor te stellen dat de exacte grenzen van de pagina extraheert. De verkregen pagina wordt vervolgens morfologisch gecorrigeerd zodat het effect van vervormingen en kromtrekken sterk wordt gereduceerd.

Het stijgende aantal gedigitaliseerde herbariumbladen biedt ook de mogelijkheid om computergebaseerde beeldverwerkingstechnieken, zoals deep learning, toe te passen. Op basis hiervan kan men automatisch soorten en hogere taxa identificeren of andere nuttige informatie uit de herbariumbladen halen, zoals handgeschreven tekst en barcodes. Herbaria bevatten ook metadata in de vorm van kleurvlakken en schalen die meer inzicht geven over de grootte en het kleurprofiel van de gedigitaliseerde exemplaren. Hoewel het proces om ze te detecteren vrij eenvoudig is, heeft elk consortium zijn eigen schalen en kleurenbalken waardoor het moeilijk is om een generiek model te hebben. Daarom hebben we in dit werk dit probleem aangepakt door een diverse dataset van schalen en kleurenbalken te

bouwen en er een detectiemodel op te trainen.

Er is ook al veel onderzoek gedaan naar de segmentatie van herbarium specimens, wat heeft geresulteerd in systemen die goed werken voor herbariumvellen met slechts één soort per pagina. Er zijn echter talloze oude herbaria boeken die meerdere exemplaren op één pagina bevatten wat het probleem exponentieel moeilijker maakt. Het ontbreken van gelabelde gegevens bemoeilijkt dit onderzoeksgebied nog meer. Dit werk pakt dit probleem aan door een augmentatie techniek voor te stellen die automatisch gelabelde multi-specimen gegevens kan genereren uit single-specimen herbaria. Het idee hierbij is om state-of-the-art modellen te gebruiken om specimen maskers te genereren uit een single-specimen dataset en er een complex gevuld herbaria blad mee te creëren met meerdere specimen maskers. Zowel de gelabelde data als de gegenereerde geaugmenteerde data worden ook openbaar gemaakt voor toekomstig onderzoek.

Naast de historische plantcollecties hebben wereldwijde consortia zoals inaturalist en plantnet de verzameling en annotatie van plantensoorten op grote schaal ook gecrowdsourced. Dit heeft niet enkel gezorgd voor een explosie van de hoeveelheid beschikbare gegevens, maar heeft het ook mogelijk gemaakt om specifieke deep learning-modellen te trainen die planten kunnen identificeren in hedendaagse content/beelden. De state-of-the-art modellen hebben meer dan honderdduizend soorten planten leren herkennen.

De technologische vooruitgang heeft, samen met de pandemie, ook geleid tot een explosieve toename van het gebruik en de creatie van multimedia-inhoud wereldwijd. Dergelijke inhoud bevat ook planten die te zien zijn vanuit verschillende standpunten. Deze multidimensionale gegevens van dezelfde plant zijn interessant bij plantfenotypering om de plantstructuur beter te begrijpen. Om dergelijke planten te identificeren en te labelen, stelt dit werk ook een context- en structuurgebaseerde werkwijze voor die planten kan heridentificeren op basis van de contextovereenkomst. Het is aangetoond dat deze werkwijze niet alleen planten kan identificeren, maar ook kan worden uitgebreid om andere objecten opnieuw te identificeren, waardoor deze breder inzetbaar is.

Het online beschikbaar maken van de specimens met de bijhorende gegevens en het samenvoegen van gegevens van verschillende instellingen maakt domeinoverschrijdend onderzoek en analyse mogelijk. Het creëert ook nieuwe mogelijkheden voor interessante use cases die de bruikbaarheid van verrijkte biodiversiteitsgegevens verbeteren. Met uitzondering van onderzoekers zijn mensen doorgaans niet gemotiveerd genoeg om verrijkte herbaria te gaan exploreren. De collecties worden vaak vermeld in een lange lijst die het zoek- en leerproces vervelend maakt. Een soortgelijke situatie is ook te zien bij kunst waar veel informatieve details in het werk dikwijls over het hoofd gezien worden. Zo hebben veel mensen geen weet van de 75 verschillende planten die in Het Lam Gods terug te vinden zijn. Naast dit voorbeeld zijn er nog tal van andere kunstwerken waarin planten

worden afgebeeld. Het verrijken van dergelijke schilderijen met plantinformatie opent de mogelijkheid om plantencollecties te koppelen aan kunstwerken waardoor beide domeinen gelijktijdig verkend kunnen worden. In dit werk wordt een werkwijze voorgesteld en gedemonstreerd aan de hand van Het Lam Gods waarbij meerdere collecties met elkaar verbonden en geïntegreerd worden. De demo werd gepresenteerd op een Flore de Gand expo in De Krook, te Gent, waar het publiek met het schilderij kon interageren en meer te weten kon komen over de aanwezige planten. Daarnaast werd ook een dynamische routeringstoepassing gecreëerd waarmee gebruikers hun eigen persoonlijke routes kunnen maken om getuige te zijn van de planten in het schilderij in hun buurt. Dit verbetert niet alleen de toegang tot de plantencollecties, maar motiveert mensen ook om meer te weten te komen over de biodiversiteit in hun omgeving.

Iemand zei ooit: '***De reis van duizend mijl begint met één stap***'. Dit werk is slechts de eerste stap van de reis om meer bewustzijn over biodiversiteit te creëren en voor iedereen toegankelijk te maken.

Table of Contents

List of Figures

List of Tables

List of Acronyms

2D	2 Dimension
3D	3 Dimension
ACK	Acknowledgment
AI	Artificial Intelligence
API	Application Programming Interface
CNN	Convolution Neural Network
CPU	Central Processing Unit
CV	Computer Vision
eWVD	Woordenboek van de Vlaamse Dialecten
FGVC7	Seventh Workshop on Fine-Grained Visual Categorization
GBIF	Global Biodiversity Information Facility
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HD	High-Definition
HOG	History of Oriented Gradients
k-NN	k-Nearest Neighbors
LBP	Local Binary Patterns
LSTM	Long Short-Term Memory
LSU	Logical Story Unit
LSVRC	Large-Scale Visual Recognition Challenge
ML	Machine Learning
MSE	Mean Squared Error
mAP	Mean Average Precision
NLP	Natural Language Processing
NN	Neural Network
OSM	Open Street Map
PCA	Principal Component Analysis

POI	Point Of Interest
R-CNN	Recurrent Convolution Neural Network
Re-ID	Re-identification
ReLU	Rectified Linear Unit
RPN	Region Proposal Network
SGD	Stochastic Gradient Descent
SIFT	Scale-Invariant Filter Transform
SVM	Support Vector Machine
TPU	Tensor Processing Unit
VGG	Visual Geometry Group
YOLO	You Only Look Once
YOLOv3	You Only Look Once version3

Part I

Introduction

1 Introduction

This chapter begins by explaining the key problems / research questions that the dissertation attempts to address. The chapter then introduces key concepts that would be necessary to understand the dissertation. It continues by summarising the main contributions and outlining the structure of this dissertation. Finally, the chapter concludes by providing an overview of the publications that were authored during this research period.

⋆ ⋆ ⋆

1.1 Introduction

Plants are fundamental to life on earth constituting the base of the food chain, interacting with living beings and other components of ecosystems and contributing to the balance of these. Extensive technological advancement and focused research has resulted in the collection, documentation and presentation of data regarding a vast majority of plant species that are currently known to mankind. Various organisations and researchers have gathered around the world, funded and created frameworks and research infrastructure aimed at providing anyone, anywhere, open access to data about all types of life on Earth. This has resulted not only in large amount of structured open data that can be used freely but has also facilitated distributed data collection across different countries that can be consolidated. Yet, this abundant information doesn't reach the larger sector of the people

since the way this information is presented does not motivate people to search and learn more about the plants.

Artificial Intelligence, and especially deep learning technologies, hold tremendous promise for characterising and understanding plant biodiversity. These technologies are increasingly being applied to detect, identify, count, or measure key biological features including leaf shape and size, species identity and distribution, and reproductive characteristics. Increasingly, such investigations have led us to rethink how we can leverage digitised natural history collections and observations by applying convolutional neural networks (CNNs) to automate large scale phenotyping and assess biodiversity distributions.

Data documenting biological and ecological information on plant species are numerous and diverse. These data span specimens observed in culture, in the field, and across natural history collections. They contain rich information often linked with textual annotations. Automating the extraction and analysis of this information offers new opportunities to explore trait diversity, trait plasticity, and the spatio-temporal dynamics of species. However, a large number of methodological and algorithmic limitations need to be overcome to speed the exploration of these data and to conduct studies on unprecedented taxonomic, temporal, and spatial scales. Although promising results have been obtained from recent advancements [Stork, 2021], we are far from unleashing the full potential of these technologies.

In this dissertation, a number of novel algorithms are introduced to automatically characterise and scale plant biodiversity data. Additionally, demonstrators have also been built to foster a farther reach of these enriched collections. This chapter is intended to provide those who are unfamiliar with the subject matter with a short overview of a number of tasks of interest and with a description of the context in which this research took place. In addition, this chapter concludes with an outline of the dissertation and an overview of the research contributions made, as well as with a list of supportive publications.

1.2 Plants in digital medium

Traditionally based on the characteristics they possess, plants can be defined as groups of organisms that have photosynthesis, cell walls, spores, and a more or less sedentary behaviour. Therefore by this definition, plants can be anything between microscopic organisms such as “algae,” to macroscopic organisms that live on land. Digital Plants refers to the different forms of these plants that can be

found on the web or the digital medium. Predominantly, macroscopic plants in digital medium are broadly distributed across the three main categories based on environment in which the plants were present during imaging. They are

- Plants in preserved environment
- Plants in natural environment
- Plants in paintings

1.2.1 Plants in preserved environment: herbarium

Herbaria are repositories of preserved plant collections that are usually in the form of dried plant specimens mounted on a sheet of paper or book. The purpose of herbaria is both to physically contain the plant collections and to act as centres for research. The plant collections themselves function as vouchers for identification and as sources of material for systematic work. Herbaria also may house numerous geographic and taxonomic references, particularly floras or manuals that may aid in plant identification. Herbaria also normally contains the local name of a plant that directly correlates with the evolving dialects of the region.

In addition to housing plant collections, many herbaria today have initiated computerised data information systems to record and access the collection information of the plant specimens, as well as to access information from other collections worldwide. Information about herbaria is contained in Index Herbariorum, which lists the names, addresses, curators, and number and types of specimens. Each herbarium listed in Index Herbariorum is assigned with an acronym. It is this acronym that is cited in publications in order to specify where the specimens were deposited. Herbaria are typically associated with universities or colleges, botanic gardens, museums, or other research institutions.

A herbarium specimen consists of a pressed and dried plant sample that is permanently glued and/or strapped to a sheet of paper along with a documentation label. The herbarium paper is high quality, heavyweight, and acid-free to inhibit yellowing. Latest herbaria have the labels stuck at the bottom left corner of the sheet. However, herbaria from the past are in the form of books and would have labels in the form of handwritten text that can be present anywhere around the specimen. An example of herbaria book during digitisation is shown in Figure 1.1. Herbarium specimens will last for hundreds of years if maintained properly. They are still the most efficient and economical means of preserving a record of plant diversity. Herbarium specimens are increasingly becoming digitised and accessible in online repositories, an important need has emerged to develop automated

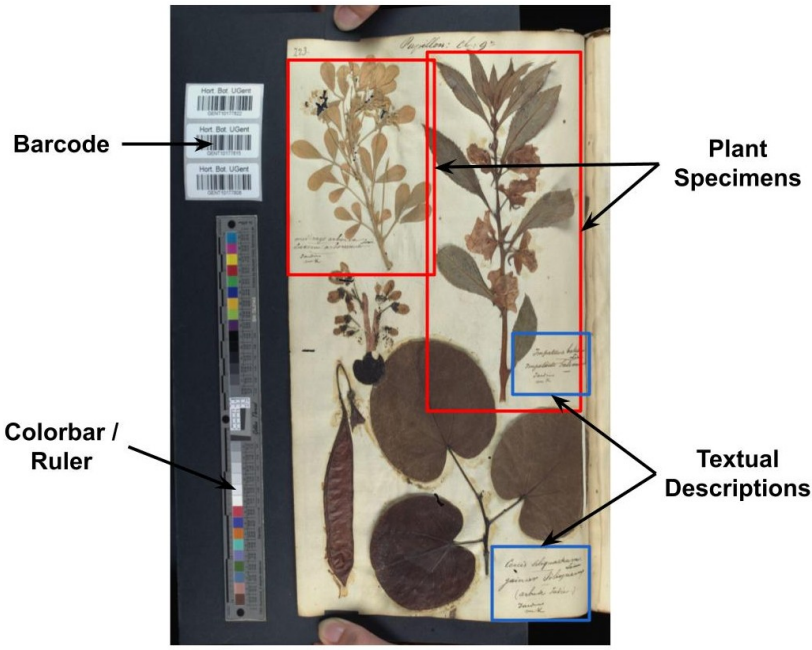


Figure 1.1: A sample page arrangement during the digitisation process of historical herbaria books. Along with the main herbaria page, bar codes, colour bars and scale are also placed to provide additional perspective regarding the size and true colour of the digitised image.

tools to process and enrich these collections to facilitate better access to the preserved archives. Particularly, automatic enrichment of multi specimen herbarium poses unique challenges and problems that have not been adequately addressed. The complexity of localisation 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. This dissertation focuses on the enrichment and extraction of specimens from multi specimen herbaria books. However there is even more to this problem.

When digitising bound historical collections such as herbaria it is important to extract the main page region so that it could be used for automated processing. The thickness of the herbaria books also gives rise to deformations during imaging which reduces the efficiency of automatic detection tasks. Therefore it makes it important to extract the exact boundary of the page. This dissertation also focuses on the extraction of the exact page boundary from the herbaria, so that deformations caused by the thickness of the book and imaging angle can be reduced.

The main datasets that were used for our experiments with the herbaria books were from three different and prominent Belgian botanist from the late eighteenth and early nineteenth century namely Charles Van Hoorebeke, Aimé Mac Leod and Julius Mac Leod. There are 78 books of Charles Van Hoorebeke with 20-40 single sided specimen per book. All these books possess a similar structure (e.g. Figure 2.11a). 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 1.1).

1.2.2 Plants in natural environment: images and videos

Extensive technological advancement and focused research has resulted in the collection, documentation and presentation of data regarding a vast majority of plant species that are currently known to mankind. Various organisations and researchers have gathered around the world, funded and created frameworks and research infrastructure aimed at providing anyone, anywhere, open access to data about all types of life on Earth. This has resulted not only in a large amount of structured open data that can be used freely but has also facilitated distributed data collection across different countries that can be consolidated.

Although such large initiatives collect plant images that have taken the identification of plants to an unimaginable level, there are cases where it still can be improved. For instance, there are tons of broadcast content and videos with recurring context that have so much plant information hidden in them. The identification of plants can be improved by re-identifying context and maximising the changing viewpoints.

This dissertation proposes a new methodology to re-identify plants in broadcast content by re-identifying recurring contexts.

1.2.3 Plants in paintings

It is interesting to see that artists have often depicted botanical imagery (i.e. plants, flowers, and trees of a region) as subjects or purposes in their paintings which ranges widely from devotional images of saints and scenes from the scriptures to portraits, still life and subjects from their time period. The use of flora in paintings proliferated especially in the fifteenth and sixteenth centuries, as artists became increasingly interested in the realistic depiction of objects from the natural world. Normally, plants / flowers have a symbolic meaning and associating it with a subject added two fold meaning to its natural decorative property. Research shows that people miss a lot of such informative details contained in a painting due to their lack of motivation, expertise and minimal time spent on viewing the artworks.

(a) (b)

*Figure 1.2: An example image of **Lilium candidum** / Madonna Lily (a) in natural environment and (b) in a painting of Van Eyck*

Paintings have plants in them and on the other hand we have abundant information about different types of plants that don't reach the larger sector of the people either.

This dissertation attempts to a find new direction that provides a foundation to explore art and horticulture simultaneously.

1.3 Computer vision and digital plants

Computer vision (CV) is a field of computer science that aims at enabling computers to see in the same way as human vision allows us to interpret the world. More precisely, this field is concerned with the theory and technology for building al-

gorithms that obtain information from images or multidimensional data [Szeliski, 2010]. A CV system, acting as a vision sensor and providing relevant information about the perceived visual input, is a basic building block for the majority of tasks explained further in this dissertation. Images are usually multi dimensional matrices composed of millions of pixels with associated colour information. It is in this matrix that algorithms attempt to find patterns so to for instance detect or recognise different specimens and objects. However, the information is too extensive and cluttered to be directly used by any learning algorithm. To understand the difficulties that arise when tackling this challenge, one should take into account not only the wide variety in which objects or plants specimens can be depicted, considering angle, viewpoint, and illumination, but also visual artefacts, such as distortions, occlusions, directional blurs, and cluttered backgrounds.

1.3.1 Feature vectors

Algorithms need to be able to deal with high dimensional inputs, given the typical resolutions used by visual data. The high dimensionality of these images is therefore reduced by computing feature vectors, i.e., a quantified representation of the image that contains the relevant information for the classification problem. During the last decade, research in this area mostly focused on the development of feature detection, extraction, and encoding methods for computing characteristic feature vectors. There are two broad categories by which feature vectors are computed. The traditional feature engineering methods involve manually designing and orchestrating feature extraction techniques such as binarisation, background subtraction, or contour detection customised for a specific application. Although such hand crafted features work really well, they are not generalised as they are problem-specific. One form of automatically generating feature vectors is by using generic algorithms such as scale-invariant feature transform (SIFT), speeded-up robust features (SURF), and histogram of gradients (HOG) that detects characteristic keypoints with their description. These descriptors capture visual information in a patch around each key point as orientation of gradients and have been successfully used for classification studies within this domain.

The next obvious step in automated feature extraction was removing an explicit decision about features to be described entirely. In the last decade, convolutional neural networks (CNNs) have seen a significant breakthrough in computer vision due to the availability of efficient and massively parallel computing on graphics processing units (GPUs) and the availability of large-scale image data necessary for training deep CNNs with millions of parameters [Barré et al., 2017]. In contrast to the above mentioned techniques, CNNs do not require explicit hand-crafted feature detection and extraction steps. Instead, they become part of the iterative train-

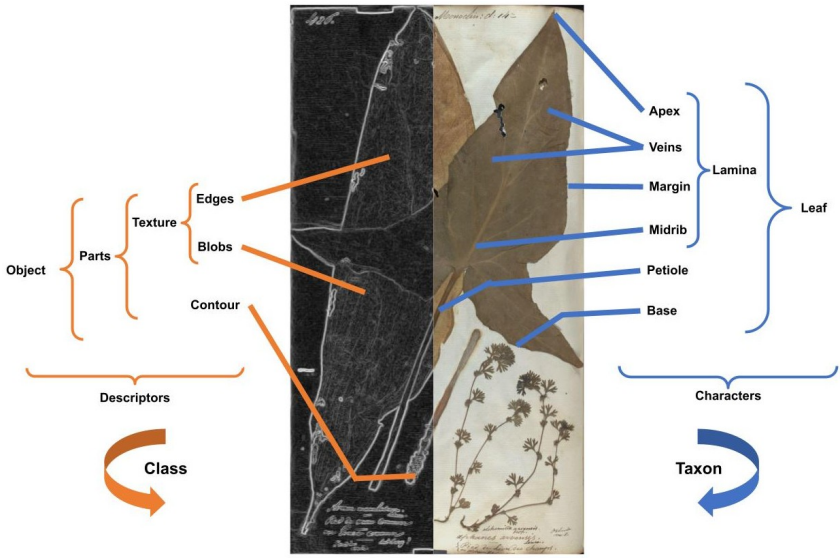


Figure 1.3: Computer vision (left) vs Botanists' (right) description of leaf specimens in a herbaria

ing process, which automatically discovers a suitable image representation (feature vector) that is learned. The fundamental concept of deep learning is a hierarchical image representation composed of building blocks with increasing complexity per layer. As shown in Figure1.3, the CNN learns these representations by aggregating smaller units at each level to form larger units that greatly increases the diversity of the resulting structure. Such hierarchical representations achieve classification and detection performances that were mostly unachievable using shallow learning methods with or without hand-crafted features [Wäldchen et al., 2018].

This dissertation deals with computer vision problems at various levels and therefore based on the requirement, the most appropriate feature extraction technique is made use to address the problem.

1.3.2 Classification, detection and segmentation

The features of the images are mainly extracted so that it could be utilised to extract high level information such as what is in the image and where. Within the domain of CV, these extractions tasks are broadly classified into image classification, object detection, and image segmentation. For the case of image classification, an algorithm is developed that is able to categorise an image based on a set of concepts, while for the case of object detection, an algorithm needs to recognise and

(a) (b)

Figure 1.4: An example output produced by (a) an object detection algorithm, and (b) an image segmentation algorithm

locate all predefined objects in a given visual input. One step beyond the detection of objects, an image or video sequence can also be segmented, which entails assigning every input pixel to one of several predefined categories. Figure 1.4 shows a conceptual example for each of the aforementioned tasks. These algorithms form the basis of more complex tasks such as detection of colour bars or segmentation of plants that are discussed in this dissertation.

1.4 Scientific contributions

This dissertation is composed of a number of publications that were realised within the scope of this PhD. The selected publications provide an integral and consistent overview of the work performed. The different research contributions are detailed in this section and the complete list of publications that resulted from this work is presented in Section 1.5.

Within this section I provide an overview of the remainder of this dissertation and explain how the different chapters are linked together. As shown in Figure 1.5, the dissertation addresses problems pertaining to digitisation and enrichment of

herbaria, linking the collections to other existing collections, plant re-identification in videos and creating the possibility to build new cross domain applications to further the reach of plant collections.

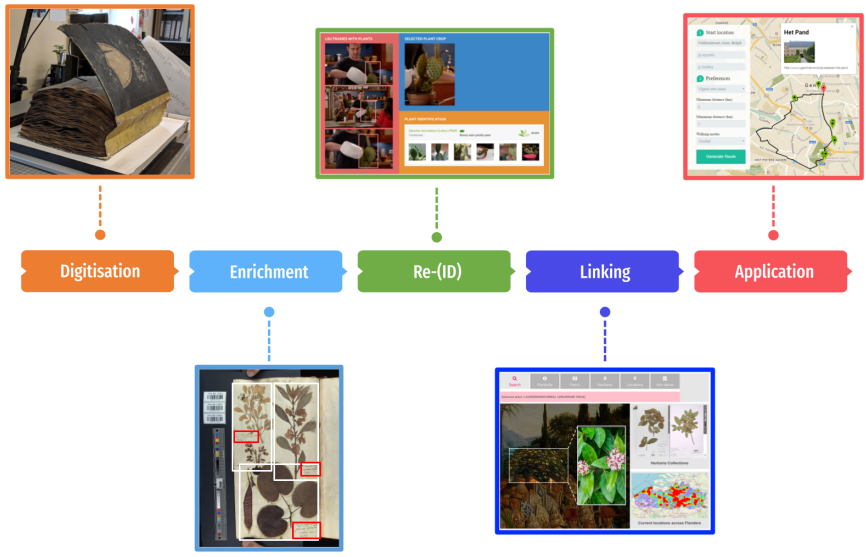


*Figure 1.5: Overview of the dissertation. The dissertation addresses problems relating to **digitisation** and **enrichment** of herbaria, plant **re-id**entification in videos, **linking** the enriched plants to other existing collections and creating the possibility to build new cross domain **applications** to further the reach of plant collections.*

CHAPTER 2: PAGE BOUNDARY DETECTION

When digitising bound historical collections such as herbaria, it is normal to include meta data information such as bar-codes and colour bars around the border region of the page during the imaging. Sometimes it also introduces significant variations in the page by inducing noise such as unimportant objects, slightly rotated, imaged under different viewing angles and perspective or blended with a scene background. Such noises interrupt the document analysis algorithms, that in turn affect the efficiency of the overall scanning procedure. The digitisation of books and bound herbaria also suffer from deformation and warping due to the thickness of the collections. Warping and deformation affects the original shape and texture of the specimens that directly affects automated detection tasks.

Chapter 2 focuses on a novel page detection algorithm, that extracts the page boundary of the herbaria books. The main challenge is to predict the hinge of

the page since the hinge is usually shared by the neighbouring pages and therefore just colour based filtering techniques are inadequate. Therefore an algorithm is proposed to accurately determine the hinge of the page. Finally It was also demonstrated that the detected page polygon could be used as a feature for reducing deformation.

CHAPTER 3: AUTOMATIC ENRICHMENT OF HERBARIA

The work in Chapter 2 acts a preprocessing step to the tasks explained in Chapter 3. Although herbaria books are digitised and the most informative regions of the page are extracted, it is important to know the contents of the page in order to further use the herbaria. Therefore, this chapter presents novel research efforts to detect and segment plant specimens from multi specimen herbaria. The species identification task works well for herbarium sheets that have only one species in a page. However, there are many herbaria books that have multiple species in the same page for which the complexity of the identification problem increases tremendously. It also involves a great deal of time and effort if they are to be enriched manually. Another major bottleneck for such tasks is the absence of labelled data.

This chapter therefore experiments and proposes models that perform well for the automatic detection of plants. The specimens from single specimen herbarium were semi-automatically labelled and a novel augmentation methodology was proposed that utilised the labelled single specimen data to create augmented complex multi specimen data. The augmentation methodology was thoroughly validated and was proved to be beneficial. Both the labelled data and generated augmented data was publicly released for further research.

CHAPTER 4: CONTEXT BASED RE-ID OF PLANTS IN VIDEOS

Automatic plant identification is helpful for the general audience in recognising plant species without the expertise of botanists. Today, these systems can identify thousands of unique varieties of plants spread across different continents. However, the identification of plants in broadcast videos is a lot more challenging due to factors such as differential lighting conditions and changing viewpoints. Also, processing every frame to detect and identify plants will be exponentially expensive. Therefore it makes it essential to come up with alternate strategies that is a lot more optimal.

This chapter proposes a novel pipeline that will not only split a video into shots and story units, but also detects and re-identifies plants at the same time. The different aspects of the pipeline were extensively evaluated.

CHAPTER 5: CROSS COLLECTION LINKING OF PLANTS IN PAINTINGS

Extensive technological advancement and focused research has resulted not only in the collection of large amount of structured open data that can be used freely but has also facilitated distributed data collection across different countries that can be consolidated. On the other hand people miss a lot of informative details contained in a painting due to their lack of motivation, expertise and minimal time spent on viewing the artworks. An interesting point to note is that there are so many famous paintings that has plants in them. Therefore a pipeline is proposed that enriches such paintings with plant information by cross linking plant data with other existing collections. The pipeline acts as a common bridging platform and brings together two completely different domains in such a way that it makes it possible to explore both the domains simultaneously. The pipeline is demonstrated using the centre panel of the Ghent Altarpiece painting that has 75 different plant species in it. The demonstrator lets the user interact with the painting and also provides additional information regarding the plants in them. The demo was presented at the De Krook library for a month and attracted a lot of audience.

CHAPTER 6: PREFERENCE BASED DYNAMIC ROUTING

Identifying the shared needs of the hugely diverse communities falling under the umbrella of humanities is not an easy task. Improving access to cultural and natural heritage data can be recognised as one such shared need. There are large amounts of geo-tagged plants data available that provides the opportunity to witness a plant live in a certain region. There are numerous routing applications that can be utilised to successfully create routes around these plant locations. However, creating a route around points-of-interest(POI) is still a very cumbersome process. To create a walking route around a neighbourhood, circular routes are preferred but such applications are not existent. Dynamically adapting the route preferences or generating a thematic route is also not present with many route generators. To address these concerns, this chapter presents the Runamic application that automates this tedious process and offers dynamic rerouting abilities. The chapter also demonstrates the possibility to dynamically route through plants in their neighbourhood based on plant POIs generated using the cross collection linking of plant collections with the painting.

CHAPTER 7: CONCLUSION & DIRECTION FOR FUTURE RESEARCH

In this last chapter, I conclude the dissertation by giving an overview of the contributions made and provide directions for future research.

1.5 Publications

The research results obtained during this PhD research have been published in scientific journals and presented at a series of international conferences. The following list provides an overview of the publications during my PhD research.

1.5.1 Publications in International Journals

1. **Thirukokaranam Chandrasekar Krishna Kumar** and Steven Verstockt. *Context-Based Structure Mining Methodology for Static Object Re-Identification in Broadcast Content.* Published in the Journal of Applied Sciences-Basel, vol. 11, no. 16, 2021, doi: 10.3390/app11167266.

2. **Thirukokaranam Chandrasekar Krishna Kumar**, Emile Deman and Steven Verstockt. *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.* Published in the ACM Journal On Computing and Cultural Heritage (JOCCH), vol. 14, no. 3, 2021, doi: 10.1145/3457184..

3. **Thirukokaranam Chandrasekar Krishna Kumar**, Kenzo Milleville and Steven Verstockt. *Automatic extraction of specimens from multi specimen herbaria.* Submitted to the ACM Journal On Computing and Cultural Heritage (JOCCH), Nov. 2021.

1.5.2 Publications in Proceedings of International Conferences

1. Kenzo Milleville, **Thirukokaranam Chandrasekar Krishna Kumar**, Thibault Blyau, Aurora Iannello, Umberto Michelucci and Steven Verstockt. *Extraction and Classification of Historical Stamp Cards using Computer Vision.* Published in ReMIX: Creation and alteration in DH, DH Benelux, Belval Campus, Esch-sur-Alzette, Luxembourg, May 9 2022, doi: https://doi.org/10.5281/zenodo.6529999

2. **Thirukokaranam Chandrasekar Krishna Kumar** and Steven Verstockt. *Page Boundary Extraction of Bound Historical Herbaria.* Published in proceedings of 12th International Conference on Agents and Artificial Intelligence (ICAART), VOL 1, Valletta, Malta, 2020, pp. 476–483, doi: 10.5220/0009154104760483.

3. **Thirukokaranam Chandrasekar Krishna Kumar**, Redouane Arroubai, Gerwin Dox, Samnang Nop, Pieter Stroobant, Jeroen Stragier, Kristof De Mey and Steven Verstockt. *RUNAMIC: Dynamic Generation of Personalized Running Routes.* Published in proceedings of 6th international congress

on sport sciences research and technology support (icSPORTS), VOL 1, Seville, Spain, pp. 98–105; SCITEPRESS (Science and and Technology Publications), doi: 10.5220/0006889600980105

4. Florian Vandecasteele, **Thirukokaranam Chandrasekar Krishna Kumar**, Kenzo Milleville and Steven Verstockt. *Video Summarization and Video Highlight Selection Tools to Facilitate Fire Incident Management.* Published in proceedings of the 16th ISCRAM Conference on Information Systems for Crisis Response And Management, Valencia, Spain, 2019, pp. 992–1001.

1.5.3 Publication of Abstracts

1. **Thirukokaranam Chandrasekar Krishna Kumar**, Kenzo Milleville and Steven Verstockt. *Species Detection and Segmentation of Multi-Specimen Historical Herbaria.* Published in Biodiversity Information Science and Standards, TDWG 2021, VOL. 5, 2021, doi:10.3897/biss.5.74060

Part II

Meta data enrichment of plant collections

2

Page extraction from bound historical herbaria

In this chapter, the pre-processing of herbaria books is explained. Since the books are bound historical books normal large scale scanning procedures could not be followed to digitise the books. Being very old, the books require special care and thus the individual pages of the book were manually imaged using a stationary camera set up as explained in the chapter. In order to obtain the high quality scan of the individual pages of the book, it makes it essential to detect and extract the informative regions of the book and page so that further processing could be performed. Additionally,since the book is bounded, the scanned pages would not always be flat and introduces a warping effect to the scanned pages. Therefore, this chapter proposes and evaluates a page detection algorithm that not only extracts the individual pages but can also dewarp the pages in case of warping of the pages.

⋆ ⋆ ⋆

This chapter is an adapted version of the following original publication:

Page boundary extraction of bound historical herbaria

Published at ICAART: Proceedings of the 12th International Conference on Agents and Artificial Intelligence, Vol 1, 2020

Abstract When digitising bound historical collections such as herbaria it is important to extract the main page region so that it could be used for automated processing. The thickness of the herbaria books also gives rise to deformations during imaging which reduces the efficiency of automatic detection tasks. In this work I address these problems by proposing an automatic page detection algorithm that estimates all the boundaries of the page and performs morphological corrections in order to reduce deformations. The algorithm extracts features from Hue, Saturation and Value transformations of an RGB image to detect the main page polygon. The algorithm was evaluated on multiple textual and herbaria type historical collections and obtains over 94% mean intersection over union on all these datasets. Additionally, the algorithm was also subjected to an ablation test to demonstrate the importance of morphological corrections.

2.1 Introduction

Since the early 1990s, libraries and museums have conducted multiple digitisation initiatives with cultural heritage documents and scientific resources on regular basis to ensure restoration and lasting preservation of historical collections. This is to protect them from further degradation caused by repetitive handling. Exponential growth in high quality image capturing devices induced by the enormous amount of rich historical collections (that are yet to be uncovered) has further led to a raising interest in historical document image analysis in recent times. Indeed, an important need has also emerged to develop automated tools to process and enrich these collections to facilitate better access to the preserved archives.

In addition to textual documents and records such as books, student registers or death records that are normally digitised on a large scale [Khan et al., 2018], there are various types of bound historical herbaria that preserve the rich horticulture of a region, that should also be digitised. In these herbaria, plants are collected, dried and stored in often difficult circumstances so that it could be used as a reference material decades later. It is important to preserve these records with technical proficiency but at the same time make it available to readers easily. Though modern scanners offer solutions for preserving these information, scanned materials are not always correctly oriented along the coordinate axes. In certain instances, the visual information is corrupted by external border noise. Sometimes it also introduces significant variations in the page by inducing noise such as unimportant objects, slightly rotated, imaged under different viewing angles and perspective or blended with a scene background. Such noise interrupt the document analysis algorithms, that in turn affect the efficiency of the overall scanning procedure. While removing background is feasible using simple segmentation techniques, other types of border noises are more challenging.

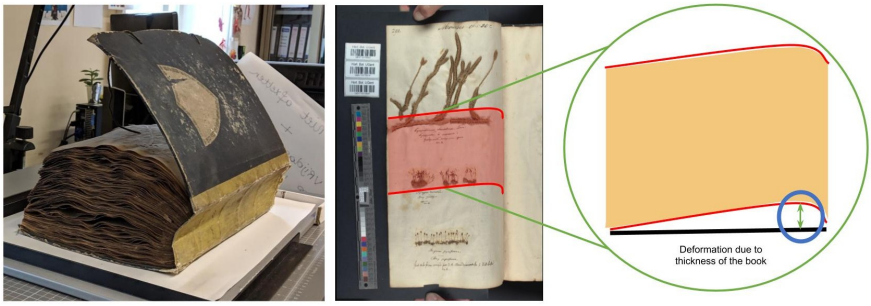


Figure 2.1: An example of a deformed page due to thickness of the herbaria book. The herbaria book and page in the example is of the Belgian botanist Julius Mac Leod.

The digitisation of books and bound herbaria also suffers from deformation and warping due to the thickness of the collections as shown in Figure 2.1. This poses a huge problem for automated detection tasks. For textual books, such warping effect decreases the text recognition process to a great extent [Mischke and Luther, 2005]. A similar problem persists for herbaria, wherein warping and deformation affects the original shape and texture of the leaf specimen. This directly affects historical plant phenotypeing and experiments involving learning the evolution of shape of leaves for which these digitised herbaria could be greatly useful. Therefore it makes it necessary to accurately detect the boundaries of the page, including the edge that is usually shared by the neighbouring page. This centre portion of the page that is usually caused due to binding is called $hinge$ in book anatomy based literature. The following terminology would be used in the rest of the chapter to refer that edge. The detection of hinge gets increasingly tougher for hand binded books since the edge would not be straight any more.

A number of approaches have been developed to remove border noise [Bukhari et al., 2012], [Chakraborty and Blumenstein, 2016b], [Fan et al., 2002], [Shafait and Breuel, 2010a], [Shafait and Breuel, 2010b]. However, as noted in [Chakraborty and Blumenstein, 2016a], most of the prior methodologies make fixed assumptions that holds good only for textual pages. Some of these assumptions are consistent text size, absolute location of border, straight text lines, and distances between page text and border. These assumptions don't hold good for herbaria type images (e.g. Figure 1.1) where such a consistent pattern is not followed. Therefore it is necessary to have an algorithm that makes use of only page and book based features such as colour, intensity, illumination and texture to detect the page boundaries.

In order to address the above concerns a multi step algorithm is proposed that detects variations in brightness and the distribution of the colour in an image to estimate the page boundary. The feature based algorithm is based on the HSV colour model since it provides additional intensity and colour depth features that can be utilised to better localise the page boundaries. A novel hinge detection algorithm is also proposed that can be used to specifically localise and extract the centre portion (hinges) of the book.

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

1. A page detection algorithm is proposed that can extract the exact page boundary of a page.

2. A traditional feature based hinge detection algorithm is proposed that can extract the centre region of the page.

3. The exact page boundary is utilised to reduce deformations that improves the text detection and localisation results.

The remainder of this chapter is organised as follows. Section 2.2 explains the algorithms in detail and reasons the approaches with examples. The data sets used are elaborated in Section 2.3. Section 2.4 discusses the initial results, while Section 2.5 concludes the chapter by proposing possible future work directions.

2.2 Methodology

Page detection is considered as the process of finding pixels and regions in an image that constitutes a page. Within the domain of digitising historical collections, page detection is predominantly applied for pre-processing of documents before hand written text detection and recognition tasks, line and character detection and segmentation of historical pictures.

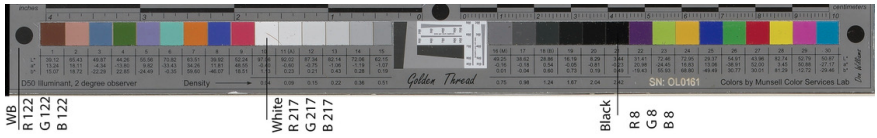


Figure 2.2: A sample colour bar used during the scanning of the herbaria. The values denote the RGB values for the white, black and white balance region of the image. These values are used to check the colour quality of the scan.

As shown in Figure 2.4, the proposed pipeline for page detection begins with the pre-processing 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 divided into two main steps namely book extraction and hinge region detection. The book extraction step filters background noise and extracts the main book region while the hinge region detection step detects the hinges and extracts the main page region. Finally, morphological transformations are performed on the extracted page in order to reduce deformations.

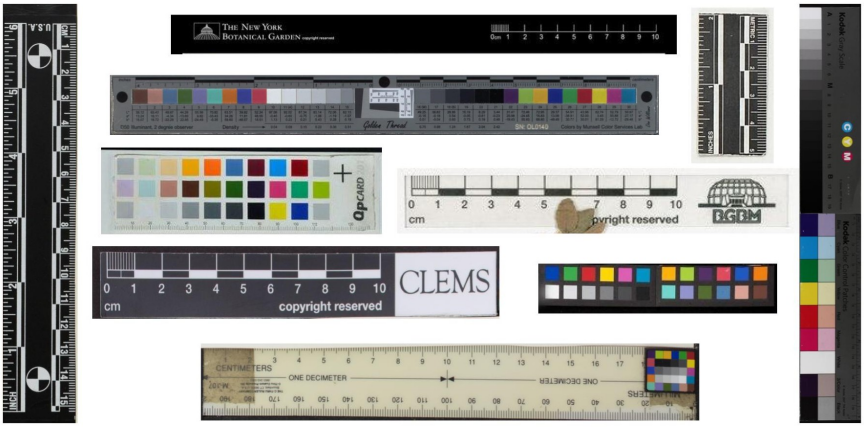


Figure 2.3: Different types of scales and colour bars used in herbaria collections

2.2.1 Pre-processing

The historical documents we have been investigating in the Flore de Gand and DISScO projects dates back to the nineteenth and the early twentieth century. Due to the factor of ageing, a large majority of these books possess a distinctive texture and colour which could be used for detection. However, due to environmental and ambient lighting conditions, the colour balance of the imaged page can be affected. Therefore it makes it essential to check the colour quality of the image. Normally, the document that is intended to be digitised is imaged with a standardised colour bar and scale in order to provide a original perspective to the document (as shown in Fig 2.2). Initially, traditional SIFT based template matching algorithm was opted to detect and localise the colour bar. However, on subsequent research with other herbaria collections, it was found that there were different types of colour bars and scales (as shown in Figure 2.3) and traditional template matching approaches were not robust enough to detect the colour bars automatically. Therefore, a yolov3 model was trained to detect the colour bars in the image. The colours in the detected colour bar are used to normalise the image to the expected white balance level. This is performed so that the image is maintained with the best possible colours that is closer to reality. Figure 2.2 depicts the RGB values

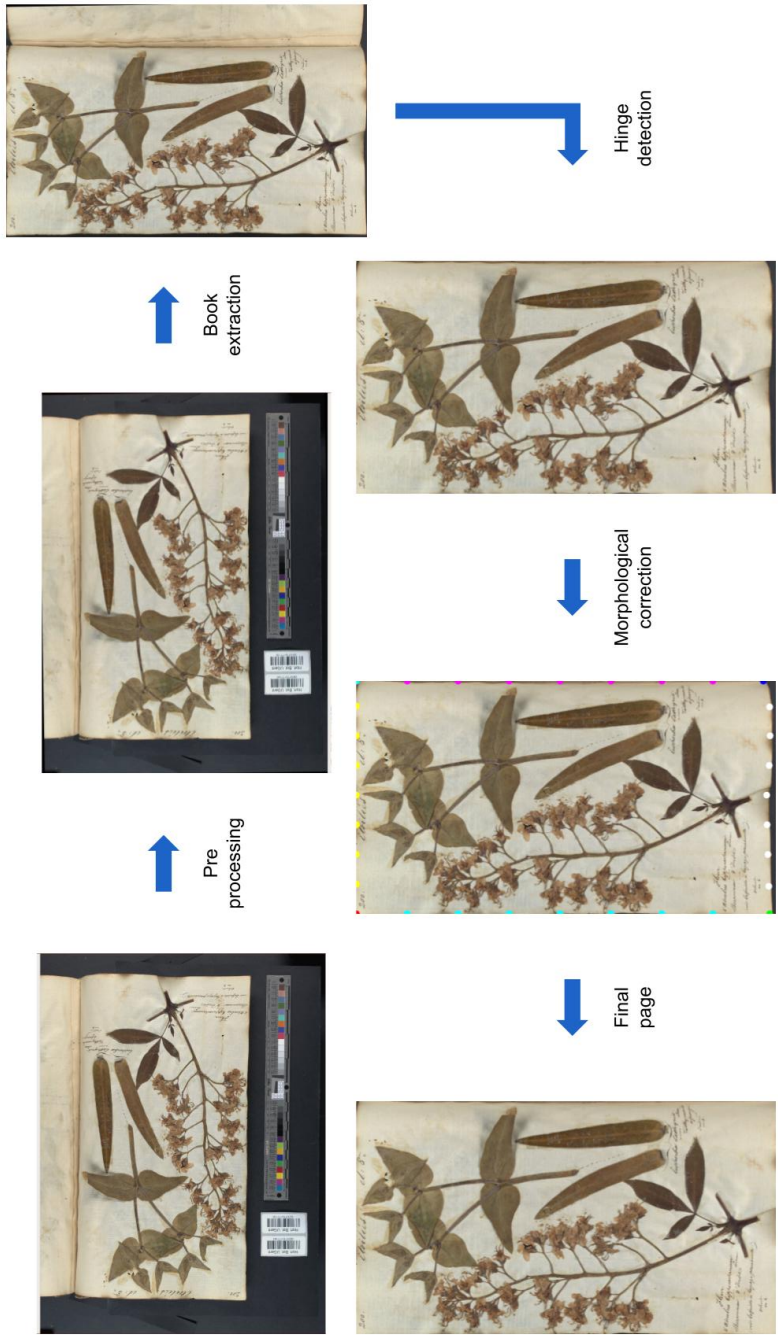


Figure 2.4: The proposed page detection pipeline

that were used to verify and normalise the image.

Additionally, it is also necessary to represent the image using a colour space [Sanchez-Cuevas et al., 2013, Albiol et al., 2001] in which some of the colour channels are invariant or at least insensitive to lighting changes, such as the H and S channel in HSV [James, 2013]. In HSV colour space, the Hue (H) and Saturation (S) values represent the colour information of the image. The Value (V) on the other hand is the measure of light intensity and denotes the extent of the colour's brightness of the image. An image can be easily converted from RGB to HSV by subjecting each pixel in the image to the transformations specified in [James, 2013].

2.2.2 Book extraction

The Saturation and Value components of HSV colour space have been adopted as the major features for background filtering and initial boundary prediction. Figure 2.5 provides a visual comparison of the original image with their Saturation and Value transformed images. The transformed images can directly be used for filtering and segmentation of the page to remove external and surrounding background noises.

Figure 2.5: The first image from the left is the original image and the following two images are its corresponding Saturation and Value transformed images. It can be seen that the hinge is more prominently perceivable from the saturation image (middle image) which is used to localise the hinge edge automatically.

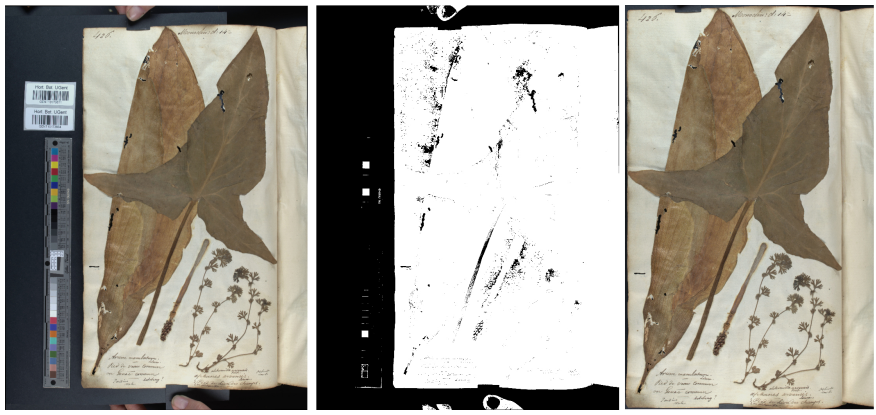


Figure 2.6: Book extraction with the corresponding segmentation mask generated based on HSV filtering. Based on the generated segmentation mask, the book is filtered and extracted as shown in the right most image.

Based on the average histogram of the Hue values for a random selection of 100 historic images from 6 different collections, it was found that more than 85% of the pixel's expected dominant colours lie between 5° and 65° of the Hue colour wheel. This seems to be logically acceptable since the dominant colours between the specified range are red and yellow and the majority of historic images has a high probability of having a slight yellow or brown tint due to the ageing factor of the document. The average histogram of Values shows a much more wider distribution. Yet, on further introspection of the average histograms, it has been witnessed that all the historic books taken into consideration, falls between the range of 30 - 93%. Based on the observed values for Hue and Value, book mask B_{mask} can be estimated for the image based on Equation 2.1.

$$B_{mask} = \left\{ \begin{array}{cc} 1, & 5° < H < 65° \\ & 30\% < V < 93\% \\ 0, & otherwise \end{array} \right\} \tag{2.1}$$

An example of the filtered mask obtained using HSV filtering is shown in Figure 2.6. Based on calculating the largest contour from the mask, the background and border noise can be eliminated and the book can be extracted by calculating the convex hull points of the contour [Goodrich et al., 2009].

2.2.3 Hinge region detection

The HSV based filtering and segmentation, results in the elimination of background noise, yet it is not sufficient enough to predict the hinge of the page. This

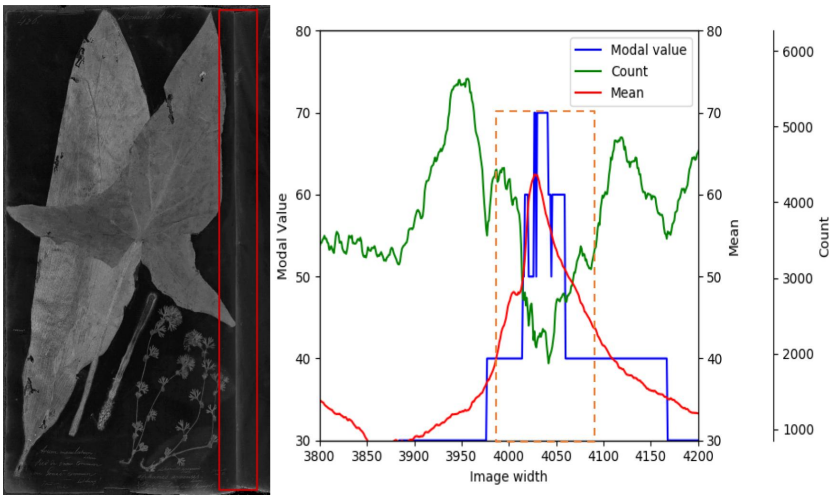


Figure 2.7: The plot shows the mean, mode and the mode count values for the columns of the image enclosed within the red bounding box. The orange dashed box shows the region where the actual boundary lies.

is because the hinge is usually shared by the neighbouring page and therefore just colour based techniques are inadequate.

Although the fourth boundary of a page in a book is shared by the neighbouring pages, it is still possible to locate this boundary because the area around the boundary tends to be darker. Logically, this is because the amount of light that could reach the hinge is gradually reduces due to binding. The Saturation value (S) signifies the amount of white light that needs to be mixed with the Hue. In the centre image of Figure 2.5 it is seen that the region around the hinge is brighter than the other parts of the page. This denotes that the Saturation values can be used for detecting the hinge.

The plot in Figure 2.7 depicts the correlation between the mean, mode and modal count values for a selected region around the boundary of the image. Since our focus here is predicting a longitudinal boundary, the mean and modal values are calculated for every column of the image saturation values. There is an evident Gaussian behaviour for the mean and modal values around the boundary region, which can be further justified by the plot in Figure 2.8.

The above mentioned phenomenon is made use of to localise and detect the hinges with near pixel level accuracy. To begin with, the peaks and valleys are estimated using a function that can calculate all local maxima based on simple

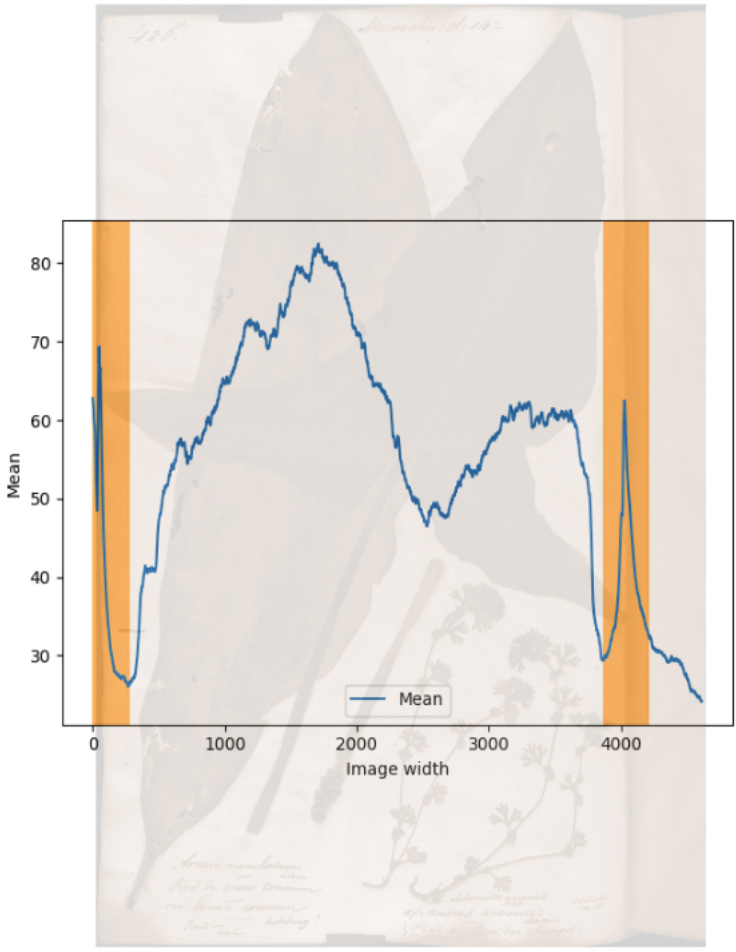


Figure 2.8: The plot shows the mean value of the columns for the image in the background. The orange strips corresponds to the region of the actual boundary.

comparison of neighbouring values [Virtanen et al., 2019]. The scipy implementation $find\_peaks$ has been utilised to detect the peaks and valleys. The prominence parameter had a greater effect on detecting the peaks and in our experiments $prominence$=10 was chosen as it performed the best for majority of images. Finally hinge regions are estimated based on the following conditions:

- Based on the peak and valley values of the mean, the peaks with the steepest slopes are selected.
- For the selected slopes, the mode and mode count values are cross verified. Based on the plot in Figure 2.7 it can be inferred that the modal count and

mean values are inversely correlated. Thus only those peaks that satisfy this condition are selected.

- Since the book is initially aligned in such a way that the hinge is always to the right side of the image, peaks present in the first half of the image are eliminated. It is also possible to automatically select the peaks based on the location of the contour but since the image is rotated and aligned during the pre-processing, this methodology is followed.
- From the remaining list of peaks, the peak with the highest prominence is chosen. The hinge region is derived as the region between the two local minima before and after the peak as indicated by the orange strip in Figure 2.8.

Note: Since the dimensions of the page do not change over a book, it is also possible to obtain them before the start of the digitisation process. In case if the dimensions of the page are already known, the peaks can be selected based on the width/height ratio of the book and the detected page in the image.

2.2.4 Page detection

Based on the obtained book mask and hinge region, the overall page is detected as follows:

- For the selected hinge region, the Saturation values (S) are applied a threshold and transformed into a mask based on the following criteria:

  Let h_r be the selected hinge region. Then

$$hs_{mean} = mean(S[h_r])$$
$$hs_{max} = max(S[h_r])$$
$$h_{mask} = \left\{ \begin{array}{ll} 0, & hs_{mean} < hs_r < hs_{max} \\ 1, & otherwise \end{array} \right\} \tag{2.2}$$

- The calculated h_{mask} is combined with the original book mask as follows:

$$B_{mask}[h_r] = h_{mask}$$

- The page region P is estimated by applying hull points to the largest contour [Goodrich et al., 2009]. The largest contour is calculated as follows.

$$P = max_{area}[B_{mask}]$$

The overall algorithm for page detection is elaborated in Algorithm 2.2.1.

Algorithm 2.2.1: PAGE DETECTION($image$)

Input: An image of the Book scan
Output: Final crop of the Page
$HSV \leftarrow RGB$

$$B_{mask} = \begin{Bmatrix} 1, & 5° < H < 65° \\ & 30\% < V < 93\% \\ 0, & otherwise \end{Bmatrix}$$

$Book = max_{area}[B_{mask}]$
if Book is not portrait
then {Rotate the mask by 90° and straighten the image
else {Straighten the image
$Mode, Count \leftarrow$ Mode of Book[axis=0]
$Peaks, valleys \leftarrow$ Find peaks and valleys for mean and count
Detect Hinge region(h_r) of the Book
if h_r

$$\textbf{then} \begin{cases} hs_{mean} = mean(S[h_r]) \\ hs_{max} = max(S[h_r]) \\ h_{mask} = \begin{Bmatrix} 0, & hs_{mean} < hs_r < hs_{max} \\ 1, & otherwise \end{Bmatrix} \\ B_{mask}[h_r] = h_{mask} \end{cases}$$

$P = max_{area}[B_{mask}]$
$P_{Dewarp} \leftarrow$Perform morphological correction on P

2.2.5 Morphological correction

The estimated page region is expected to be a rectangle or square for a normal book based on its dimensions. Yet, the obtained page suffers a mismatch due to a number of external factors. In order to compensate for this mismatch, the detected page has to be interpolated and transformed such that their shapes match. The morphological correction is performed as follows:

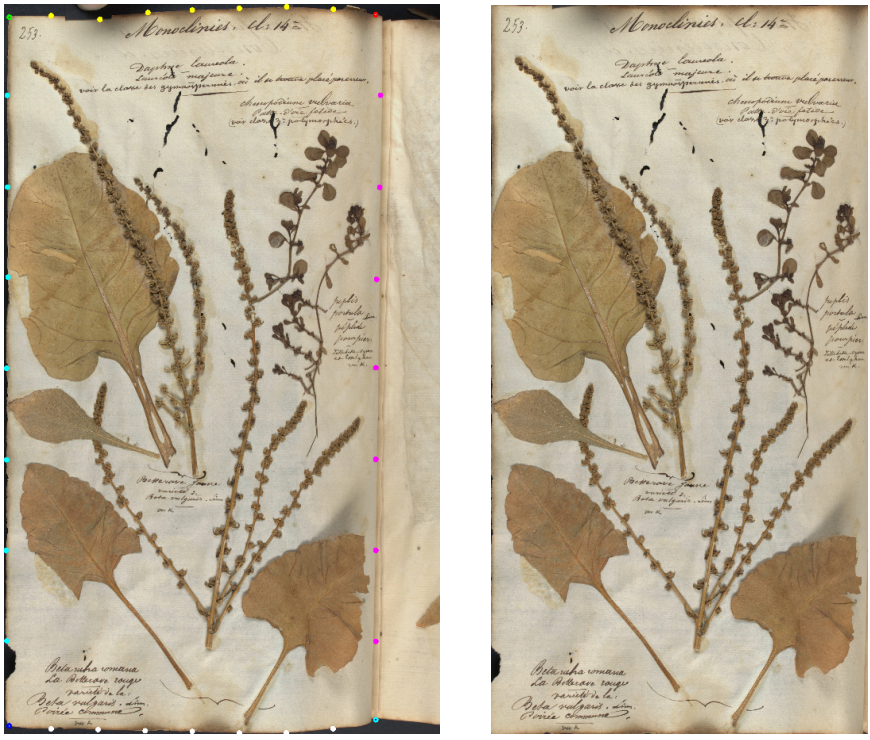
- For the estimated page region P, enclosing bounding box is calculated. The enclosing bounding box would be the reference bounding box R as shown in Figure 2.9. In case if the book dimensions are known, the dimensions are used to estimate the reference bounding box. The morphological correction is performed if the area of P is less than the area of R.

Figure 2.9: Steps involving morphological correction. Based on the page mask, equidistant hull points are chosen along the perimeter of the mask. However, for a flat page, the same points would be a rectangle (red points). Therefore, using the hull points, the entire image is remapped to the expected rectangle.

- In order to learn the shape representations, i equiv distant points are chosen along every side s of P and reference bounding box R. The value is dependent on the actual size and resolution of the image. In our algorithm, 12 points were chosen on each side. 12 was chosen since for i =12, the selected points were neither too close nor too far away resulting in better interpolation results.

- P is interpolated between P_{si} and R_{si} and the resulting co-efficients are re mapped on to R_{si}. Since the results of both cubic and linear interpolation looked identical default linear interpolation was preferred in our algorithm.

The implementation of morphological correction is demonstrated in Figure 2.10. Since a precise shape of the page boundary polygon could be obtained, this feature can be used for reducing the deformation of the page. Figure 2.10 showcases one such use case for flattening of the page to reduce deformation using simple morphological transformation. For this task, the numpy implementation $griddata$ was used to learn the current representation of the page and was transformed into the original representation (before deformation) using the opencv geometric image transformation function $remap$. Normally, text lines would be detected to reduce deformation [Mischke and Luther, 2005]. But, for herbaria type collections, where text patterns would be limited, it would be hard to use the text line features. In those scenarios, the page boundaries can be used as features to reduce deformation.



(a) (b)

Figure 2.10: An example of morphological correction for flattening the page in order to reduce deformation. (a) is the original image with chosen hull points for flattening and (b) is the subsequent flattened image.

2.3 Dataset and evaluation

2.3.1 Dataset

The page detection algorithms were initially developed and parameters were chosen based on randomly selected pages from the books of Julius mac Leod and Charles van Hoorebeke. The final algorithm was evaluated on the rest of the pages.

In order to evaluate the generalisation of the algorithm, the same algorithm was also evaluated on the ICDAR 2019 Competition Baseline Detection (cBAD) dataset [Diem Markus and Basilis, 2019]. The dataset consist of documents with varying levels of layout complexity extracted from 7 archives (bottom row in Figure 2.11). There were documents from two tracks and random subset was used from both the tracks.

(a) (b) (c)

(d) (e) (f) (g)

(h) (i) (j) (k)

Figure 2.11: Sample results of the page detection algorithm. (a) is a page from Charles Van Hoorebeke, (b) - (d) are from Mac leod herbaria. The rest are from cBAD baseline - dataset with different complexity obtained from 7 different archives.

2.3.2 Evaluation

Precision, recall and IoU (Intersection over Union) metrics are normally used to evaluate the efficiency of the boundary detection algorithms. These metrics should be used in order to compare performance of the algorithm with similar models. As proposed in [Tensmeyer et al., 2017] the page detection algorithm was also evaluated using manually labelled polygons.

Normally a page edge could be a few pixels thick and it is important to evaluate how much of the edge actually falls within this region. A pixel level comparison is therefore performed for the detected boundary to evaluate the closeness of the estimation with the ground truth. For this type of evaluation, the page boundary regions were manually annotated for 50 images of variable complexity as shown in Figure 2.11 bottom row. Finally, the test examples were also manually evaluated using two human participants.

2.4 Results and discussion

Table 2.1: Evaluation of the proposed algorithm on multiple datasets

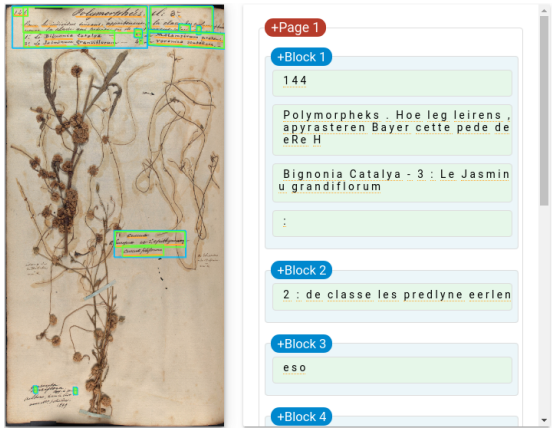
Dataset	Precision	Recall	IoU
Mac Leod	0.963	0.951	0.966
Charles Van Hoorebeke	0.935	0.948	0.94
cBAD	0.937	0.973	0.945

Table 2.2: mIoU performance comparison of our method with state-of-the-art (PageNet) and baseline (GrabCut) algorithms.

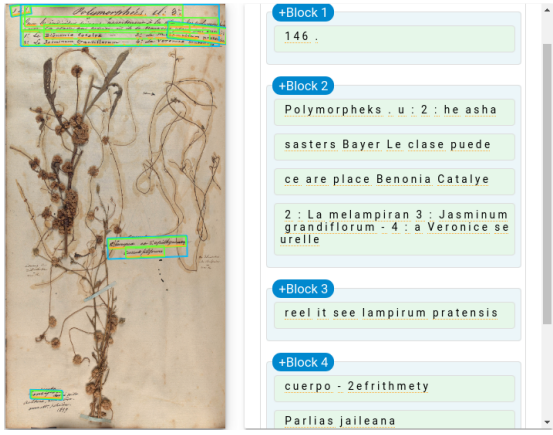
Dataset	PageNet	GrabCut	Ours
Mac Loed	0.962	0.893	0.966
Charles Van Hoorebeke	0.938	0.904	0.940
cBAD	0.947	0.864	0.945

Table 2.1 shows the precision, recall and IoU scores for the different datasets. It could be seen that the overall IoU score is minimum 94% over all the datasets combined. Even though the following results were only marginally better than the state of the art models (Table 2.2), the generated page boundary was a polygon unlike quadrilaterals generated by the other models. The polygons provide a more

realistic boundary(as shown in Figure 2.11) that is required by the morphological correction algorithm to reduce deformations.



(a)



(b)

Figure 2.12: Comparison of text detection and recognition results using Google Cloud Vision API. (a) depicts the recognition results for original non-corrected page while (b) depicts the results for a morphologically corrected page. It can be seen that the detection and grouping of text (light blue bounding box), number recognition (the right page number is 146) and text recognition (la melampirum, jasminum grandiflorum, veronice seurellera) results are much improved in (b).

2.4.1 Ablation study

An ablation study was performed to evaluate the importance of morphological correction. To do so, the page samples before and after morphological transfor-

mations of the Mac Load test data set were chosen and text detection / recognition was applied. The Google Cloud Vision API was used to obtain text detection and recognition. The results of text detection and recognition were quantitatively observed. The number of right predictions for both text detection and recognition tasks were manually verified and counted for both the samples of the image. Since the image quality between the two images is similar and the same text recognition model setting is used, it makes the obtained results comparable. It was observed that the morphological correction improved both the localisation and prediction of hand written text by 25% on average for each image. It was also noticed that page numbers, headings and text that were close to the boundaries had much better results than before. An example of the result is shown in Figure 2.12.

2.5 Conclusion and future work

A novel page detection algorithm has been presented which eliminates border noise by segmenting the main page region from the rest of the image. The importance of using HSV colour model for historical document processing was elaborated. With less assumptions, it was showed that the page detection could also work for complex page structures. It was also demonstrated that the detected page polygon could be used as a feature for reducing deformation. Finally, the page with reduced deformations was proved to perform better in automatic text detection tasks.

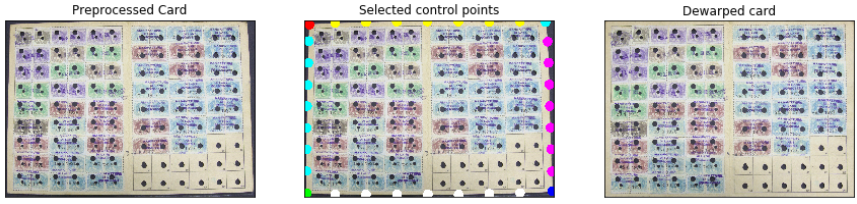


Figure 2.13: Performance of our proposed algorithm on Italian Worker cards. The worker cards are two side foldable cards where stamps were used to denote the payment of the workers. Our algorithm was used to extract the boundary of the cards for further processing.

In order to further portray the versatility of the algorithm, the algorithm was utilised in the context of our AI4EU STAMP project to pre-process digitised Italian worker cards from the 1910s to 1930s. As shown in Figure 2.13, the algorithm performed as expected. However, the thresholds had to be manually selected for the proper functioning of the algorithm. As our future work, I would explore the possibilities to automatically select the thresholds so that the algorithm will work on any collection without manual intervention.

Acknowledgement

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 (FWO) research infrastructure under the grant number FWO I001721N for the DISScO-Flanders project.

3

Segmentation and detection of species in multi-specimen herbaria

This chapter deals with the challenges of automatic enrichment of multi-specimen herbaria. The state of the art approaches for detection and localisation of herbaria sheets with single specimen works close to expectation. However, the problem is exponentially difficult when there are multiple specimens present in the same page. The problem gets even tougher since there are no labelled data available for training deep learning models for such complex scenario. This chapter addresses these problems by proposing an augmentation algorithm to automatically generate labelled multi-specimen data from single specimen herbaria. It utilises the existing approaches to generate single specimen segmentation masks and creates a complex augmented herbaria sheet with multiple single specimen masks. The chapter reports a 38% improvement in the results for the segmentation task while training using the generated data.

⋆ ⋆ ⋆

This chapter is an adapted version of the following original publication:

Automatic extraction of specimens from multi specimen herbaria

Submitted to ACM Journal on Computing and Cultural Heritage (JOCCH), 2022

Abstract 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 herbarium poses unique challenges and problems that have not been adequately addressed. The complexity of localisation 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, experiments have been performed to identify the models that perform well for the plant specimen localisation problem. The major bottleneck for performing such experiments was the lack of labelled data. This problem is also addressed, by proposing tools and algorithms to semi-automatically generate annotations for herbarium images. Based on the experiments, it has been witnessed that segmentation models perform much better than detection models for the task of plant localisation. The 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.

3.1 Introduction

The 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 [Bras et al., 2017, Carranza-Rojas et al., 2018] or to extract other useful information from the herbaria sheets, such as detecting handwritten text, colour bars, scales, and barcodes. The state of the art species identification task works well with an accuracy of 84% for herbarium sheets. However, this is limited only to sheets that have one specimen on a page. There are many herbaria books that have multiple species on the same page (as shown in Figure 3.1) for which the complexity of the detection and localisation 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, a pipeline has been proposed that can automatically detect, identify, and enrich plant specimens in multi-specimen herbaria.

As shown in Figure 3.2, 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

(a) (b)

Figure 3.1: 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 page of interest, such as bar codes and colour bars that need to be removed (see Figure 3.1). Next, the plants visible on the page need to be extracted separately. Finally, a text recognition model would be used to recognise the text on the page and link it to a database of taxa. The feasibility of the pipeline was experimentally validated. This work is part of an ongoing project and therefore the focus of this paper is limited to the extraction of the plant species. Three different methods were investigated, 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.

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

1. A modular pipeline is proposed that can automatically 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 compared for the localisation and detection of plant species.

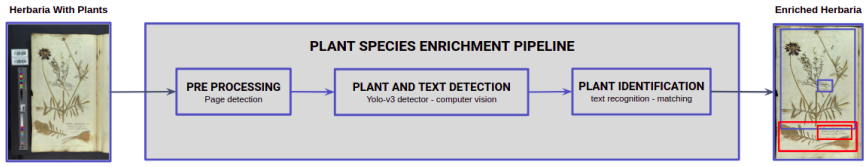


Figure 3.2: Proposed pipeline for automatic enrichment of the herbaria books.

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, the labelled data used for training and validating our models is also provided.

The remainder of this chapter is organised as follows. Section 3.2 reviews the related work. Subsequently, Section 3.3 presents the feasibility of the building blocks pipeline used for detection and identification of plant species. The dataset along with the algorithms used to generate it are presented in Section 3.4. Section 3.5 discusses experiments and their results while Section 3.6 concludes this chapter and explains the future directions.

3.2 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) [Liu et al., 2016] and 'You only look once' (YOLO) [Redmon et al., 2016] and two-stage, region-based CNN detectors, such as Faster R-CNN [Ren et al., 2015]. 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 [Girshick, 2015] 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 generalise well to a new dataset on few amounts of labelled samples. In [Triki et al., 2020], 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, generalising 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 [Ronneberger et al., 2015] 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 [White et al., 2020], 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 $F1$ score of 0.96, validated on 80 of the images. In [Hussein et al., 2020], 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 [He et al., 2017] 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 [Triki et al., 2021] to segment and localise 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.

3.3 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 localising and identifying plants in multi-specimen herbaria.

3.3.1 Data

The herbaria dataset used in Chapter 2 is reused in this chapter. They are currently hosted within the virtual herbarium of Plantentuin Meise [1]. The goal of this chapter is to have a deeper understanding on the methods that can automatically extract plant specimens from books such as Julius and Aime Mac Leod that have multiple plant specimens in a single page (e.g. Figure 3.1).

[1] Meise Virtual Herbarium : Mac Leod

3.3.1.1 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) [Dutta et al., 2016]. 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.

3.3.2 Feasibility experiment: pre-processing

Within the domain of historical digitisation, it is necessary to pre-process the documents before handwritten text detection, text recognition, line detection, character detection, and specimen segmentation tasks are performed. The page detection and morphological correction algorithm explained in Chapter 2 is utilised to pre-process the bound herbaria.

3.3.3 Feasibility experiment: plant specimen detection

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).

As seen in Figure 3.3, Text, PageNumber and Title classes had no problems and the model with the limited training data was already able to detect 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 can also be seen in Figure 3.1 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. Therefore due to the irregular boundaries of the plant specimens and limited data availability we decided to train both detection and segmentation models on single specimen herbaria.

3.3.4 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

(a) (b)

Figure 3.3: Feasibility experiment results of the Yolov3 model trained on 70 labelled herbaria pages. The text regions are well detected whereas the plant boundaries require a lot more training examples to perform as expected.

with the 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.

An example of text recognition result is shown in Figure 3.4. As shown, the

[2]Google Vision API :Try it here

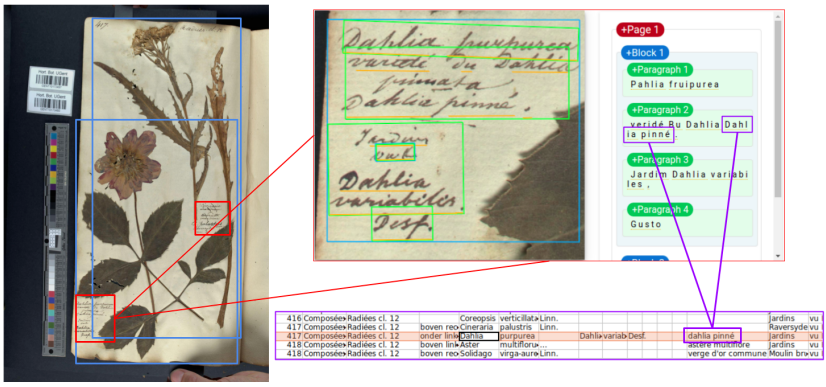


Figure 3.4: Sample text recognition results using Google Cloud Vision API. The plant specimen to the left belongs to the Dahlia flowering plant family. The text had Dahlia variabilis and Dahlia Pinne that was recognised by the API correctly.

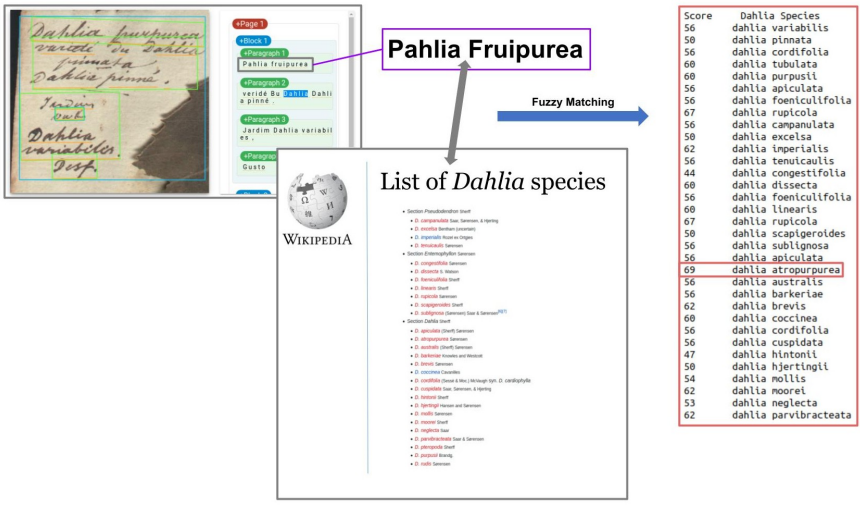


Figure 3.5: Sample fuzzy matching results using the text recognition result obtained from Google Cloud Vision API and list of species in Dahlia *family obtained from wikipedia. As seen, although* Dahlia purpurea *was detected as* Pahlia Fruipurea*, fuzzy matching on all the species in dahlia family showed that the highest score was obtained by* Dahlia atropurpurea *which was the correct species.*

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 with an expected list of possibilities. One option is to perform 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 3.5. There are still problems on how partial recognition would be chosen which would be further investigated and improved in my future works.

3.3.5 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 localise specimens in a multi-specimen herbaria. As identified in Section 3.3.3, the major bottlenecks for improving the localisation of specimens include the lack of labelled data and selecting the right model. The rest of the paper is structured to address these problems.

3.4 Dataset creation, Labelling and Augmentation

Detection models expect bounding box coordinates of the object as explained in Section 3.3.1.1. 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. 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 Algorithm3.4.1, 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 noise. This contour represented

Figure 3.6: Sample segmentation masks that were selected for the data set. The segment masks were semi automatically generated using Algorithm 3.4.1.

the mask of the plant specimen. However, this was not a perfect algorithm and produced some inaccurate masks. Therefore, the generated masks were manually verified and only the perfect ones were chosen 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. Figure 3.6 shows sample specimens that were part of the chosen data set.

During training and testing, the images were down scaled 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 testing 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.

Algorithm 3.4.1: MASK DETECTION AND LABELLING($image$)

Input: Pre-processed image of the herbaria page
Output: Page Mask
I [HSV]←I [RGB]
Generate masks using k-means colour clustering
$Specimen_{mask}$ ←cluster that has the specimen
$Mask$ ← image mask with all zeros
$Specimen_{mask}$ ← Dilation followed by erosion (filtering noise)
$Contour$ ← Find rectangle enclosing *largest Contour* from $Specimen_{mask}$
Generate clean *Mask* using *largest Contour*
Visualise *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

3.4.1 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 [Bochkovskiy et al., 2020] 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 3.7.

On the other hand, a modified form of mosaic augmentation inspired from [Ghiasi et al., 2021] was used to train the instance segmentation models. The

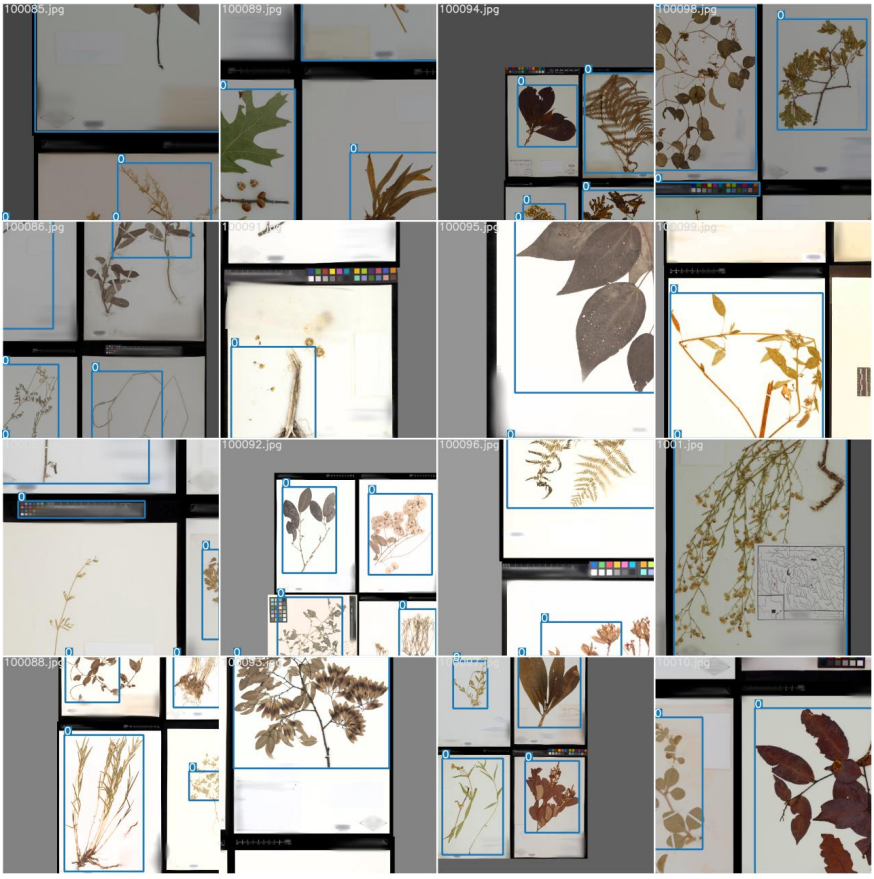


Figure 3.7: Mosaic Augmentation used for training object detection model (batch size 16). Based on the size of the batches, random crops of the image are stitched together in a form of mosaic and is used for training.

copypaste augmentation was not successful in this scenario due to large intra-class variations. Also, copypaste augmentation dint use geometrical transformations. Therefore a modified version of the copy paste augmentation was proposed that performs better for this scenario. The idea here was to use three plant specimens from three different images that 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 3.8.

Figure 3.8: Modified version of mosaic augmentation used for training segmentation models. As seen in the image, an augmented image is generated by placing three plant masks on a empty page image.

3.5 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 de-

tection and localisation. All the experiments were performed on a Linux Intel(R) Core(TM) i5-7440HQ CPU system with a RAM capacity of 64GB; the GPU was NVidia GeForce 980, and the operating system was Ubuntu version 18.04. The entire pipeline was implemented in python 3.6 and Pytorch deep learning library.

3.5.1 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).

3.5.2 Automatic mask generation: pixel-wise segmentation

As discussed in Section 3.4, 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 [He et al., 2015] backbone and the other an EfficientNetB0 [Tan and Le, 2019] backbone. [White et al., 2020] successfully implemented these backbone models for creating segment masks for herbaria sheets and that is the reason why we used the following backbone models. Both of the backbones were pretrained on Imagenet, allowing for a faster generalisation to our dataset compared to training from scratch. The models were trained on a total of 1500 images and validated on 375 images.

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 of selecting crops without any specimen (positive) pixel in them. Therefore

Table 3.1: Results of both image segmentation models on the validation set.

Model	IoU	$F1$
Resnet50	0.951	0.974
EfficientNetB0	**0.956**	**0.977**

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 over fitting [Chen et al., 2020]. 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 [White et al., 2020], 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 ($F1$) function, where the model with the lowest validation loss was saved.

Figure 3.9: 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.

Table 3.1 lists the performance of both models on the validation set containing 128 images. The Resnet and EfficientNet models achieved $F1$ 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 3.9 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 3.9), they could be separately extracted with an additional post processing step. However, many herbarium images contain overlapping specimens, which are not trivial to separate in post processing.

3.5.3 Automatic specimen localisation: detection vs segmentation

To detect and localise 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 initialised and used for training. The models were trained on a total of 1500 images and validated on 375 images.

Figure 3.10: 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.

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 initialised and

used for training. The usage of pretrained weights speeds up the training process [White et al., 2020]. The models were trained on a total of 1500 instances and validated on 375 instances.

Table 3.2: Results of object detection and segmentation on the validation set

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

Table 3.1 lists performance of both models on the validation set containing 128 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 3.10. It could be seen that the Yolov3 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.

3.5.3.1 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.

Sample results are shown in Figure 3.11 and 3.12. 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 was 0.51 which was 38% higher than the non augmented images, which had a mAP of 0.37 as shown in Table 3.3. With more augmented data and resources, the performance of the segmentation model could further be improved.

Table 3.3: Results of ablation experiment performed with and without data augmentation on Mask R-CNN model.The training was limited to 500 epochs.

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

3.6 Conclusion and future work

In this chapter the idea of having an automatic enrichment pipeline for multi-specimen herbaria has been studied. The feasibility of different blocks has been

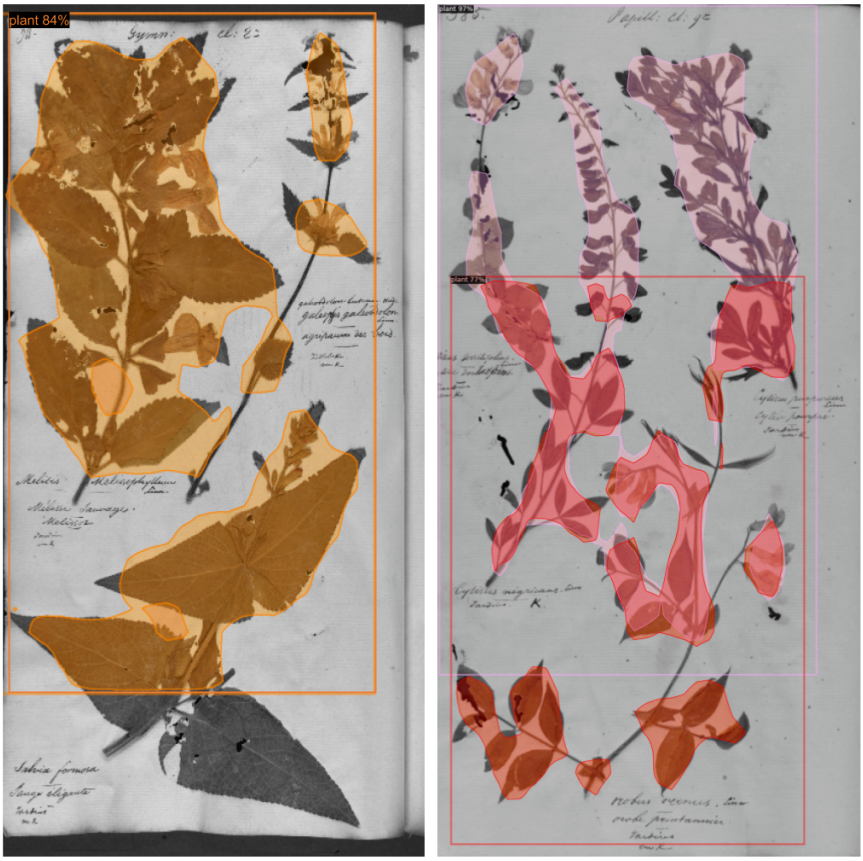


Figure 3.11: Segmentation results of the model trained using non augmented images

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 a pixel-wise segmentation model (U-net) that further refined the specimen masks as explained in Section 3.5.2. The proposed augmentation technique has proved to be extremely beneficial for plant localisation and 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 localisation in complex multi-specimen scenarios since they performed considerably better than the detection models.

Despite having promising results, it is believed that this work merely marks the beginning of this direction. Accurate plant specimen masks are required for

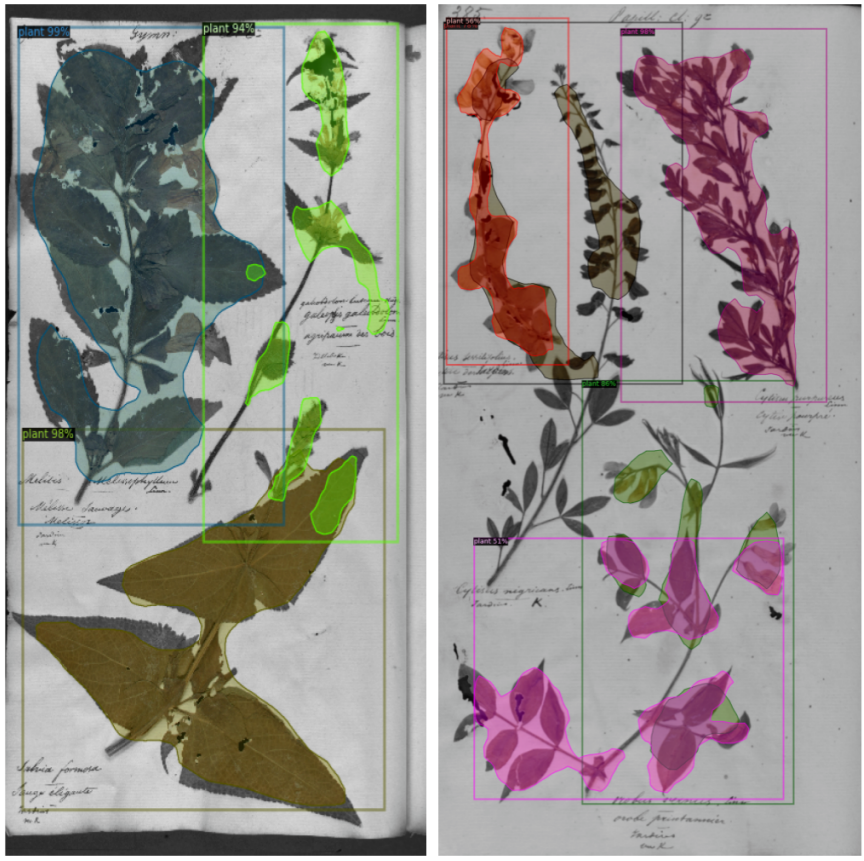


Figure 3.12: Segmentation results of the model trained using augmented images for the same images used in Figure 3.11.

performing automatic trait extraction and therefore further work would be focused 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, it is intended 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.

Acknowledgement

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 (FWO) research infrastructure under the grant number FWO I001721N for the DISSCO-Flanders project.

4

Context based Re-ID of plants in videos

Moving slightly away from the herbaria collections the thesis was focusing so far, this chapter addresses a problem with the automatic enrichment of plants in videos. This chapter proposes a pipeline, that can not only split a video into shots and story units but also utilises it to re-identify objects in them. Plant videos and soap videos that have plants in them are carefully chosen and the pipeline is validated. In order to recognise the detected plants, models have been trained and evaluated and are compared with the existing plant recognition APIs.

⋆ ⋆ ⋆

This chapter is an adapted version of the following original publications:

Context-based structure mining methodology for static object re-identification in broadcast content

Published in Journal of Applied-Sciences-Basel , vol. 11, no. 16, 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.

Published in ACM Journal on Computing and Cultural Heritage (JOCCH), vol. 14, no. 3, 2021

Abstract Technological advancement, in addition to the pandemic, has given rise to an explosive increase in the consumption and creation of multimedia content worldwide. This has motivated people to enrich and publish their content in a way that enhances the experience of the user. In this paper, we propose a context-based structure mining pipeline that not only attempts to enrich the content, but also simultaneously splits it into shots and logical story units (LSU). Subsequently, this paper extends the structure mining pipeline to Re-ID objects in broadcast videos such as SOAP. We hypothesise the object Re-ID problem of SOAP-type content to be equivalent to the identification of reoccurring contexts, since these contexts normally have a unique spatio-temporal similarity within the content structure. By implementing pretrained models for object and place detection, the pipeline was evaluated using metrics for shot and scene detection on benchmark datasets, such as RAI. The object Re-ID methodology was also evaluated on 20 randomly selected episodes from broadcast SOAP shows New Girl and Friends. We demonstrate, quantitatively, that the pipeline outperforms existing state-of-the-art methods for shot boundary detection, scene detection, and re-identification tasks.

4.1 Introduction

Automatic plant identification is helpful for the general audience in recognising plant species without the expertise of botanists. Since 2011, long term efforts have been made by the biodiversity informatics community in organising plant identification campaigns such as PlantCLEF and ImageCLEF [Goëau et al., 2011]. The campaigns coupled with the emergence of novel deep learning architectures has paved the way for the development of dedicated mobile applications such as LeafSnap [Kumar et al., 2012] or Pl@ntNet [Affouard et al., 2017] that has exponentially improved the performance of Image-based plant identification systems. Today, these systems can identify thousands of unique varieties of plants spread across different continents. Interestingly, videos have become the main source of visual information these days. Due to the advances in storage and digital media technology, recording and accumulation of large volumes of video has also become very easy and many popular websites like YouTube, Yahoo Video, Facebook, Flickr and Instagram allow users to share and upload video content globally. Today we stand at the point where the videos that arrive on the internet increase exponentially on a daily basis. Apart from this, there are tons of broadcast channels with enormous amounts of video content, shot and stored every second. These videos contain an interesting amount of plants in them that can be enriched and re-used. However the identification of these plants is a lot more challenging due factors such as differential lighting conditions and changing viewpoints.

Apart from the domain specific problem, with such large collections of videos,

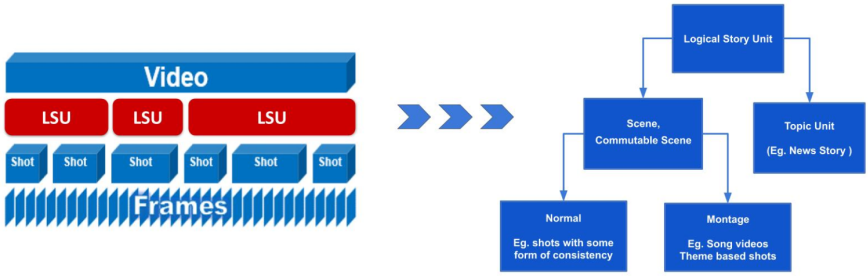


Figure 4.1: Pictorial representation of the structure of video, detailing the position and definition of a logical story unit (LSU). As shown in the flow diagram, a LSU can either be a scene or a topic unit. This chapter predominately focuses on normal scene and topic unit type videos.

it is very difficult to locate the appropriate video file and extract information from them effectively. Additionally with such enormous amounts of data, even the suggestion list increases tremendously, making it even more difficult to make an efficient and informed decision. Therefore, the large sizes of files, the temporal nature of the content, and the lack of right indexing methods to leverage non-textual features, makes it really hard to handle, catalogue and retrieve videos efficiently [Bagdanov et al., 2007]. To address these challenges, efforts are being made in every possible way to bridge the gap between the low level binary video representations and high level text based video descriptions (e.g. video categories, types or genre) [Lu and Grauman, 2013, Mahasseni et al., 2017, Goyal et al., 2017, Han et al., 2018, Meng et al., 2016, Plummer et al., 2017]. Due to the absence of structured intermediate representations, powerful video processing methodologies utilising scene, object, person, or event information does not exist yet. Therefore, in this chapter we address these problems by proposing a framework involving an improved semantic content mining approach that obtains frame level context and object information across the video. The proposed architecture extracts semantic tags such as objects, actions and locations from the videos and utilises them to not only obtain scene/shot boundaries but also Re-ID objects from the video.

Since this chapter deals with several video features/aspects, it is important to clearly define the structure and components of a video in the way they are referenced further in this chapter. Any video can essentially be broken down into several units. First of all, a video is a collection of successive images, more specific frames shown at a particular speed. Thus a frame is one of the many still images that makes the video. Subsequently, a group of uninterrupted coherent frames constitute a shot. Enriching every frame of a video would be computationally expensive and practically inefficient. Every frame belongs to a shot, that lasts

for a minimum of at least 1 second and based on the frame rate of the broadcast video, the number of frames can be anywhere between 20 to 60 frames per second. Thus we find it logical to consider shot as fundamental unit of the video based on which, the entire video can iteratively be enriched with data such as scene types, actions and events.

Humans on the other hand, tend to remember different events or specific scenarios from a video that they use during a video retrieval process. Such an event can be a dialog, action scene or, generally, any series of shots unified by location or a dramatic incident [Minerva M. Yeung, 1997]. Therefore, an event on the whole should be treated as an elementary retrieval unit in advanced video retrieval systems. Multiple terms appear in literature [Vendrig and Worring, 2002] representing temporal video segments on a level above shots and below sequence. These are scenes, logical unit, logical story unit and topic unit . The flow diagram on Figure 4.1 shows how this space could be well defined [Petersohn, 2010]. A logical story unit (LSU) can therefore be a scene or a topic unit depending on the type of content. Our proposed pipeline can automatically segment videos into logical story units.

Researchers often address semantic mining and structure mining problems separately because they were applied to different domains. However, during the last decade, image recognition algorithms have improved exponentially and deep learning models together with GPU/TPU computational hardware, allow very accurate real-time detectors to be trained and served. This has paved way to define complex pipelines that could be defined and reused across multiple domains. We have made use of such technological advancements and defined a versatile semantic extraction pipeline that is proved to address multiple video analytic problems at the same time. In summary, the main contributions of this chapter can be listed as follows:

1. A flexible pipeline is proposed that can derive high level features from detection algorithms and semantically enrich the video by performing automatic video structure mining. This pipeline consolidates the frame level place and object tags using time-efficient deep neural networks in such a way that it could be used for further enrichment tasks such as Re-ID.

2. Within the pipeline, a novel boundary detection algorithm has been implemented which can cluster the temporally coherent and semantically closer segments into shots and LSUs.

3. A novel multi object Re-ID algorithm is also proposed that can generate object timelines based on context similarity in SOAP and broadcast content.

4. A study has also been performed to evaluate the performance of plant identification models and existing plant identification apps and API to automatically identify plants in videos.

The remaining of this chapter is organised as follows. Section 4.2 reviews the related work. Subsequently, Section 4.3 presents our methodology that explains in detail the algorithms used for semantics extraction, boundary prediction and object Re-ID. The experimental set up and model selection are presented in Section 4.4. Section 4.5 discusses the results while Section 4.6 concludes this chapter and discusses the future work.

4.2 Related Work

This work elaborates the role of semantics in video analysis tasks such as video structure mining and Re-ID. Spatial semantics includes object, location and persons in a frame while temporal semantics ranges from actions, events and their interaction across the video. Therefore, understanding a video requires the ability of a system to automatically comprehend such spatio-temporal relationships. In the following we discuss approaches for semantic extraction, LSU/shot boundary detection and Re-ID methodologies.

4.2.1 Semantic extraction

4.2.1.1 Image classification and localisation

Image classification and object recognition tasks have been investigated for a long time. Yet, there were no suitable general solutions that were available. This was mainly attributed to the quality of training data and accessible computational hardware. Also, the classification accuracy for lesser number of classes was observed to have a better performance accuracy than the data set with more classes [Zhou et al., 2016]. But, the image classification task was exponentially boosted majorly due to open competitions such as Large Scale Visual Recognition Challenge (LSVRC) and MIT-Places-365. This led to the development of region proposal networks (RPN) based deep neural networks such as AlexNet, GoogleNet and Vision Geometry Group (VGG) that revolutionised image classification tasks. This opened doors in all directions of image classification and annotation. We use the VGG-16 trained on the MIT-Places-365 for obtaining the place / location of a frame because it is much more generalised and the architecture could be reused for further tasks such as Dense Captioning of a frame that also has VGG-16 as its base architecture.

In addition to the classification tasks, the success of the above mentioned challenges has also fueled research on localisation and detection tasks. Speed and accuracy have been the major areas of focus and based on that, there are two major types of object detection models namely, (1) region based convolution models like R-CNN and Faster RCNN that splits the image into a number of sub images and (2) convolution models such as Single Shot Detector (SSD) and You Only Look Once (YOLO) that detects object in a single run [Redmon et al., 2016]. Even though the Faster RCNN are slightly better in accuracy, the latest version of YOLO, (YOLOv3 [Redmon and Farhadi, 2018]) performs the detection of objects up to 20 times faster and almost similar / acceptable accuracy. Thus our pipeline has a pretrained YOLOv3 model that has been used for detecting objects and persons in a frame.

4.2.1.2 Video annotation

There has also been research pertaining to video annotation. [Altadmri and Ahmed, 2009] proposed an event based approach to create text annotations such that it infers high level textual descriptions of events. This method does not take into account the temporal flow or correlations between different events in the same video. Thus the approach does not have the ability to interact or fuse multiple events into scenes or activities. As explained in the previous section, it is important to search for and retrieve continuous blocks of video often referred to as scenes or story units.

Stanislav Protasov et. al [Protasov et al., 2018] proposed a pipeline with keyframe based annotation of scene descriptions while [Ji et al., 0] proposed a sentence generation pipeline for providing descriptions for keyframes based on the semantic information. Even though the techniques had acceptable results, the annotation still lacked information and also faced information losses. Torralba et al. [Torralba et al., 2003] on the other hand proposed a solution for semantic video annotation that consist of per-frame annotations of scene tags. The per-frame annotations, are computationally expensive and often redundant. Therefore I incorporated a pipeline that takes into account the drawbacks of the previous methodologies. The pipeline obtains all possible spatial information ranging from the location to objects and persons in the form of textual descriptions for every n^{th} frame of the video. n depends on the frame rate of the video and is adjusted in such a way that textual descriptions are obtained for a minimum of 4 frames per second. This was based on the speed vs accuracy trade off where in choosing more number of frames per second signification increases the processing time while processing less number of frame had a higher impact on the performance.

4.2.2 Boundary detection

Shot and scene detection is one of the long studied problems in video structure mining. There have been a lot of different approaches based on different features used and different clustering methods available. In the following we discuss latest approaches for shot and LSU detection.

Based on the existing works for shot boundary detection, we find a striking similarity pattern that prevails in all the works. We have come to a conclusion that boundary detection is performed based on calculating or learning the deviation of features over subsequent frames. Widely used features include RGB, HSV, or LUV colour histograms [Odobez et al., 2003], background similarity [Goyal et al., 2017], motion features [Kwon et al., 2000], edge ratio change and SIFT [Mitrović et al., 2010], and spectral features. [Odobez et al., 2003] uses a spectral clustering algorithm to cluster shots, while [Kwon et al., 2000] proposes a new adaptive scene segmentation algorithm that uses adaptive weighting of colour and motion similarity to distinguish between two shots. They also propose an improved overlapping links scheme to reduce shot grouping time. Recently, deep features extracted using CNN were also employed that obtained significant state of the art results [Gygli, 2017]. They had an end to end trainable CNN model that was trained using a cross entropy loss to detect shot transitions. In this work we employ frame level object, person and location type semantic descriptions as features to estimate shot boundaries.

[Protasov et al., 2018] and [Ji et al., 0] also proposed techniques for scene detection. The former utilises a scene transition graph to cluster similar shots to scenes while the latter proposes to use Jaccard similarity for obtaining similarity between shots. As per survey [Del Fabro and Böszörmenyi, 2013], the LSU detection task is understood as a three-stage problem. In the first step frames are grouped into shots. In the second step location, person and object descriptions are consolidated to obtain shot level descriptions. At the third stage, shot level descriptions are used to cluster the shots into story units using a similarity metric and assumptions about the film structure. I have proposed an algorithm based on this methodology for shot and LSU boundary detections.

4.2.3 Plant identification

There are a number of studies that only focus on leaves to identify plant species. Leaf is the most common, easily accessible and explainable plant organ. Therefore automated identification systems based on learning the variations of leaf characteristics works well on many occasions. [Kumar et al., 2012] proposed a mobile app, called LeafSnap, to enable users to identify trees from photographs of their

leaves. The UK version of the app LeafSnap-UK includes 157 species varieties and achieves a top-1 classification accuracy of 73% and a top-5 classification accuracy of 92%. Yet, these systems heavily depend on only leaves making it unsuitable for those species whose leaf characteristics change tremendously across different seasons round the year.

In the year 2014, [Joly et al., 2014] proposed Pl@ntNet as a crowd-sourced project to accumulate and acquire real world botanical image data at a greater pace. It has both a web and mobile front-end which enable users to identify the plant in a photograph and share their observations. Both web and mobile platform of Pl@ntNet allows the collection of different plant organs (i.e., leaves, fruits, barks, flowers) and also of the complete plant in natural neighbourhood condition. This paved way for multi organ plant identification approaches. [Joly et al., 2014] also evaluated a multi organ identification approach on about half of the plant species of France (2200 species), showing top-5 identification accuracy of up to 69% for single images. Yet, both LeafSnap and Pl@ntNet were designed with constraints based on a preferred set of hand-crafted features to identify plant images and therefore suffered scalability issues.

The automated plant recognition field has shown great interest because identification of plants by conventional means is difficult and time consuming. Therefore, as a final step it is important for plant identification approaches to get rid of the hand-crafted features and become model-free. In the last years, the deployment of deep CNNs has especially led to a breakthrough in fine-grained visual categorisation. For instance, [Lee et al., 2015] proposed a CNN based approach for plant identification using leaf images and achieved an average accuracy of 99.7% on a dataset with 44 species. [Zhang et al., 2015] presented a CNN model to classify the Flavia dataset and reported an accuracy of 94,69%. Plant identification challenges like PlantCLEF has also played a vital role in the improvement of the plant identification models. Such challenges have created a platform that lets people around the world compare the performance of their models on a common dataset to see which strategies succeed [Goeau et al., 2017]. The top scoring model uses state-of-the-art deep convolutional neural networks [Lasseck, 2017]. To improve identification performance it uses an ensemble of several models trained on different datasets with multiple image dimensions and aspect ratios. Pl@ntnet has also considerably improved their identification engine by being an integral part of the PlantCLEF initiative and using latest CNN based approaches. [Goëau et al., 2017].

4.3 Methodology

Based on the motivations explained in Section 1, we propose a pipeline that utilises the semantic descriptions and their co-occurrences across the video to address fundamental video processing challenges pertaining to structure mining and object Re-ID tasks. The proposed pipeline is shown in Figure 4.2. We follow a step wise approach to explain the implementation of the pipeline, that can be enumerated as follows:

1. *Semantic extraction*
2. *Structure mining and similarity estimation*
3. *Object Re-ID*
4. *Plant identification*

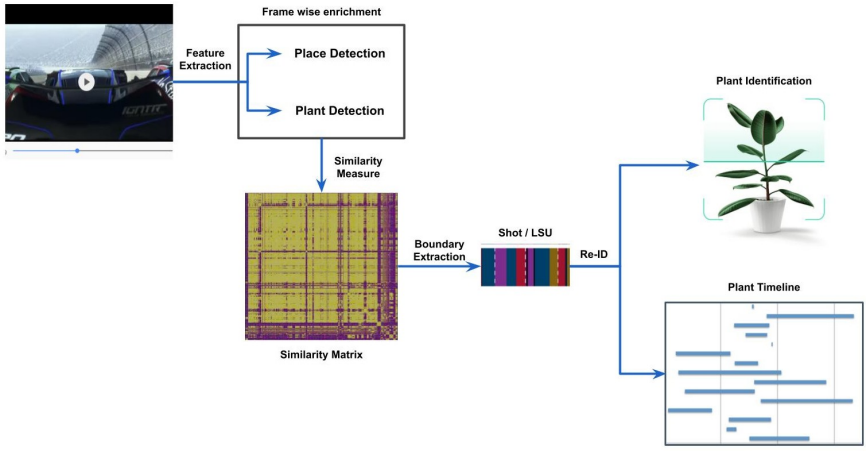


Figure 4.2: Overview of the proposed pipeline. Given the input video, the framework extracts visual features to obtain frame level semantics. The enriched semantic information would then be used for search and retrieval of video segments, predict shot and scene boundaries and to also create plant timelines.

4.3.1 Recognising objects, places and their relations

In order to work with the high level semantic features, it is important to have thorough information regarding the composition of each frame (e.g. objects, persons, and places in the frame). Since broadcast videos don't carry that much of frame level semantic information, it makes it necessary for our pipeline to have a good

model that can predict with high accuracy the objects and places in a frame. As seen in Figure 4.2, the frame level semantic extraction is a common step for all the tasks dealt in the chapter starting from shot / LSU boundary prediction to object timeline generation.

4.3.1.1 Feature extraction

We make use of low level and mid level visual information for predicting the necessary high level features that are used to determine the semantic composition of a logical story unit. In our approach we use the objects, person and location tags as the high level features for detecting the LSU boundaries. To obtain the object and person annotations, the latest version of YOLO object detector [Redmon and Farhadi, 2018] pre-trained on COCO data-set [Lin et al., 2014] is used. COCO stands for Common Objects in Context. The data set comprises of 1.5 million object instances covering 80 object classes. Along with the object detector, the place or the location of the scenes are predicted using the ResNet-50 CNN architecture pretrained on the places-365 data set [Zhou et al., 2016]. In total, Places contains more than 10 million images comprising 400+ unique scene categories [Zhou et al., 2017].

4.3.2 Shot boundary detection

Once we extract the visual features of the video frames, we utilise it to estimate the similarity between frames. This in turn is used to predict the overall structure of the video. Broadcast videos generally have a frame rate of 24fps. As explained in Section 4.1, almost all shots have redundant frames and thus for computational advantage, we process every sixth frame of the video (4 frames / sec). This makes the pipeline six times faster while still preserving its semantic structure. Then, we cluster the temporally similar frames to form shot and story units.

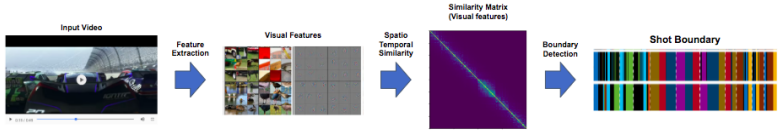


Figure 4.3: Overview of the framework for Shot Detection. Shot is defined as a group of continuous frames without a cut. To predict the shot boundaries, the framework utilises only the frame level visual features from the given input video.

Spatio-temporal visual similarity modeling In contrast to other approaches that use clustering for boundary detection, we build a similarity matrix that jointly

describes spatial similarity and temporal proximity. The generic element S_{ij} defines the similarity between frames i and j as shown in Equation 4.1.

$$S_{ij} = exp\left(-\frac{d_1^2(\psi(x_i), \psi(x_j)) + \alpha \,.\, d_2^2(x_i, x_j)}{2\sigma^2}\right) \tag{4.1}$$

where, $\psi(x_i)$ and $\psi(x_j)$ are the list of visual tags for the i^{th} and j^{th} frame respectively. d_1^2 is the cosine distance between frame x_i and x_j, while d_2^2 is the normalised temporal distance between frame x_i and frame x_j. σ denotes the standard deviation while the parameter α tunes the relative importance of semantic similarity and temporal distance. The effect of alpha on the similarity matrix is shown in Figure 4.4.

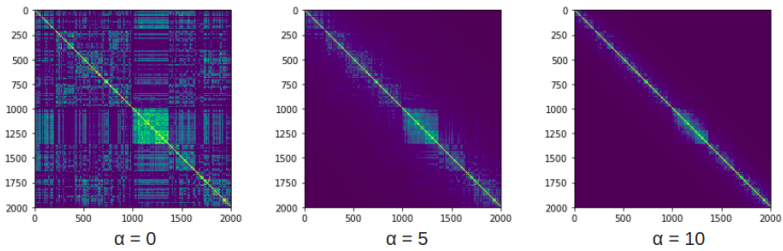


Figure 4.4: Effect of α (from left to right 0, 5 and 10) on similarity matrix S_{ij}. Higher values of α enforce temporal connections between the nearby frames and increases the quality of the detected shots.

As shown in Figure 4.4, the effect of applying increasing values of α to the similarity matrix is to raise the similarities of adjacent frames, therefore boosting the temporal correlations of frames in the neighbourhood. At the same time, too high values of α would lead to the boosting of temporal correlation of very close neighbouring frames thereby failing to capture gradual shot changes. The final boundaries are created between frames that do not belong to the same cluster. An experiment was conducted with the videos of RAI dataset where values from 1 to 10 were provided for α and its effect was studied. We found that α value of 5 performed well on average for both gradual and sharp shot changes. Therefore, in our experiments, we use α value as 5 for our shot boundary detection experiments since it provides the right amount of local temporal similarity required for the prediction of boundaries.

As seen in Equation 4.1, semantic composition based frame similarity estimation is composed of the following two sub parts namely:

- Semantic similarity scoring scheme

- Temporal model analysis

4.3.2.1 Semantic similarity scoring scheme

We use the cosine similarity principle to measure inter frame similarity. By cosine similarity we measure the cosine angle between the two frame vectors. The cosine similarity between the i^{th} and the j^{th} frame is calculated by taking the normalised dot product as follows:

$$sim(x_i, x_j) = ||\psi(x_i)|| \cdot ||\psi(x_j)|| \tag{4.2}$$

where, $\psi(x_i)$ is the normalised vector based on the list of visual tags for frame x_i. This results in a spatial similarity matrix. The similarity measure is converted into a distance measure based on the following Equation:

$$d_1^2(\psi(x_i), \psi(x_j)) = 1 - sim(x_i, x_j) \tag{4.3}$$

An example of utilising the spatial similarity matrix to retrieve top 4 similar frames from a video is shown in Figure 4.5.

Figure 4.5: Effectiveness of the spatial similarity matrix. In this example, the spatial similarity matrix is utilised to retrieve top 4 similar frames from a video. It can be seen that the context in the top four frames are very similar to the context in the query image. The video used is season 5 episode 21 of FRIENDS TV show.

4.3.2.2 Temporal model analysis

As per Equation 4.1 the temporal proximity is modelled using d_2^2, which is the normalised temporal distance between frames x_i and x_j. The normalised temporal distance can be defined by Equation 4.4

$$d_2^2(x_i, x_j) = \frac{|f_i - f_j|}{l} \tag{4.4}$$

where f_i and f_j are the index of frame x_i and x_j respectively and l is the total

number of frames in the video.

Algorithm 4.3.1: SHOT BOUNDARY DETECTION($FrameList$)

Input: List of frame level objects and places tags
Output: Shot boundaries
$shots$ =[ ]
for $i \leftarrow 1$ **to** n
do {
  for $j \leftarrow 1$ **to** n
  do {
    $place\_sim(x_i, x_j) = ||\psi(x_i)|| \cdot ||\psi(x_j)||$
    $obj\_sim(x_i, x_j) = ||\psi(x_i)|| \cdot ||\psi(x_j)||$
    $sim(x_i, x_j) = \frac{w_1(place\_sim) + w_2(obj\_sim)}{w_1 + w_2}$
    $d_1^2(\psi(x_i), \psi(x_j)) = 1 - sim(x_i, x_j)$
    $S_{ij} = exp\left(-\frac{d_1^2(\psi(x_i), \psi(x_j)) + \alpha \,.\, d_2^2(x_i, x_j)}{2\sigma^2}\right)$
  }
}
for $i \leftarrow 1$ **to** n
do {
  if $S_{i,i+1} \leq threshold$
  then {shots.append(i)
}

4.3.2.3 Boundary prediction

Based on Equation 4.1, the lower the value of S_{ij} the more dissimilar frames x_i and x_j are. Thus we calculate the shot boundary by thresholding S_{ij}. In our experiments, 0.4 was used as the threshold value. The threshold value was calculated based on repeating the experiment with values between 0 and 1 with a step of 0.1 and choosing the best value. The entire shot boundary detection algorithm is shown in algorithm 4.3.1.

4.3.3 Context based logical story unit detection

Based on experiments, we have understood that, a normal broadcast content such as a SOAP episode or news, often have multiple angles pertaining to the same story unit and in more than 90% of the cases, the angles return back multiple times throughout the video. Therefore, as shown in Figure 4.6, the context based similarity estimation begins with the shot detection and based on the estimated shot boundaries, frame level semantic descriptions are merged as follows:

$$L_{ij} = \frac{w_1(place\_sim) \; + \; w_2(obj\_sim)}{w_1 + w_2} \tag{4.5}$$

where w_1 and w_2 are the weights for place and object descriptions. In our experiments we have given more importance to the place descriptions than object descriptions mainly because the current state of the art object detection model

does not have the ability to predict all the objects in a frame. Moreover, the pre-trained place detection model has the ability to capture the overall context of the shot location and therefore has been deemed more important. Therefore we have maintained w_1 and w_2 as 2 and 1 respectively in all our experiments.

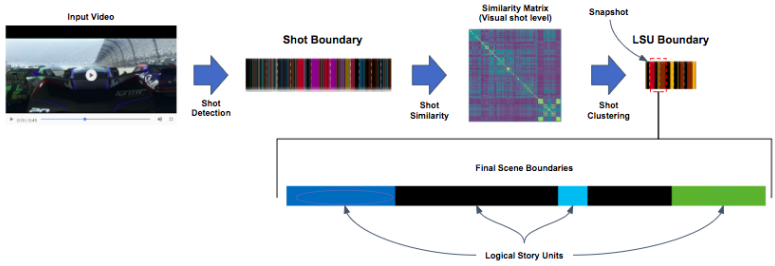


Figure 4.6: Overview of the LSU detection module. Given the input video, the framework extracts visual features to predict the logical story unit boundaries based on semantic similarity between temporally coherent shots. The final decision boundary is based on thresholding the distance between consecutive shots.

The shot level similarity measure is calculated based on the joint similarity estimated using Equation 4.5. The final LSU boundary is based on the similarity threshold of the n continuous shots.

Figure 4.7: An example of shot similarity. The video used is taken from RAI video dataset (23353). The figure also shows the key frame of the shots within a selected LSU (red box).

Based on experiments, we have understood that, a normal broadcast content such as a SOAP episode or news, often have multiple angles pertaining to the

same story unit and in more than 90% of the cases, the angles return back within ± 3 different angles. Thus in our experiments we have used $n = 3$. Also, as explained earlier, all the experiments are carried out with a similarity threshold of 0.4. An example of the similarity matrix of a video from RAI is shown in Figure 4.7. The final similarity matrix is used along with the Re-identification algorithm to generate object timelines.

4.3.4 Object re-identification

We propose an algorithm that formulates unique object ids using the LSUs and frame level object detections such that recurring objects are provided with the same id. The algorithm we propose is based on the following hypothesis:

- If two shots S_a and S_b are similar, then objects present in S_a and S_b are most likely similar.

Explanation

Multimedia broadcast content like SOAP, news,or talk shows, often have reoccurring locations that have objects present in them. Then, based on the above hypothesis, the objects are the same if they are present in the same location. For example, in Figure 4.9, Image 1 is frame 26070 of the video and image 2 is frame 27604. Although they are close to 1500 frames apart, they both pertain to the same location and thus the objects in them are the same. An important point to note is that, the hypothesis holds good only for stationary / static objects. This means, if there are dynamic objects present in the shots (e.g. person) the hypothesis would not work. Thus, our approach focuses only on static object re-identification and this chapter will address problems only pertaining to it.

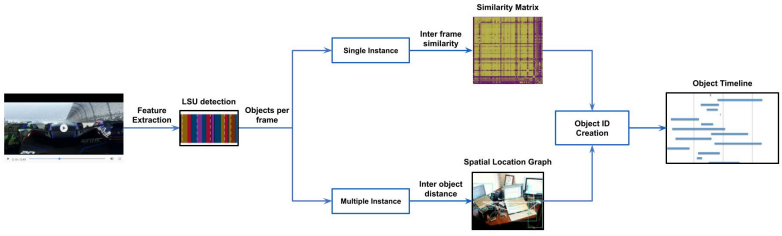


Figure 4.8: Proposed pipeline for multi object Re-ID. Given the input video, we estimate LSU and objects per frame for the video. Based on the number of occurrences of the object in a frame, the objects are categorised as single and multi instance objects. Subsequently using the inter-frame similarity and graph based algorithms, object IDs are created and visualised.

Based on the number of occurrences of a same class object in the same frame, the Re-ID algorithm is composed of two sub parts namely,

- Single instance
- Multiple instance

4.3.4.1 Single instance

If there is just a single occurrence of the object in every frame it appears though out the video, then by hypothesis 4.3.4, the id for object O at frame n is given by Equation 4.6 as follows:

$$O_{id}^{n} = \left\{ \begin{array}{ll} id = 1, & n = 0 \\ O_{id}^{a}, & S_{a}^{n} > threshold, \\ id + 1, & S_{a}^{n} < threshold, \\ & \textit{where (a=1:n-1)} \end{array} \right\} \tag{4.6}$$

where in O_{id}^{n} is the object id at frame n, S_{a}^{n} is the context similarity measure between the frame a and n as calculated in Equation 4.5.

4.3.4.2 Multiple instance

If there are multiple occurrence of the object, we propose a graph based approach to rightly localise the object in the frame. An example of this problem is shown in Figure 4.9. In such cases, where multiple objects of the same class exist, it is not only important to know whether shot / LSU of the frames are similar but also to know the spatial position / location of the object in the frame so that the object could be Re-IDed correctly.

Therefore, based on the bounding box co-ordinates of the detected objects, a location graph is estimated using spatial distances between the objects as shown in Figure 4.10. The idea here is to generate and compare the graphs such that the IDs of the objects can be matched.

Spatial distance estimation Although the 2-D Euclidean distancing measure works really well between frames within similar angles across similar LSUs, there are cases where the angle and zoom changes across similar LSUs. The topological information contained within the frame is also lost making it impossible to obtain a realistic distance estimation. Thus to compensate for the topological information we propose to use depth maps in combination with the location graph to estimate a more realistic spatial distance between the objects in a frame. To

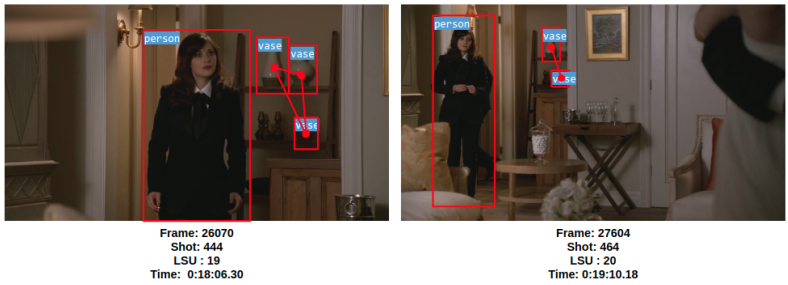


Figure 4.9: Example of multiple instance object class. This example is taken from the season 4 episode 16 of New girl TV SOAP show. In the left side image (frame 26070) there are three different objects of the same class (vase) detected while in the right image (frame 27604), there are two objects of the same class detected.

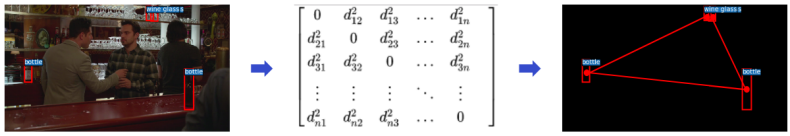


Figure 4.10: Spatial location graph generated for a frame using the centre of the bounding box co-ordinates and Euclidean distance between them.

obtain the depth information, we use Dense Depth [Alhashim and Wonka, 2019] pre-trained on NYU Depth V2 dataset [Nathan Silberman and Fergus, 2012]. The estimated depth is used as the third dimension and thereby the Euclidean measure is re calculated as shown in Figure 4.11.

Let x and y be the centre points of the objects O_x and O_y respectively in a frame, then the distance between them is given by:

$$distance = |\vec{(x)} - \vec{(y)}| \tag{4.7}$$

The estimated depth on the hand, has a range of values that are clipped between 10 and 1000 where 10 is the closest and 1000 being the farthest. If the depth values at points x and y can be represented as $\delta(\vec{(x)})$ and $\delta(\vec{(y)})$, the depth between the objects can be estimated by:

$$depth = |\delta(x) - \delta(y)| \tag{4.8}$$

Finally from Equation 4.7 and 4.8, the actual distance between the objects can be calculated as follows:

$$D_x^y = \sqrt{(distance)^2 + (depth)^2} \tag{4.9}$$

Algorithm 4.3.2: MULTI OBJECT RE-ID($FrameList, ShotList, LsuList$)

Input: Objects_list per frame, shot boundary & LSU similarity
Output: Object IDs per frame
$shots$ =[ ]
for $object \leftarrow object\_list[0]$ **to** $object\_list[len(object\_list)]$
do
- **if** count(objects) in all_frames≤ 1
  - **then** single_instance.append(object)
  - **else** multi_instance.append(object)

for $object \leftarrow single\_instance[0]$ **to** $single\_instance[len(single\_instance)]$
do
- $id = 1$
- **for** $i \leftarrow 1$ **to** $class$
  - **do**
    - **for** $i \leftarrow 1$ **to** n
      - **do**
        - **if** $i == 0$
          - **then** $object_{id} = id$; id = id + 1
          - **else** **if** Sim($frame_n, 1 : frame_{n-1}$ >threshold
            - **then** Let frame a be the frame; most similar to frame n; $object_{id} = O^a_{id}$
            - **else** $object_{id} = id$; id =id+1

for $object \leftarrow multi\_instance[0]$ **to** $multi\_instance[len(multi\_instance)]$
do
- $id = 1$
- **for** $i \leftarrow 1$ **to** $class$
  - **do**
    - **for** $i \leftarrow 1$ **to** n
      - **do**
        - **if** $i == 0$
          - **then** $object_{id} = id$; id = id + 1
          - **else** **if** Sim($frame_n, 1 : frame_{n-1}$ >threshold
            - **then** Let frame a be the frame; most similar to frame n; $object\_list = graph\_compare$ $(G_n[O^{'}_{class}], G_a[O^{'}_{class}])$; **if** $object\_id$ **in** $object\_list$
              - **then** $object_{id} = object\_id$
              - **else** $object_{id} = id$; id =id+1
            - **else** $object_{id} = id$; id =id+1

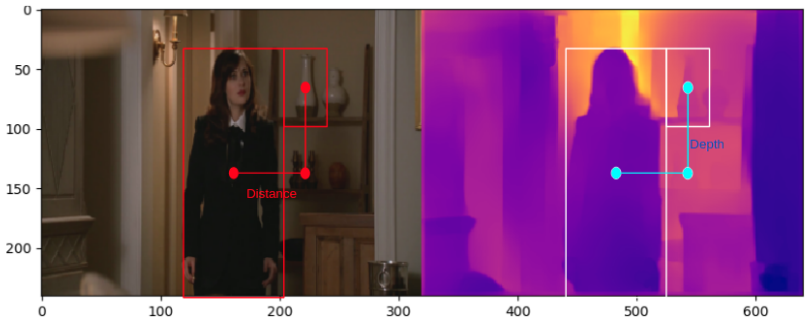


Figure 4.11: Comparison of frame 26070 with its estimated depth. Using the depth and distance measures, the actual distance between the two objects can be estimated.

Spatial location graph For every frame with multiple instance objects, the spatial location graph is estimated based on estimating the pairwise distance between the objects in the frame using Equation 4.9. Let $G_i(O, D)$ and $G_j(O, D)$ be the graphs with objects (O) as nodes and their distance (D) as edges for two similar frames i and j. The objects in frame j are matched with the objects of i based on comparing the distances between the objects in j and i such that the difference between the distances is always minimal. For instance, if frame i has 3 objects O_i^1, O_i^2, O_i^3 of which O_i^1 and O_i^2 belongs to the same class. D_i^{12}, D_i^{13} denotes the distance between object 1&2, object 1&3 respectively. Frame j on the other hand, has two objects O_j^1, O_j^2 in which O_j^1 belong to the same class as O_i^1 and O_i^2. To re-identify objects O_1 in frame j, sub graph distances of $G_i^{'}[O^1]$ and $G_i^{'}[O^2]$ are compared with $G_j^{'}[O^1]$. O_j^1 is equal to the object (O_i^1 or O_i^2) in frame i for which the difference between distances is minimal. The overall object Re-ID algorithm is shown in Algorithm 4.3.2.

4.4 Experiments

For providing a more comprehensive overview of the strength of the pipeline, the pipeline was separately evaluated on benchmark task specific datasets. All the experiments were performed on a Linux Intel(R) Core(TM) i5-7440HQ CPU system with a RAM capacity was 64G; the GPU was NVidia GeForce 980 with 4G memory and the operating system was Ubuntu version 16.04.

The entire pipeline was implemented in python 3.6 and PyTorch deep learning library. The datasets and evaluation metrics used for evaluating our pipeline are explained in the following sections.

4.4.1 Dataset and pre-processing

4.4.1.1 Shot and LSU detection

In this work, a thorough, more objective and accurate performance evaluation has been carried out to evaluate the pipeline for shot boundary detection, LSU boundary detection and object Re-ID.

To evaluate the proposed approach for shot and LSU boundary detection, we test the pipeline on the benchmark RAI data set. The data set is a collection of ten challenging broadcasting videos from the Rai Scuola video archive, ranging from documentaries to talk shows constituting both simple and complex transitions.

We evaluate our approach for Object Re-ID on randomly selected SOAP episodes. For fair evaluation, we chose to validate our approach on two different SOAP broadcast content namely *New Girl* and *Friends*. We selected 10 episodes from season 4 of *New Girl* and 10 episodes from season 3 of *Friends* as our final dataset for object Re-ID.

4.4.1.2 Plant species identification

The Pl@ntnet consortium has a wide collection of floral images for numerous species spread across the world. A small subset of the Western European collection consisting of the specific species of plants present in the painting was obtained and all the experiments were performed on them. There are 75 species in the painting and the distribution of the number of images present in the dataset per species (class) is shown in Figure 4.12. The dataset consist of a total of 101455 images and with a 70-20-10 training-validation-test split, there were 71021, 20393 and 10041 images respectively for training, validation and testing. Since the plant recognition is directed towards users searching for a plant using images taken at its natural setting, efforts were made to ensure that the dataset has images entirely taken at natural environment.

Additionally, it is also possible for users to search for a plant using specific parts/organs of a plant. The visually available parts of most terrestrial plants are leaf, flower, fruit and stem/bark. In order to facilitate searching using any of the plant organ, the dataset has to be consolidated in such a way, i.e., the main focus of each image should belong to one of the fore mentioned plant parts. This would make the trained model more generalised to identify plants with much better accuracy. Our dataset was collected in a similar fashion and the distribution of the dataset based on plant part information is shown in Figure 4.13. It can be seen that leaf, flower, fruit and bark constitute to more than 88% of the dataset.

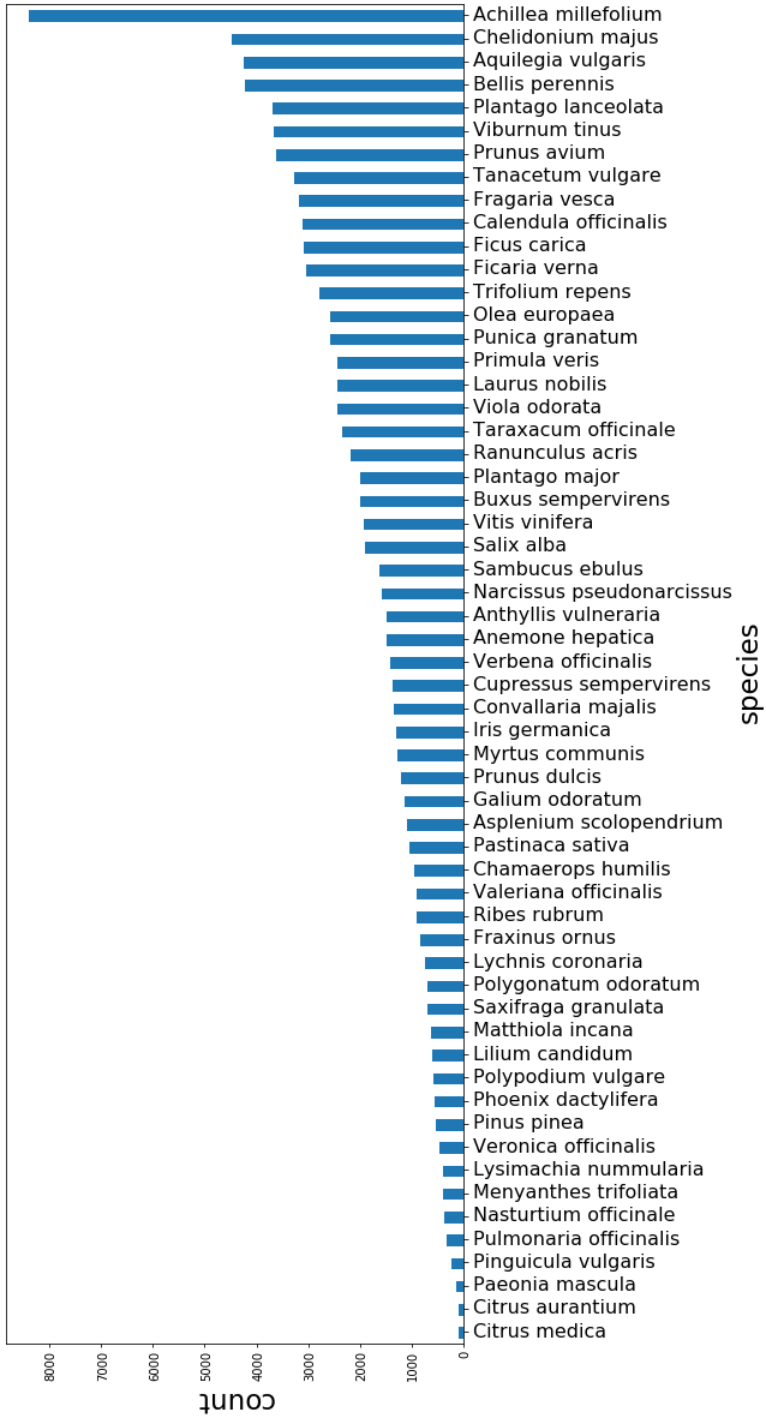


Figure 4.12: Class distribution of our dataset

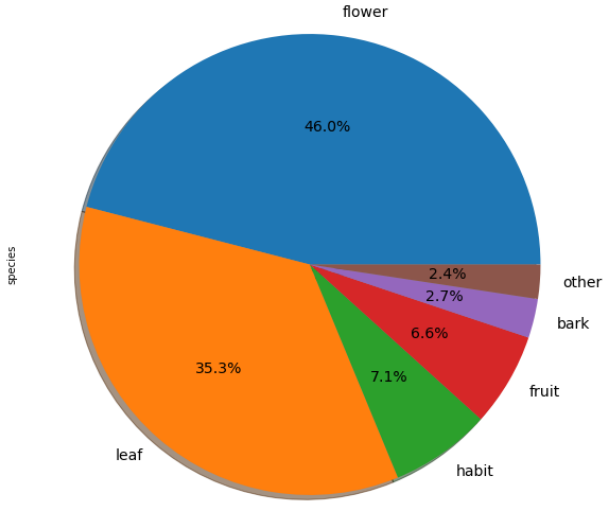


Figure 4.13: Distribution of different parts of the plant within our dataset

As it can be seen in Figure 4.12, the dataset has a linear imbalance with 9 classes having less than 500 images per class. This means that after the split, the training examples would be less than 100 images per class. This is comparatively low for the dataset and therefore the experiments were performed after re balancing the dataset using the imbalance-dataset-sampler[1] package such that the trained model was not biased to a specific class. The experiments were also performed on a sampled dataset which included classes that had minimum of 100 training images.

4.4.2 Evaluation metrics

We evaluated the pipeline based on three tasks: (1) accuracy of the shot boundary detection; (2) accuracy of the LSU boundary detection; and (3) accuracy of object Re-ID algorithm.

For all the experiments, we use the precision, recall and f1-score for the evaluation of our results. Precision, recall and f1-score are computed based on the matched shots / LSU with the ground truth. Furthermore, the results were graphically visualised and analysed for better insights.

[1] Imbalance dataset sampler - *https://github.com/ufoym/imbalanced-dataset-sampler*

The precision measure refers to fraction of rightly predicted boundary from total predictions whereas recall measure denotes the fraction of boundaries rightly retrieved. If $groundtruth$ refers to the list of ground-truth values and $prediction$ refers to the list of automatically predicted values, then precision and recall can be expressed as in Equation 4.10.

$$\begin{aligned} precision &= \frac{|groundtruth \cap prediction|}{|prediction|} \\ recall &= \frac{|groundtruth \cap prediction|}{|groundtruth|} \end{aligned} \tag{4.10}$$

F-score on the other hand combines precision and recall measures and is the harmonic mean of the two. Traditional F_{shot} can be defined as follows:

$$F_{shot} = 2 \,.\, \frac{precision \,.\, recall}{precision \,+\, recall} \tag{4.11}$$

As mentioned in earlier sections, the precision, recall and f1 measure would not suffice to validate the accuracy of the LSU boundary detection algorithm. The reason behind it is that humans and algorithms have different ways to perceive story units. Humans can relate changes in time and location to discontinuities in meaning whereas an algorithm solely depends on visual dissimilarity to identify the discontinuities. This semantic gap makes it impossible for algorithms to achieve fully correct detection results. Therefore, as suggested in [Vendrig and Worring, 2002] I use the $coverage$ and $overflow$ metrics to measure the extent to which our LSU boundary detection algorithm performs with respect to human labelled LSUs, using visual features.

Coverage C measures the quantity of frames belonging to the same scene correctly grouped together, while Overflow O evaluates to what extent frames not belonging to the same scene are erroneously grouped together. Formally, given the set of automatically detected scenes s = $[s_1, s_2, ..., s_m]$, and the ground truth g = $[s_1, s_2, ..., s_n]$, where each element of s and g is a set of shot indexes, the coverage of scene s is proportional to the longest overlap between s_i and g_t:

$$coverage = \frac{max_{i=1...n} \#(s_i \cap g_t)}{\#(g_t)} \tag{4.12}$$

$$overflow = \frac{\sum_{i=1}^{m} \#(s_i / g_t) \,.\, min(1, (s_i \cap g_t))}{\#(g_{t-1}) + \#(g_{t+1})} \tag{4.13}$$

F_{scene} combines coverage and overflow measures and is the harmonic mean of the two. But since for coverage, the closer the value is to 1, the better the performance and for overflow the closer the value is to 0, the better the performance,

we use $1 - overflow$ for calculating the F_{scene}. Thus F_{scene} can be defined as follows:

$$F_{scene} = 2 \,.\, \frac{coverage \;\; x \;\; (1 - overflow)}{coverage \;\; + \;\; (1 - overflow)} \tag{4.14}$$

For the experiments, pertaining to object Re-ID, we make use of $Accuracy$ metrics. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. In our scenario, the predicted observations are labelled as $True$ if they are correctly predicted and $False$ if otherwise. Therefore, if total number $True$ samples are denoted by $True$ and total number of $False$ samples are denoted by $False$, then Accuracy can be calculated as follows:

$$Accuracy = \frac{True}{True + False} \tag{4.15}$$

4.5 Results and discussion

4.5.1 Shot boundary detection

In this study, to evaluate shot boundary detection, we have compared our framework with the state of the art CNN based fast shot boundary detection [Gygli, 2017]. We have used 10 random Internet archive videos from the RAI data set. Table 4.1 compares the precision, recall and F-score with the state of the art. Experimental results depict that the state of the art model performs extremely well on normal transitions, while performing comparatively lower for complex transitions. Our approach on the other hand, has obtained similar precision values for both complex and normal transitions. Eventually, on an average, our approach has outperformed the state of the art with an f1 measure of 0.92.

Table 4.1: Performance comparison for shot detection using boundary level metrics.

	Gygli et al [Gygli, 2017]			**Our Approach**		
Video	**Precision**	**Recall**	**F$_{shot}$**	**Precision**	**Recall**	**F$_{shot}$**
23353	0.95	0.99	**0.96**	0.877	0.99	0.945
23357	0.91	0.97	0.939	0.874	0.99	**0.940**
23358	0.92	0.99	**0.954**	0.775	0.99	0.873
25008	0.94	0.94	**0.94**	0.849	0.99	0.918
25009	0.97	0.96	**0.965**	0.726	0.98	0.841
25010	0.93	0.94	0.935	0.955	0.99	**0.977**
25011	0.62	0.9	0.734	0.863	0.99	**0.927**
25012	0.66	0.89	0.758	0.890	0.890	**0.89**
Average	0.853	0.948	0.899	**0.861**	**0.986**	**0.912**

4.5.2 LSU boundary detection

In this study, we also evaluated LSU boundary detection by comparing the results against two different algorithms for scene detection: [Sidiropoulos et al., 2011], uses a variety of visual and audio features that are integrated in a Shot Transition Graph (STG) while [Baraldi et al., 2015] used a spectral clustering algorithm and Deep Siamese network based model to detect scenes. We used the same 10 videos from the RAI data set for validation. Table 4.2 tabulates the coverage and overflow measures calculated based on the above methods. Experimental results depict that the model [Sidiropoulos et al., 2011] has the highest coverage value of 0.8 but it also has a very high overflow measure . [Baraldi et al., 2015] on the other hand has a comparatively better overflow measure and has better overall performance than [Sidiropoulos et al., 2011]. Although our approach achieved a lower coverage measure, it has obtained a very good overflow measure, which has resulted in higher F_{score}. Our approach with an average F_{score} of 0.74, outperformed the other methods in comparison by more than 10%.

4.5.3 Object Re-ID

In this study, to evaluate Object Re-ID, we have applied the algorithm on 10 random episodes from season 4 of *NEW GIRL* and 10 random episodes from season 3 of *FRIENDS* TV shows. The dataset does not possess ground truth labels. Thus the approach was manually validated as follows: If the object was Re-IDed correctly it was marked $True$ else it was marked $False$. The $True$ and $False$ were consolidated per object class for all the episodes of *New Girl* and *Friends* separately and the object classes that at least have a minimum of 20 occurrences in all the episodes of SOAP put together were chosen to estimate the accuracy. The accuracy was then calculated for each SOAP separately. Table 4.3 shows the accuracy results for the object Re-ID applied on two SOAP series. Experimental results depict that the object Re-ID performs at an average accuracy of 0.87.

4.5.4 Ablation study: effect of depth

In order to evaluate the importance of *depth* information in spatial distance estimation, test were conducted by selecting random frames of different angles from similar LSUs and distance was estimated with and without depth information. For example, as shown Figure 4.14, distance and depth were measured for two different frames. Depth based distance based on Equation 4.9 and normal Euclidean distance between the objects person and vase were estimated. On comparing the depth based distance and Euclidean distance between the two frames, it was seen that the error between the depth based distance is very less compared to error between the Euclidean distances. The experiment was repeated to 10 different

Table 4.2: Performance comparison for LSU detection using frame level metrics.

	Lorenzo et al [Baraldi et al., 2015]			Sidiropoulos et al [Sidiropoulos et al., 2011]			Our Approach		
Video	**Coverage**	**Overflow**	**F**$_{scene}$	**Coverage**	**Overflow**	**F**$_{scene}$	**Coverage**	**Overflow**	**F**$_{scene}$
23553	0.82	0.40	0.69	0.63	0.20	0.70	0.66	0.0083	**0.79**
23557	0.77	0.24	**0.76**	0.73	0.47	0.61	0.65	0.2016	0.72
23558	0.77	0.37	0.69	0.89	0.64	0.51	0.73	0.1346	**0.80**
25008	0.42	0.06	0.58	0.72	0.24	**0.74**	0.41	0.0100	0.58
25009	0.95	0.76	0.39	0.69	0.53	0.56	0.67	0.124	**0.76**
25010	0.66	0.40	0.63	0.89	0.92	0.15	0.66	0.012	**0.79**
25011	0.70	0.14	**0.77**	0.94	0.92	0.15	0.61	0.048	0.74
25012	0.53	0.15	0.65	0.93	0.94	0.11	0.63	0.0400	**0.76**
Average	0.70	0.30	0.66	**0.8**	0.63	0.43	0.63	**0.074**	**0.74**

Table 4.3: Performance evaluation of object Re-ID.

	New Girl (10 episodes)		Friends (10 episodes)	
Class	**True**	**False**	**True**	**False**
bed	29	0	152	0
bottle	604	153	51	14
refrigerator	23	0	56	0
sofa	76	0	306	13
dining table	202	11	87	12
vase	43	8	143	45
bowl	59	0	78	39
tv	-	-	51	0
cup	-	-	69	20
car	74	13	-	-
handbag	61	0	-	-
potted plant	20	0	-	-
Count	1212	187	993	143
Accuracy		**0.866**		**0.874**

scenarios from 10 different episodes and depth based distance error was estimated to be at least 6 times lesser than the Euclidean distance error on an average.

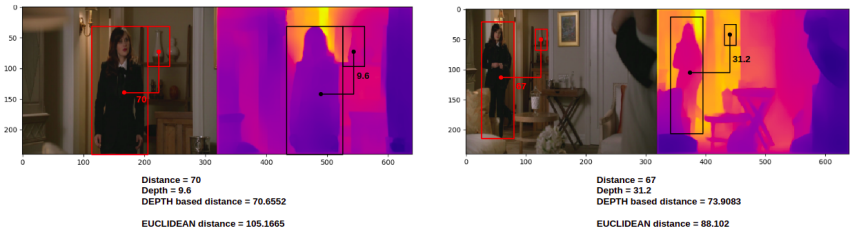


Figure 4.14: An example of ablation experiment to study the effect of depth in spatial distance estimation. Depth based distance is found to be more comparable and less erroneous when compared to the normal 2-D Euclidean distance.

4.5.5 Plant species identification

For the task of plant species recognition, existing models and off the shelf plant recognition apps were tested and compared. In order to ensure proper training and evaluations, a dataset consisting of the plants present in the Van Eyck painting (Chapter 5) was curated as explained in the following subsection.

4.5.5.1 CNN based prediction model

Various backbone CNN models such as VGG-16 (Deep-leafsnap) [Simonyan and Zisserman, 2014], Resnet-18 (Leafnet) [Barré et al., 2017], Resnet-50 [He et al., 2015], Densenet-121 [Huang et al., 2016] and Inception-v3 (Plantnet) [Szegedy et al., 2015] were trained from scratch using the training and validation dataset. The training / validation data was further grouped into chunks of 128 (batch size) and therefore one training epoch consisted of 550 forward and backward training iterations [Simonyan and Zisserman, 2014]. The total training was performed for 100 epochs, while the hyper parameter, learning rate was adjusted after 40 epochs.

Table 4.4: Evaluation of multiple CNN-backbone models on the collected dataset

Model	Entire Dataset (61 classes)		Sampled Dataset (52 classes)	
	Top1 acc	**Top5 acc**	**Top1 acc**	**Top5 acc**
VGG 16	69.249	81.894	75.448	90.723
Resnet 18	75.201	93.081	84.890	96.663
Resnet 50	75.413	93.143	85.448	96.723
Densenet 121	83.298	96.146	**89.193**	**97.659**
Inception v3	**85.665**	**96.587**	87.132	97.038

The results of the experiments on both the entire dataset and sampled dataset is shown in Table 4.4. As expected, the latest model inception_v3 performed the best with a top1 and top5 accuracy of 86% and 97% respectively. On the other hand, for the sampled dataset, that had a minimum of 500 images per class, densenet-121 performed the best with the top1 and top5 accuracy of 89% and 98% respectively. Interestingly, the top5 accuracy of inception_v3 was almost similar for both datasets.

4.5.5.2 Existing plant recognition apps and API

In order to utilise the power of an already existing API, a study was made to see the most suitable plant recognition application that is available for use. The list of applications are shown in Table 4.5.

The most famous plant identification app that is also featured by National Geographic is iNaturalist. It has a large community of people that help with the collection of both flora and fauna data across the world. A part of their collected data is also available for download for non-commercial purposes. iNaturalist is largely a community led identification platform that recently developed computer vision models for identification. Though their android application has a really high accuracy, it is still not possible to use their API for identification. Pl@ntnet on the

other hand offers an identification API that provides access to Pl@ntnet identification engine. Pl@ntnet app and API are a product of the Pl@ntnet consortium, that has over 22000 different plant species across the world. It has a reported accuracy of 0.85 on the top5 retrievals. Plant.id and Flora Incognita are relatively newer plant identification applications that can identify 11000 and 4000 different species respectively. Plant.id does not have an android application and offers only an API service.

Table 4.5: Popular plant recognition applications in Western Europe

Plant Identification applications			
Application	**Number of species**	**Reported Performance**	**API**
Flora Incognita	2700	0.85	NO
iNaturalist	98310	0.91	NO
Plant.id	11000	0.90	YES
Pl@ntnet	29253	0.86	YES

In order to select the best app to be used within the demonstrator, Pl@ntnet and plant.id APIs were quantitatively evaluated by collecting 150 images from the 75 classes in the painting (2 images/class) and 50 random plant images from other plant classes. Pl@ntnet and Plant.id had very similar performances (0.883 and 0.874) but since Pl@ntnet had a lot more plant classes, Pl@ntnet was preferred for the demonstrator.

4.5.6 Demonstration: plant re-identification

In order to demonstrate the proposed pipeline, the pipeline was utilised to perform plant re-identification within soap and broadcast content. To do so, 10 videos were chosen from Youtube that has plants in them. The videos included soap and plant explanation videos. The plant species in the video were manually labelled and both identification and re-identification of the plants were verified. The pipeline started with the estimation of shot and LSU followed by Re-ID of objects that belong to the class "*Plant /PottedPlant*". Upon having unique plant IDs, plant crops from each of the frames were obtained. The plant identification was applied to every crop and the specimen with the highest votes was assigned as the identified plant. A sample result of the plant identification pipeline is shown Figure 4.15. It was noted that, although soap episodes had a detection from multiple angles, it was still hard for the pl@ntnet API to identify the specimen. The primary reason was due to the fact that the crops obtained were of low quality and had a lot noise. It should also be noted that some of the detected plants were non-real plants which were disregarded. The pipeline worked with an accuracy of 88% on plant videos

by identifying 8 out of 9 plants.

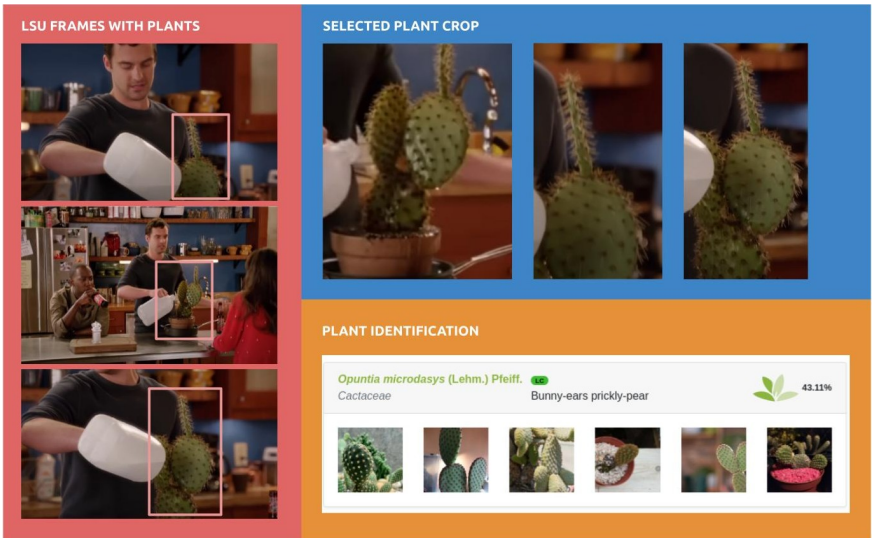


Figure 4.15: Sample result of re-identification of plants in videos. This example is taken from season 1 episode 14 of New Girl TV SOAP. The detected plant crops are identified as cactaceae by the Pl@ntnet API.

4.6 Conclusion and future work

We have proposed a flexible pipeline for annotation, structure mining and Re-ID of objects in broadcast videos by exploring its semantic composition. The high level features extracted from the low and mid level visual features provide useful information about various aspects of the video. A video mining approach has been followed to infer high-level semantic concepts from low-level features extracted from the videos. The results of the video data mining has been further improved, by exploiting temporal correlations within the video and constructing new features from them. Boundary prediction algorithms were proposed for clustering and segmenting videos based on its structure. Furthermore, object Re-ID was also explored and adapted to Re-ID static objects in videos. This helps in creating object timelines that could be interesting for a variety of applications. Our experiments show that our approach is general enough for all broadcast videos of different genres and languages. However, on further introspecting the failure cases, it was found that the selection of similarity threshold played a vital role in the overall accuracy of the pipeline. Therefore, for future work, we would look into adapting the similarity threshold automatically which would further improve

the efficiency of the pipeline. Also, multi-modal features and effective methods to fuse multi-modal information will be investigated. In addition, we would also further optimise the spatial location graph to also include dynamic / moving objects. Finally, the framework will be evaluated on a large scale and the models would be improved accordingly.

Acknowledgement

The research activities as described in this chapter were funded by Ghent University, imec and Flanders innovation & entrepreneurship (VLAIO) agency.

Part III

Demonstrating the cross domain applicability of enriched data

5

Cross-collection linking of botanical imagery in paintings to explore a region's plant richness and diversity over time

The previous part of the book discussed meta data enrichment of historical herbaria and broadcast videos for the detection and identification of plants. The following part introduces two demonstrators that can make the enriched data reach a larger set of audience. This chapter introduces a flexible cross collection linking architecture and explains its importance within the context of interdisciplinary domains.

⋆ ⋆ ⋆

This chapter is an adapted version of the following original publication:

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.

Published in ACM Journal on Computing and Cultural Heritage (JOCCH), vol. 14, no. 3, 2021

Abstract As people on average only spend 20 seconds to watch an artwork, they mostly miss a lot of informative details that is contained within it. As an example,

Figure 5.1: Historical paintings from the 14th and 15th century that has one or more botanical imagery in them

the 75 different plants that can be found in the Ghent Altarpiece is something not a lot of people are aware of. Within this chapter we present a methodology, based on cross-collection linking, to create awareness about the botanical imagery in van Eyck's masterpiece and to inform people about their region's plant richness and diversity over time. As such, this chapter is a nice example of how the interdisciplinary fields of cultural heritage and botany can go hand in hand to facilitate its dissemination to the general public. The plants in the painting can be queried by their name or by a picture taken with a mobile device - a plant recognition app is used to evaluate the pictures taken from the plants. Currently, we link the detected plants to herbaria, observation data, GBIF (Global Biodiversity Information Facility) plantinfo and recent wikimedia commons pictures, but other links can also be easily integrated with the platform. Finally, we also studied nowadays plant observations (volunteered geographic information) in more detail and reveal which region currently has most of van Eyck's plants/flowers.

5.1 Introduction

Plants are fundamental to life on earth constituting the base of the food chain, interacting with living beings and other components of ecosystems and contributing to the balance of these. Extensive technological advancement and focused research has resulted in the collection, documentation and presentation of data regarding a vast majority of plant species that are currently known to mankind. Various organisations and researchers have gathered around the world, funded and created frameworks and research infrastructure aimed at providing anyone, anywhere, open access to data about all types of life on Earth. This has resulted not only in large amount of structured open data that can be used freely but has also facilitated distributed data collection across different countries that can be consolidated. Yet, this abundant information doesn't reach the larger sector of the people since the way this information is presented does not motivate people to search and learn more about the plants.

In the same way, people miss a lot of informative details contained in a painting due to their lack of motivation, expertise and minimal time spent on viewing the artworks. A study made in 2001 by Jeffrey and Lisa Smith at the Metropolitan Museum of Art discovered that the average time spent on a piece of art is 27 seconds. A similar study in 2017 [Carbon, 2017], has shown that the average viewing time of a painting was 32.9s on contemporary art exhibitions. This is quite similar to the time measured in 2001. Another study, done by the Louvre museum unveiled that the Mona Lisa painting, as famous as it is, doesn't really hold visitors attention as much as one might think. The average viewing time for the most famous painting in the world is a mere 15 seconds. The viewing time is significantly affected by viewing distance and crowd and therefore the viewing time of famous artworks is a lot lower. This greatly affects the true essence of a masterpiece.

It is interesting to see that artists have often depicted botanical imagery (i.e. plants, flowers, and trees of a region) as subjects or purposes in their paintings which ranges widely from devotional images of saints and scenes from the scriptures to portraits, still life and subjects from their time period as shown in Figure 5.1. The use of flora in paintings proliferated especially in the fifteenth and sixteenth centuries, as artists became increasingly interested in the realistic depiction of objects from the natural world [Meagher, 2007]. A plant could be depicted either as an attribute, giving clues to the identity of the subject (2nd painting from the right in Figure 5.1), or as providing a moral or philosophical annotation to the subject (2nd painting from the left in Figure 5.1) in addition to the plant's inherent aesthetic property. This makes it possible to enrich such paintings with plant information by cross linking plant data with existing collections. This in turn opens up a new direction that provides a foundation to explore art and horticulture simultaneously.

In order to create awareness about the plant richness within art masterpieces, we propose a pipeline (discussed in Section 3) that enriches these paintings automatically by linking it with various distributed plant collections. In order to demonstrate the pipeline we have also built a web demonstrator[1] (shown in Figure 5.2) with one such famous painting, that not only forges a sense of continuity between the pictorial and the real world but also depicts figures, plants and animals with an enormous sense of precision, The Ghent Altarpiece[2]. The Ghent Altarpiece also known as "The Adoration of the Mystic Lamb" ranks among the most significant works of art in Europe. Painted in the 15th century by brothers Hubert and Jan van Eyck, the detail and close attention to landscape and nature

[1] Web Demonstrator - *http://floredegand.be/lamgods/ ©www.artinflanders.be* - Art in Flanders vzw • original painting @ Sint-Baafskathedraal

[2] Google Culture : The Adoration of the Mystic Lamb

is at a very high level. The numerous recognisable species of plants are minutely depicted with high levels of botanical accuracy. The list of plants found in the Ghent Alterpiece, with their corresponding botanical and common names, is made available at our project website and are also described in detail in the book *The Ghent Altarpiece in Detail* by M. Martens and *Een wonderbaarlijke tuin - Flora op het Lam Gods* by P. Van den Bremt and H. van Crombrugge. Additionally, to better demonstrate the pipeline, the collections would be focused specifically for the region of Belgium, but of course it can be easily extended to other countries as well.

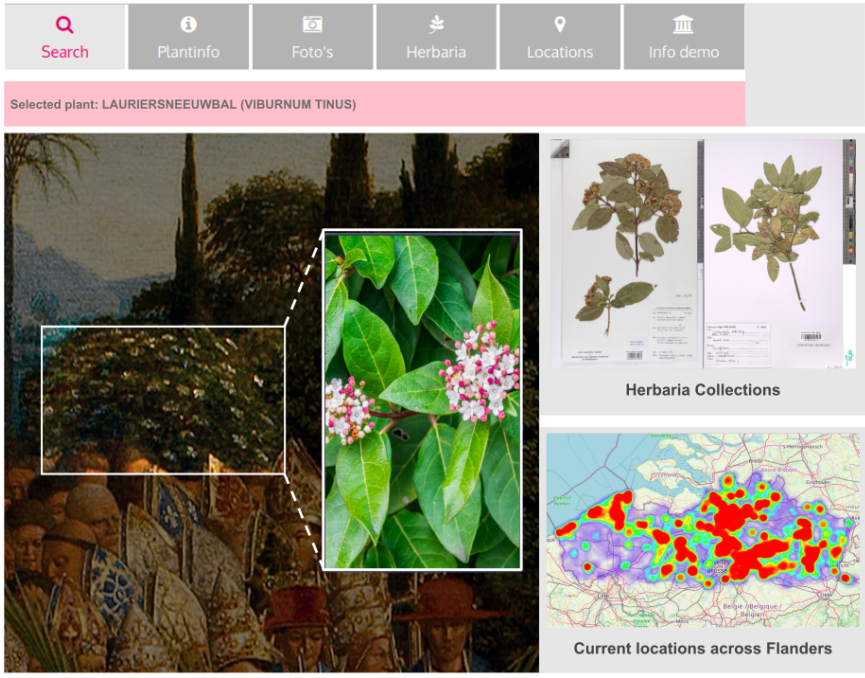


Figure 5.2: Interactive painting with plant highlighting © www.artinflanders.be - Art in Flanders vzw, photo Hugo Maertens.

5.1.1 Research goals

Apart from the interactive web application, contribution of this chapter includes the following research objectives.

1. Motivate people to study artwork (average time people look to paintings is very low).

2. Motivate people to visit places which they probably would not visit and collect data about the biodiversity. We decided to link it to nowadays plant diversity in Flanders and want to stimulate people to go find these plants in their neighbourhood.

5.2 Related work

5.2.1 Linked data and cross collection linking

The heritage domain has constantly been enriched by digitising and collecting data from different institutions across the world. For instance, it has been witnessed that, short descriptions of the digitised cultural heritage items are sometimes improved by linking them with Wikipedia articles which normally include in-depth descriptions and links [Agirre et al., 2012]. This is a very low level example of linked data. On the other hand, efforts and projects such as The Europeana project [3], one of the most prominent examples of digital heritage in recent times, has collected and catalogued millions of digitised museum artefacts owned by numerous institutions across Europe under one roof. In such scenarios, institutions that are providing access to their data face a drawback of creating a gap between the data and itself [Boer et al., 2012]. A case study on aggregation of linked data within the Europeana network showed that the institutions that were already using linked data were able to share their data easily through their libraries which proved to be beneficial due to increased interoperability [Freire et al., 2019]. This also removes the gap between the owners and their data since the data access rights are held with the owning institution, even though the data is made available to other institutions. Subsequently, it can also be seen that the collection and vocabulary statistics of the Rijksmuseum obtained while converting their collection to linked data shows strong positive insights while using linked data [Dijkshoorn et al., 2018].

5.3 Proposed framework

The general architecture of the proposed framework is shown in Figure 5.3. First, the famous painting to be explored is digitised such that we have a high quality digital image to work with. Subsequently, the plants in the painting are annotated using the VGG Image annotator. The annotated painting is automatically enriched by linking it to various horticulture collections and is presented in a web-based platform that allows users (e.g., researchers, local historians, plant enthusiasts and the public at large) to view, explore and learn more about the plant species. Finally, to facilitate improved interaction, the platform has also been designed to search for

[3]The Europeana platform - Europe's digital cultural collection for responsible, accessible, sustainable and innovative tourism

plant species in the painting using plant pictures taken with a mobile device. In this section, we will now further discuss each of the building blocks in detail.

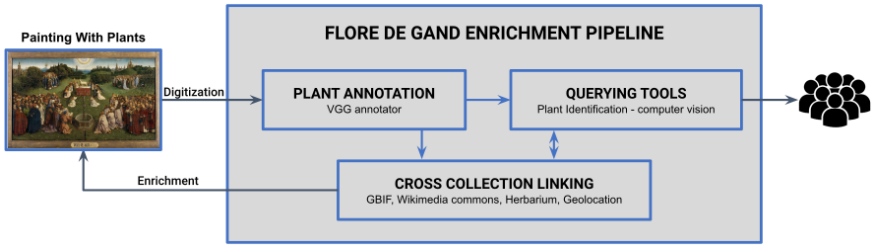


Figure 5.3: The overall architecture of the proposed framework.

5.3.1 Digitisation and annotation

The digital revolution has already transformed visitors' experience in museums, and it can also transform art historians' experience for the better. The digital image used in the demo is available at Lucas - Art in Flanders, the online image library of the Flemish art heritage. The image was a result of the Macro-photography efforts of KIK/IRPA - CSFP and Lucas web. The image has also been imaged using infrared and x-radiography methodologies. Interestingly, Google Art and Culture also joined hands with Lukas - Art in Flanders and Cathedral of Saint-Bavo and digitised the masterpiece[4] in order to let the future generations explore the painting with unprecedented detail. A robotic art camera was used to take 4,000 high-resolution close-ups of the artwork and used those to create the highest ever resolution image ever made for the panels. There are more than 8 billion pixels in that image that lets the user visualise even the tiniest of the details.

The digital image that was used in this study was annotated using the VGG Image annotator[5]. This resulted in a structured JSON file containing the plant names and their corresponding region/coordinates in the painting.

5.3.2 Querying tools

The structured JSON file containing the plant names and corresponding region/coordinates is linked to a drop-down list with Javascript. On selecting a plant name in the drop-down list, the corresponding plant is highlighted on the painting and a zoomed crop is shown to study it in detail. Apart from the drop down list, the

[4]https://artsandculture.google.com/exhibit/RAJydMZDbjc8LQ
[5]http://www.robots.ox.ac.uk/~vgg/software/via/

pipeline also lets the users query the plant species in the painting using dialect names and images of the plant from their neighbourhood as explained below.

5.3.2.1 Dialect plant names to ease search process

As end users will not always be familiar with the botanical name of a plant, a link has been made to the dialect names of these plants over different regions in Flanders. As such, people can query for plants in their own dialect. The mappings between the original plant names and their dialect alternatives are based on the E-WVD Dictionary of the Flemish Dialects (https://e-wvd.be/). Finally, in order to maximise usability/retrievability, a fuzzy string matching algorithm matches the query input to the plant names dataset. An example of all plant dialect names for *Taraxacum officinale* (Dandelion) and the geographical spreading of these names across Belgium is shown in Figure 5.4.

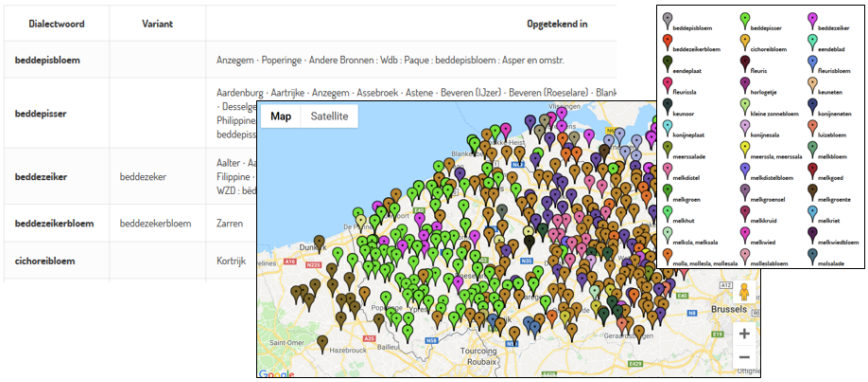


Figure 5.4: e-wvd.be overview of dialect plant names and their geographical spreading for the 'Paardenbloem' plant (Taraxacum officinale).

5.3.2.2 Plant recognition apps to search using plant pictures taken in your neighbourhood

The process of plant identification is an important component of the proposed pipeline since it contributes towards increased engagement of the user. Given the advancement in technological resources, automatic plant species identification has gained a lot of popularity in the last couple of years resulting in enormous innovations. The current demonstrator allows user to search for plants in the painting by using plant images from their neighbourhood. In Section 4.5.5 more details have been provided about the study that was performed on the performance of CNN based plant identification models and other existing APIs for identification of plants. Although the trained CNN models performed well, the Pl@ntnet API

was used in the demonstrator to identify plant species in the search images. This was because Pl@ntnet API had the ability to recognise a wide variety of plant species including species that were not in the painting and this was utilised to provide the users with the feedback in case the recognised plant species was not part of the painting.

5.3.3 Cross-collection linking with plant species

Plant occurrences in the paintings are enriched by linking the plant species with multiple collections. Image datasets, herbaria and observations are some of the sources to which the painting is linked to. Based on the selection made using the querying tools, the pipeline collects linked information about the chosen plant from the fore mentioned sources and presents it to the user. An overview of the collections used in the pipeline is provided below.

5.3.3.1 Observations across Flanders

Waarnemingen.be is a Belgian web platform (which also exists in the Netherlands) where people can register observations of plants. As shown in Figure 5.5 and 5.6, the platform allows to perform a spatio-temporal study of the spreading of a particular plant (such as the Greater Celandine). The platform collects and shares

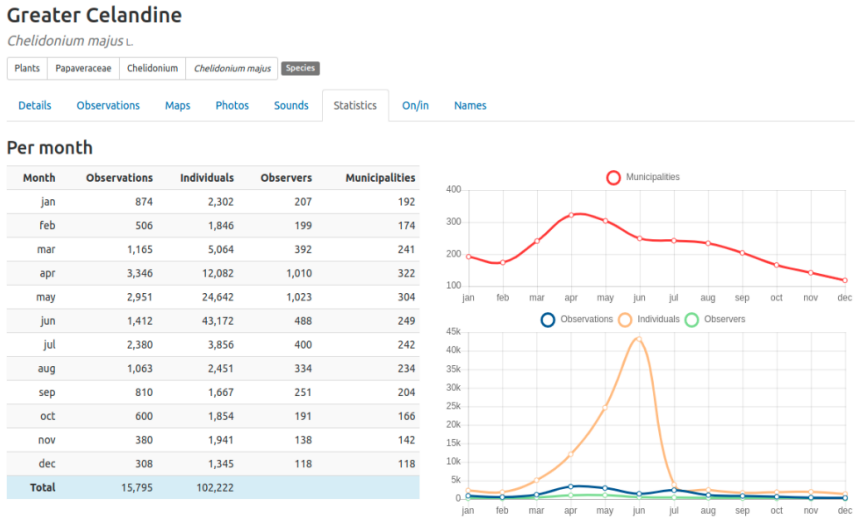


Figure 5.5: Waarnemingen.be with per month observations of the plant Chelidonium majus within Belgium for the period between 2000-2020.

data based on the guidelines of GBIF[6]. GBIF stands for Global Biodiversity Information Facility which is an international network that collects global biodiversity information. A study made with the data for the observations between the years 2000 and 2020 revealed that for the 75 plants that can be found on the Ghent Alterpiece, 6% of the total observations reported were found in Belgium. In total, 306979 observations of the 75 plant species have been reported with an average of 5200 observations per species type. The top 5 most popular plant species from the painting that were found in Belgium are *Ranunculus acris, Plantago lanceolata, Plantago major, Achillea millefolium* and *Trifolium repens*. Also, some exotic plants such as *Chamaerops humilis, Lilium candidum, Cupressus sempervirens* and *Prunus dulcis* could not be found in Belgium within this time period. It is also possible to utilise the geographic information to learn more about the distribution of plants across a region. A sample result of the developed heatmap tool that was used to visualise the heatmaps of plant distribution within Belgium is shown in Figure 5.6.

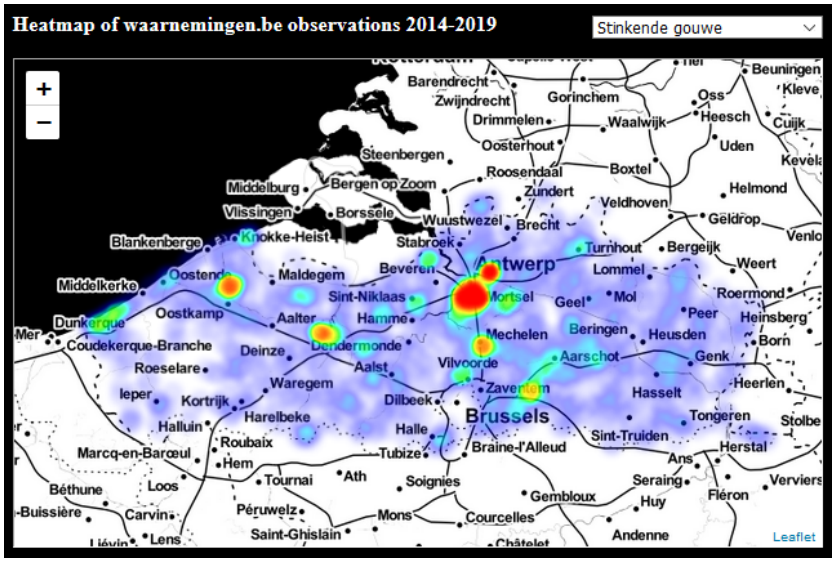


Figure 5.6: Outcome of the Heatmap tool that can be used to visualise the distribution of the plant species of a region. The heatmap shows the distribution of Chelidonium Majus plant across Flanders region in Belgium.

[6]GBIF - *https://www.gbif.org/occurrence/charts?dataset_key=bfc6fe18-77c7-4ede-a555-9207d60d1d86* • sample observations collected all over the world (including those of waarnemingen.be)

5.3.3.2 Availability in botanical gardens

Waarnemingen.be mainly contains species information found in the neighbourhood that are often reported by volunteers. However, Belgium also has an enormous density of flora that are spread across botanic gardens, arboreta and private plant collections that have their own unique identity, profiles and specialisations. The *PLANTCOL*[7] project provides a user interface to query data from such botanical gardens and arboreta in Belgium. The PLANTCOL database currently contains registered information such as the number of accessions in the collections photographs from 9 collections / gardens in Belgium. The list of gardens and collections are shown in Figure 5.7(a). PLANTCOL allows individuals to query their database with the scientific name of the plants using HTML forms. The form is sent to the server via an HTTP POST request while the server sends back the result as an HTML page. For example, to get the amount of accessions, first an HTML POST request is made with the scientific name of the plant. The corresponding response of the server (result page) is then parsed using the XML Path language (XPath) to obtain the requested name and accession count. In the demonstrator, along with the count of plant species within these gardens, a link to these individual accessions for each garden is also provided. This makes it possible for the user to look for the plant in their local botanical garden.

5.3.3.3 Recent photographs of the plant

Wikimedia Commons[8] is a large scale collection of crowd sourced media files. Real world plant photographs from this collection are linked to the plants in the painting, as shown in Figure 5.7(b). This can be interesting for the users to compare the painted plant with its real-life counterpart. Wikimedia Commons makes public domain and freely-licensed media available for everyone and those images can be embedded directly into your application. MediaWiki, the software that powers Wikipedia to collect and share information also powers Wikimedia Commons. There is a simple REST API (Commons API) that allows client code to communicate directly with the MediaWiki software in order to interact with the Wikimedia Common's content. For instance, in our scenario, a request is made with the scientific name of the plant and the server response contains links to the images of the plant. Those image links are visualised using a pop up window in our demonstrator.

[7] PLANTCOL Database - *https://www.plantcol.be/*
[8] Wikimedia Commons - *https://www.commons.wikimedia.org/*

Figure 5.7: Cross collection results for the plant Chelidonium Majus that are added to the demonstrator (a) The first image to left depicts the count of the plants present in the respective private / botanical gardens as obtained from the PLANTCOL database (b) The middle image is an extract from the Wikimedia commons collection (c) The right image is an example of the herbaria obtained from the Botanical Collections of Meise Botanic Garden.

5.3.3.4 Herbaria

Botanical Collections[9] is a collection of digitised herbaria that is promoted by several institutions and maintained by Meise Botanic Garden. The collection contains high resolution scans, data and descriptions for more than 1 million herbaria from all over the world. The collection is indexed by barcode and can be queried by scientific plant name. This makes it possible to retrieve a number of images of herbaria for the requested plant and show them to the user - an example is given in Figure 5.7(c). There is no API available and users make queries via the single page website built on Angular. This means query parameters are sent to the server via an HTTP GET request in JSON format. By imitating the JSON structure and inserting the scientific plant name in it, it is possible to make a request. The server sends the results back in JSON format from which the herbaria barcodes can be easily extracted. The barcodes can then be used to construct URLs that point to the herbaria sheets. The images are served via the Internet Imaging Protocol, so the desired image size can be specified - this avoids the problem of downloading the images at maximum resolution (typically 4000px by 7000px).

5.4 Discussion

The popularity of art has already made big organisations like Getty and Google to invest and digitise art works in a way that end users can zoom into the minute details and consume the art work immensely. The methodologies proposed in this

[9]Botanical Collections of Meise Botanic Garden - *https://www.botanicalcollections.be/*

work improves the experience of the user further by automatically enriching the artworks. This work mainly deals with the cross collection linking of horticulture data contained within an artwork, but this could be easily extended to other types of data within the artwork, such as animals, places or objects.

Geographing is the marriage of Picture, Place, and Time by adding Global Positioning System (GPS) coordinates to your digital photos. Adding location information to photo's creates a historical document of the location where the photo was taken. Years from now people can use these photographs to see the changes on our planet over time. GBIF, iNaturalist, Pl@ntnet and Waarnemingen.be are some of the groups that already perform geographing from volunteered people on plant collections. This already provides interesting insights such as fruits and flowering months, foliage etc based on the observations (e.g. Figure 5.5). Waarnemingen.be additionally allows users to use the location information of where the plants can be found in the neighbourhood. This can be used in combination with the dynamic routing tool explained in the next chapter [Thirukokaranam Chandrasekar et al., 2018], to automatically generate hiking and biking routes through these plant occurrences. The proposed pipeline can be further extended to also include geographing since it already allows people to use plant images to query the plants in the painting.

5.5 Conclusion and future work

A novel cross domain linking methodology has been demonstrated which brings together two completely different domains and acts as a common bridging platform to make exploration of both the domains possible. The pipeline is flexible such that these methodologies can be extended to other paintings and linked collections. Finally, the platform also paves way to let people explore their neighbourhood by providing the flora around their neighbourhood.

Despite having delightful results, we believe that this work merely marks the beginning of the idea of combining two completely different domains to excite and glorify on one another. Apart from extending the framework to other paintings, we want to make use of the resulted cross domain data to train vision models and automate our annotation pipeline. Famous paintings also have various interesting elements such as historical monuments and place references in them which provides us with new domains that we want to explore.

Acknowledgement

The research activities described in this chapter were funded by The Department of Culture, Youth & Media, Flanders (Belgium) for the *Flore de Gand* project.

6

Runamic: POI based dynamic routing application for exploring plant richness in the neighbourhood

This chapter demonstrates a dynamic route planning framework that can generate walking, running or cycling routes based on user preferred points of interest (POI). Apart from regular and popular POIs of a region that are publicly available, as explained in the previous chapter, POIs can also be obtained based on cross collection linking of data from different domains.

⋆⋆⋆

This chapter is an adapted version of the following original publication:

RUNAMIC: dynamic generation of personalised running routes

Published in proceedings of the 6th international congress on sport sciences research and technology support, volume 1: icSPORTS, Seville, Spain, 2018

Abstract This chapter presents a novel dynamic routing application called Runamic. Runamic allows to generate dynamic routes, starting from any point in the city, whilst taking user preferences into account. A user can choose how long he would like to walk alongside his preferred points of interest, such as plants, parks and/or

tourist hot spots. Based on the preferences, characteristics of the road network/environment and feedback based on the previous walks, a route is generated. The suggested tours are loops and as circular as possible, in order to avoid too many turns and overlapping route segments. Finally, it is important to mention that, based on different actuators, the routes can be changed dynamically. After the walk, the user has the ability to rate the generated route, which will influence future route generation. User evaluation shows very positive results on the dynamic routing aspect, and in most cases the graph and R-tree based algorithm generates nice circular routes.

6.1 Introduction

The internet revolution and biodiversity conservation awareness programs have resulted in countries forming international research networks that provide anyone, anywhere, open access to data about all forms of life on Earth. For example, GBIF gives access to more than a billion occurrences of plant data worldwide, that can be accessed in a consistent format according to internationally recognised standards. One of the major contributors to this knowledge are amateur naturalists who share geo-tagged smartphone images of the plant specimens from their environment. This implies that apart from having additional information regarding the specimen, it is also possible to obtain the exact location of a plant specimen in a neighbourhood. However, the way this data is presented makes it less interesting and hard to locate and view the plant. As explained in the previous chapter, cross collection linking tools not only provides the ability to search for plants in the painting but also provides additional information such as geo-data of the plant spread due to the linking. This in turn provides the opportunity to also witness the plants that are in the painting, within a region, live.

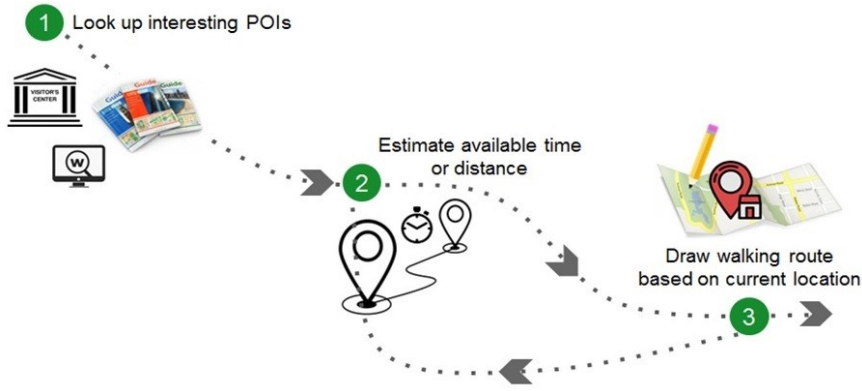


Figure 6.1: Traditional route planning methodology

A variety of applications and devices are already available that can be used to facilitate search and navigation. Based on their route planning ability, these applications could be broadly classified into two categories namely database based route planning and automated route planning. Of course, instead of using one of these digital tools, users can also consult the route books or physical maps that are offered by local tourist offices to plan their routes (Figure 6.1).

Figure 6.2: List of problems in traditional route planning applications

The above mentioned traditional route planners are useful and definitely facilitate the possibility to create routes for recreational walking such as tourism. However, they all have several problems limiting the user experience. As shown in Figure 6.2, the options provided by traditional route planners will always be limited to a couple of predefined routes passing by a set of pre-selected points of interest (POI). The problem with such a limited list of routes is that routes might have been created months or even years ago which might not be relevant for the present day. Also, such routes have fixed lengths and modifying it would mean the users would have to estimate the time and distance of the new route segments manually. This might not only take a lot of time but might also prove to be highly cumbersome for users who are not familiar with the region. Furthermore, the path and route conditions also play a vital role with regards to overall experience. For repetitive activities such as hiking or walking, users often prefer a variation in their route to keep them motivated. Better experiences not only result in more fun, but have also proven to be a form of motivation that paves way for an improved interaction and involvement [Mitchell, 2013]. Finally, pre-defined routes do not include user preferences and since every user has a unique perspective, generalising it would be a highly difficult process. For example, not many people would prefer a route near a busy motorway when compared to a route around a more pleasant

Figure 6.3: Runamic: A Dynamic Route Generation Application. The figure shows an active routing screen of the application. The green line segment in the figure denotes the currently followed route while the blue one denotes a shorter route, requested dynamically.

neighbourhood with enthralling landscapes [Ulrich, 1979]. Therefore, it would be better if people are presented with a list of different sets of POIs or themes around which a route could be generated and modified.

In order to explore the plant POIs in the neighbourhood, a dynamic routing application (Runamic) is proposed that automatically generates circular, preference based routes with dynamic rerouting capabilities. The remainder of this chapter is organised as follows. The following Section 6.2 compares the existing related work in our context of usage. Section 6.3 presents the different types of data that are used as input for the route generation algorithm. Next, Section 6.4 presents the routing algorithm and discusses the optimisations that were investigated to im-

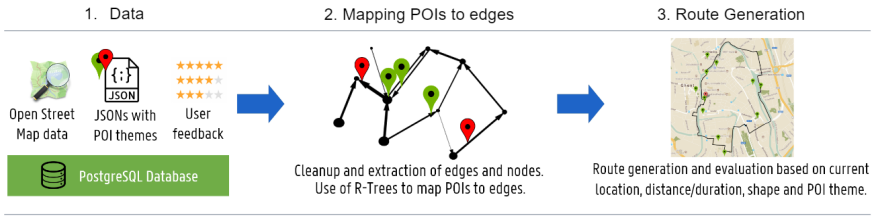


Figure 6.4: General workflow of Runamic framework

prove its performance. Subsequently, Section 6.5 discusses our one-step process to add new types of (dynamic) POIs and Section 6.6 focuses on the Android app. Finally, user engagement and evaluation is covered in Section 6.7 and Section 6.8 concludes this chapter and points out directions for future work.

6.2 Related Work

A user can, for example, create a route using a route plotting application and share this route with the user base like in [RouteYou, 2017] or the user can also choose between various databases of predefined/shared routes such as in [Contours, 2017] for walking or [RunKeeper, 2017, Runtastic, 2017, MapMyRun, 2017] for running. Another possibility is to use a smartphone or a GPS tracker to log the runners location and upload the route to a database [Gavin Maurice, 2011]. Such database based route planning applications work on high quality existing routes, but they are inflexible and not personalised. Availability of these routes may differ strongly between urbanised and rural areas.

A more intelligent method of planning one's route would be a logical choice for the problems discussed above. Existing automated route planning solutions have some flexibility, but this is typically limited to some predefined user profiles [Stroobant, 2016]. The user can draw a route by adding route markers on the map, which are then connected by a shortest path algorithm that can be configured to prefer or avoid certain types of roads [PlotARoute, 2017]. These types of methods are not very user friendly for dynamic changes, since this scheme requires moving/adding markers while performing an activity. Google Maps is also a similar type of application that gives a simple algorithm-based point to point route generation, where the user provides his start position and destination. It is dynamic, yet customisation is still limited to only transportation modes. Finally, there are applications like route suggestions in [RouteYou, 2017] and [RouteLoops, 2017] that generate routes based on the theme or the preference of the user profile. The route planning is completely automatic, yet it is not dynamic.

Apart from these existing applications, there are a number of theoretical solutions such as [Hochmair and Fu, 2009, Hrncir et al., 2014, Su et al., 2010, Turverey et al., 2010], that find the shortest path between two nodes based on user-specified preferences or properties. Works like [Hrncir et al., 2015, Song et al., 2014] addresses the multi-criteria aspect of bicycle routing by looking for a set of Pareto optimal routes. [Storandt, 2012] focuses on the problem of finding Pareto optimal routes by prioritising one of the many routing criteria. This application, on the other hand uses a weighted combination of criteria to search for optimal paths between two nodes.

6.3 Runamic data

6.3.1 Map data

The first part of the workflow, shown in Figure 6.4, is collecting data of the road network and processing this data so that it can be used within the route generation algorithm. Open map data of Open Street Map [OSM, 2018] is utilised to obtain the required nodes and edges structure of the road network such that an appropriate city graph could be generated. For acquiring the OSM data of a specific region (e.g., the city of Ghent), we use the OSM - Osmosis tool [OSM-OSMOSIS, 2018]. This tool serves as an access point for OSM map data and allows to generate custom selected parts of the OSM-map based on the coordinates of the bounding box of the preferred region. The result of this query is the OSM-map data (nodes, ways, relations and tags) for the selected region that gets stored in a temporary XML-file. This XML file contains the following elements:

- *Nodes* represent locations or points on the map.
- *Ways* represent a polyline referencing the nodes that exist on the polyline.
- *Tags* describe specific features of map elements (highways, parks etc).
- *Relations* represent relationships that can't be described by nodes and ways.

On acquiring the data, pre-processing of the data is started by filtering out all the roads which are not suitable for pedestrians. We use the OSM tags associated with the "highway" tag to filter out all the unnecessary data. This is done with the help of the sub tags such as "motorway", which corresponds to extremely busy roads that are not suited for pedestrians. Further optimisations, such as using the activity loggings/heatmaps of a large database of recreational activities (e.g. RouteYou) are currently under development. The result is saved in a new XML-file which is then migrated to our PostgreSQL-database. In this migration step the

nodes and the ways are added to two separate tables as nodes and edges in our database. Since roads can intersect other roads at non-end nodes, they are transformed by splitting them at intersections before they are added to the database. As a last step, we also remove intermediary node information and only keep the start and ending node of the segments.

6.3.2 Points of interest (POIs)

To solve the lack of information required to generate themed routes or touristic routes, POIs were added to the database. Some of the common interest points, like waterways or green region points, can be acquired from the OSM tags. However, the OSM data does not have tags for all possible points of interest in the city. Thus, to enrich the ”interest points” data, a new platform was developed which allow stakeholders to upload their own set of custom POI points in the provided json structure. For example, the tourist administration of the city could upload all tourist attractions points as themed sets for specific types of attractions. In a similar way, cross collection linking gathers the live locations of plants in a region that can be uploaded as a set. The upload to the POI database is also a simple one step process. Furthermore, additional tools were also developed for automatically generating these POI-data from popular existing open data formats, as further discussed in Section 6.5.

6.3.3 Mapping POIs to edges

The POIs are uploaded to the server in the form of JSON and represented as nodes in the database. Since nodes cannot refer to ways (roads) directly, these POIs are stored as locations without any association to nearby roads. To connect these two elements, we need to match each of the POI nodes to the closest way. We realise this by using an R-tree data structure [Guttman, 1984] by which imaginary bounding boxes are built around the ways. These boxes form the base of the R-tree structure. This is used to speed up the search for the road that is closest to the POI.

6.4 Route generation

The actual creation of randomised cycles has been implemented as a server application, in order to be able to deal with increasing map sizes, to report on the usage, and to improve the running time of the algorithm.

6.4.1 Basic generation

The cycle generation algorithm is based on several recent papers [Stroobant, 2016, Stroobant et al., 2018], which specify how cycles can be generated by first gen-

erating a path from the origin to a random point at a distance (the "rod"), and then generating a new path that connects the origin alongside a detour with the rod ("closing" the rod). This algorithm works in essence, but a few issues appeared not long after implementation, which are described in the subsequent subsections.

6.4.2 Feeding extra costs to prevent going back the same way (poisoning)

In order to create a detour that isn't identical to the rod itself, we mark the rod and the space around it with an extra cost. The work [Stroobant, 2016] computes the distances using a graph search, saving all edges in a certain radius. This is memory expensive. Instead, the server uses a linear approximation between the nodes at 12% and 37% of the rod, since they convey the general position and direction of the rod the most accurately. In the second step of the algorithm, only a part of the rod is reused, and statistically this part is exactly half as long as the original rod. Therefore, the 2 selected nodes are the first and third quartile of the expected useful sub rod. Additionally, instead of using a linear function for poisoning, corresponding to

$$\frac{L_{\text{perceived}}}{L_{\text{actual}}} = 1 + \max\left(0,\ \frac{D_{max} - D}{D_{max}} \cdot \alpha\right)$$

we use an exponential decay function for poisoning:

$$\frac{L_{\text{perceived}}}{L_{\text{actual}}} = \exp\left(\left(1 - \frac{D}{D_{max}}\right) \cdot \alpha\right)$$

where $L_{\text{perceived}}$ and L_{actual} are the perceived and actual route lengths while D and D_{max} are the current and maximum Dijkstra short path graph respectively. Originally this change was made to create larger cycles, but this choice also turned out to be the cheaper option when increasing the coverage.

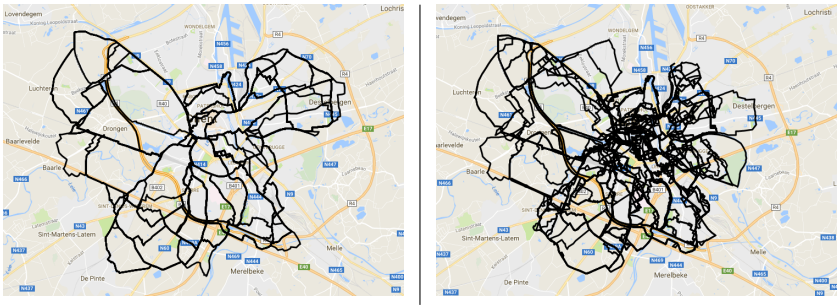
6.4.3 Increasing the coverage

We enhanced Dijkstra's algorithm [Cormen, 2001] with Pareto fronts: instead of saving a best cost for every node in the graph, a vector of Pareto optimal solutions is stored for every node. This allows us to use multidimensional cost functions and to find the shortest path from start to end for every linear combination of the input. This is done because sometimes the usage of an algorithm only yields routes that are slightly too large or too small, and adding a second dimension yields an increased amount of routes with a higher variance, and thus a higher probability of generating a cycle that satisfies our constraints.

As the cost function, we use two randomly chosen instantiating of the hyper parameters, which are a small distance apart, namely 8% of the total route length. Larger differences than that will slow down the execution too much.

6.4.4 Routing through POIs

When we try to route through POIs, a first attempt would be to reduce the cost of the road that contains the POI. However, if this road is short, then reducing the already low cost of going through that road would not influence the cycle generation. A first attempt to fix this is to set the cost to a negative value. Unfortunately, in this case the routing algorithm will stop working, especially when a negative cost cycle is created. Alternatively, we could reduce the cost of all routes close to the POI as well, but this would encourage the routing algorithm to route close to the POI but potentially miss it, which might not be what we want. So, an alternative technique is proposed: whenever a POI is hit, a potential cost is saved in the cost structure, and every subsequent road hit will be cheapened and the potential cost reduced. This forces the algorithm to route through the POI's without introducing negative cycles. In practice, we use a potential with an exponential decay. This has the added advantage that we can simply approximate the cost factor using the potential function itself.



Without road burn | With road burn.

Figure 6.5: Comparison of 100 generated routes with and without road burn. Based on the previous route generations, every generated route is penalised(road burn) to avoid repetitive usage of a same road. Therefore the one with road burn chooses a lot more road segments than the one without road burn.

6.4.5 Randomisation

Since we don't want to navigate the user(s) by the same routes over and over, we decided to add a randomisation aspect to our routing methodology. At this point, it is possible to create a random cycle, and if we look at the randomness from the perspective of the route itself, we see that most routes only have a small part in common, and that they are quite distributed. However, if we look at the randomness from the perspective of the distinct roads, we observe that some roads are included quite often in the routes, while other roads are never included at all. To address this issue, we introduce two features:

- The cost of traversing a road in Dijkstra's algorithm is multiplied with a random number between 0.1 and 0.9. This change allows the algorithm to create shortcuts through small alleys;
- Every route creates road burn, i.e., a small cost that is added to the route segments to avoid repetitive usage of a central road.

As shown in Figure 6.5, this change improved the seeming randomness of the cycle generation by a large amount.

6.5 POI generation and upload

In order to facilitate the generation of POIs and to easily upload POI datasets to the RUNAMIC server, three POI management tools were developed. The first tool, shown in Figure 6.6, allows stakeholders to draw dynamic POI regions (e.g. regions that can only be run during the day because of safety reasons). Both positive and negative POI regions can be created and additional metadata can be added to define the validness information of the particular polygon or rectangular region.

Subsequently, this tool automatically creates the POI data (in JSON format) that can be uploaded to our server. The second tool extracts location entities from text data with a geographic entity recognition (GER) algorithm and transforms the location data into POI coordinates using a geocoder. Figure 6.7 shows some results of evaluating this tool on a collection of historic pictures from the Cege-Soma dataset. As seen, this tool allows us to easily map non-spatial data when it contains sufficient location information in its textual descriptions. Finally, as explained in the previous chapter, *Waarnemingen.be* provides geo location of plant species across Belgium. Scripts were also developed to automatically generate RUNAMIC POI datasets from the obtained plant observations.

6.6 Android application

To test out and visualise route generation, an Android application was developed (shown in Figure 6.8). The application has been published to the Google Play Store for user testing and provides all the tools to go running/walking. Apart from the normal routing capabilities, the app also allows users to link their heart rate monitors for a wholesome interaction. The general flow of the app is as follows: first there is a main screen where the map gets rendered. Then there is a route settings tab where the user specifies his preferences (Figure 6.8a) and POIs for the route generation (Figure 6.8b). Previous runs are stored in the application and get shown in the history tab. Finally, a profile tab is present, where the user can see all

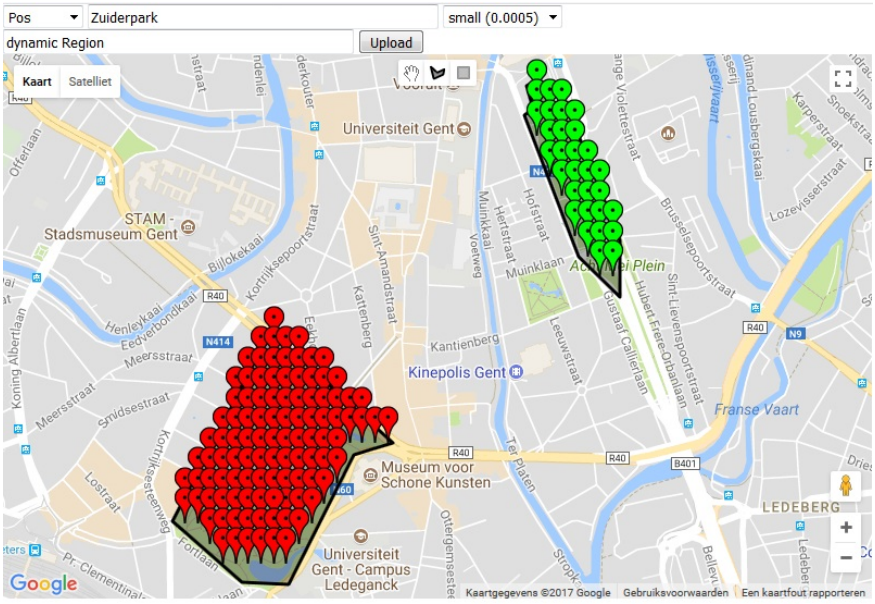


Figure 6.6: Tool for drawing dynamic POI regions

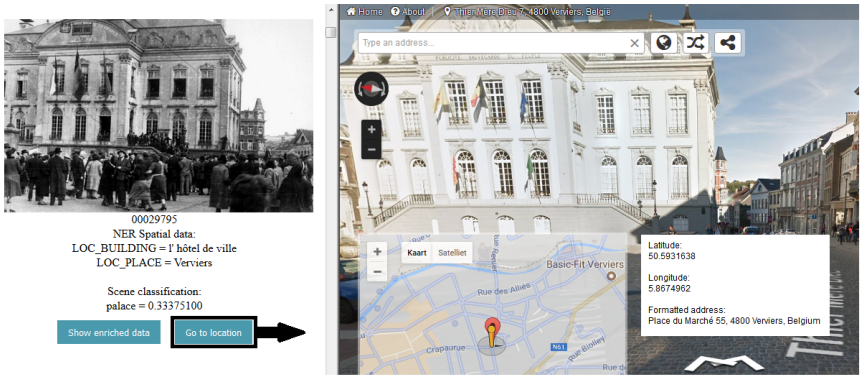
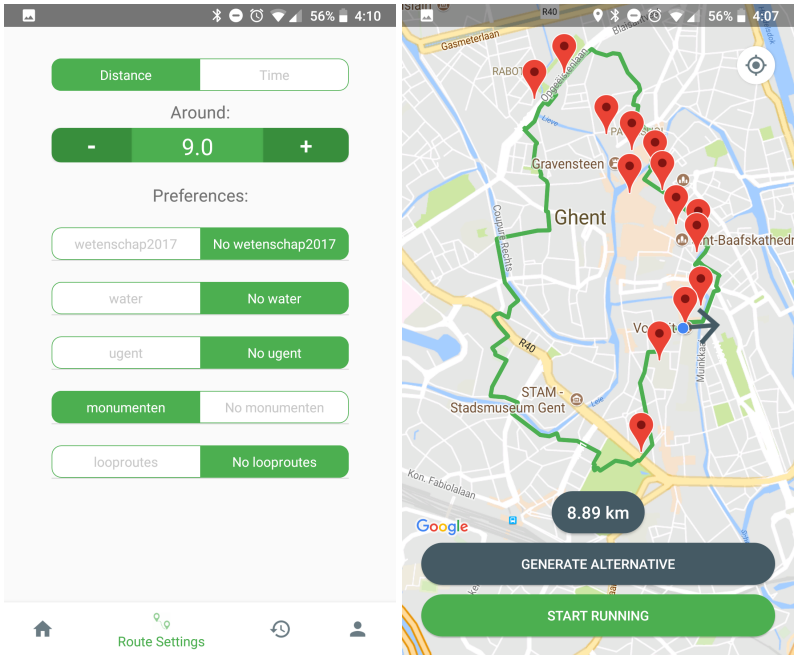


Figure 6.7: Geographic entity recognition (GER) tool to extract locations from textual metadata of images.

his statistics (speed,distance,average heart rate) and can change the settings of the application.

6.6.1 Route preferences

The route preferences in the route settings tab consist of two main elements. The first element is the distance/time the user would like to run. The user can choose



(a) (b)

Figure 6.8: (a.) The preferences screen used for obtaining user preferences. (b.) The main routing screen where the generated route along with the nearest preferred POI points are displayed.

the preferred time or distance. For the time parameter, a level system (beginner, intermediate, expert) which determines the average speed is used. In a later stage this parameter would be determined based on the user base, but for now they are default parameters. The second element allows to make a selection among the available POI themes. When the application boots up it requests all the possible POI themes from the server. The user then chooses which ones he would like to add to generate his personalised request.

6.6.2 Route generation

When the user has chosen the variables for his route, he can go back to the home screen and generate a route. The application sends a request to the server containing the current location and the preferences of the user. The server then generates a route and sends it to the application. The body of the response contains all the nodes for the route, directions and the nearby/used POI nodes. The application renders the route on the map, and allows the user to generate an alternative tour.

Finally, if the user likes the route he can start running. While running, the app saves statistics and provides audio instructions. When the user finishes, he gets the option to give the route a rating. The application sends this rating to the server and the weight of the edges get updated for future route generations.

6.6.3 Dynamic routing

In the application the user has the option to change his route while running. If a user changes his route, the application dynamically adapts the planned tour. For example, if the user takes another turn than instructed, a new route will be requested continuing from his current position but returning to the original end point. Other actuators for dynamic routing are:

- **Heart rate** : When the heart rate of the user reaches a lower level than the minimum threshold, a longer route gets generated. When the heart rate of the user reaches a higher level than the maximum threshold, a shorter route gets generated.
- **Average speed** : Similar to the heart rate the route can be manipulated based on the average speed of the user.
- **Buttons** : In the application, the user can also touch the plus or minus button to increase or decrease the length of the route dynamically whilst running.

6.7 Evaluation

The first version of the application server was tested using a web-application that was used to navigate to various Ghent University locations, during the 200 years of Ghent University celebrations. Following the enormous support to the initiative, the android application was formally released and was made available for download in the Android Play Store from the 17th of October, 2017 in order to perform a first hand evaluation on user engagement and performance. The app was limited to the region of Ghent for feedback and development purposes, but can easily be expanded to any area. The app had an extremely welcoming opening with approximately 1000 unique users logging in and using the app in the first two months (as can be seen in Figure 6.9). This in turn proves the necessity for a more personalised routing application in performing day to day running and walking activities. There were people of different age groups and the feed backs obtained from them were greatly positive.

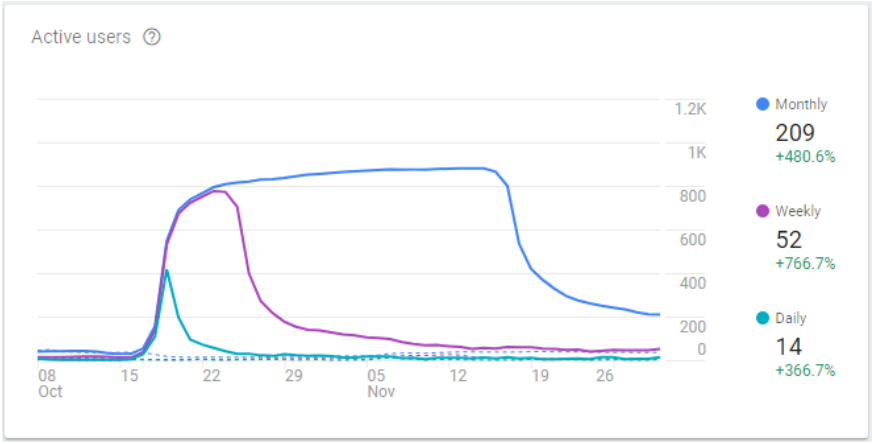


Figure 6.9: Active Users graph for the first evaluation period

6.8 Conclusion and future work

In this chapter we have introduced a mobile application which allows generation of dynamic routes starting from any point in the city based on user preferences. The application has been rolled out for public usage and its functionalities have been limited to the city of Ghent for evaluation, testing and developmental reasons. Currently the application contains a collection of POI datasets covering a variety of themes ranging from nature (parks, water) to tourism (monumenten) and a special POI theme of plants obtained from the cross collection linking of the Van Eyck painting. In most cases, the graph and R-tree based algorithm generates nice smooth routes, leading to very positive user feedback on the dynamic routing aspect.

Part IV

Conclusion

7

Conclusion and future perspectives

In this chapter, we review and summarise the main achievements that led to the results and provide future perspectives on possible advancement studies at the end of this chapter.

⋆ ⋆ ⋆

This dissertation attempts to address and solve some of the problems faced during the enrichment and utilisation of plant specimens for cross domain applications. The main contributions of this dissertation are outlined as follows:

A novel page detection algorithm has been proposed in Chapter 2 which eliminates border noise by segmenting the main page region from the rest of the image. The importance of using HSV colour model for historical document processing was elaborated. With less assumptions, it was showed that the page detection could also work for complex page structures. Although, the numbers show that the performance of the algorithm don't drastically improve the current state of the art, the estimated final page polygon was more usable for performing morphological transformations. It was demonstrated that the detected page polygon could be used as a feature for reducing deformation. Finally, the page with reduced deformations was proved to perform better in automatic text detection tasks.

In Chapter 3 the idea of having an automatic enrichment pipeline for multi specimen herbaria has been studied. The major roadblocks have been identified

and the feasibility of different building blocks has been tested. There was also a problem with the lack of labelled data for the detection and segmentation of plant specimens in herbaria. This has been addressed by proposing a semi-automatic labelling algorithm. The algorithm was further improved by using a pixel-wise segmentation model (U-net) that further refined the specimen masks as explained in Section 3.5.2. To generate labelled data for the complex overlapping scenario, an innovative augmentation technique has been proposed that was proved to be extremely beneficial for localisation and segmentation of plant specimens. It has contributed to a 38% increase in the overall performance of the segmentation model. Finally, it was observed that the instance segmentation models were better suited for the task of plant localisation in complex multi specimen scenarios since they performed considerably better than the detection models.

Chapter 4 proposed a flexible pipeline for annotation, structure mining and Re-ID of plants in broadcast videos by exploring its semantic composition. The high level features extracted from the low and mid level visual features provide useful information about various aspects of the video. A video mining approach has been followed to infer high-level semantic concepts from low-level features extracted from the videos. The results of the video data mining has been further improved, by exploiting temporal correlations within the video and constructing new features from them. Boundary prediction algorithms were proposed for clustering and segmenting videos based on its structure. Furthermore, Re-ID was also explored and adapted to Re-ID plants in videos. This helps in creating plant timelines that could be interesting for a variety of applications. All the more, the overall general applicability of the Re-ID algorithm was also demonstrated using other common objects of interest such as tables and sofas.

There are multiple collections (such as herbaria) or famous paintings where people are not motivated to explore them in detail. Chapter 5 proposes a novel cross domain linking methodology which brings together two such completely different domains and acts as a common bridging platform to make exploration of both the domains possible. A demonstrator has been built that makes use of van Eyck's painting to let people explore the art and horticulture together. The pipeline is flexible such that these methodologies can be extended to other paintings and linked collections.

Chapter 6 has introduced a mobile application which allows generation of dynamic routes starting from any point in the city based on user preferences. The application has been rolled out for public usage and its functionalities was limited to the city of Ghent for developmental, evaluation and testing reasons. To validate the application to generate routes through custom POIs, a web-app was created

with Ghent university buildings as POIs. The web app was used for navigating to different Ghent University buildings during the 200 years of Ghent University celebrations. Likewise, based on cross collection linking methodology explained in Chapter 5, the distribution and location of plants across Belgium was obtained. This list of plant occurrences can be used as POI and routes can be generated which will enable people to explore the plant richness in their neighbourhood.

7.1 Direction for future research

7.1.1 Improvements within the scope of this work

Each of the algorithms explained in the dissertation could further be improved in various directions. They are explained in the following section.

In the context of the AI4EU STAMP project, the page detection algorithm was utilised to pre-process digitised Italian worker cards from the 1910s to 1930s. Although it performed as expected, the thresholds had to be manually selected for the proper functioning of the algorithm. One possible direction for the improvement of the page detection algorithm is to explore the possibilities for automatic selection of thresholds. This would increase the generic applicability of the algorithm and can make it work on a majority of book collections without manual intervention.

The work explained in Chapter 3 merely marked the first steps towards automatic localisation and identification of multi specimen herbaria. In order to perform automatic trait extraction, accurate plant specimen masks are required and therefore further focus would be required to fine-tune the models and obtain better instance masks. Another possible direction is to obtain different parts of each specimen, such as the leaves, flowers, and fruits. 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, it is intended to extend the model towards extracting rulers, colour bars, and other objects around the plant specimen that would further improve the quality of the enriched collection.

The experiments in Chapter 4 show that the boundary detection algorithm is general enough for all broadcast videos of different genres and languages. However, on further introspecting the failure cases, it was found that the selection of similarity threshold played a vital role in the overall accuracy of the pipeline. Therefore, adapting the similarity threshold automatically will be a way to further improve the efficiency of the pipeline.

The idea of combining two completely different domains to excite and glorify on one another is really new and further work is essential to witness its true potential. Apart from extending the framework to other paintings, making use of the resulted cross domain data to train vision models and automate the annotation pipeline is an interesting direction to look into. Famous paintings also have various interesting elements such as historical monuments and place references in them which provides us with new domains that we want to explore.

There is always need for labelled data and expanding the algorithms explained in Chapter 3, 4 to generate labelled data from herbaria and videos could be an interesting line of research. The labelled painting already acts as starting point to further enrich other paintings. Enriching plants in painting to generate similar art works using state of the art generative techniques such as GAN throws interesting perspective to the research mentioned in this dissertation.

Finally, since the dissertation deals with multiple problems across different stages of enrichment, bringing them together to create workflows could be an interesting direction. For example, creating a seamless automatic pipeline for mass digitisation projects by coupling the tools for page extraction, specimen localisation and species identification. The possibility to create such workflows could be further investigated.

7.1.2 New research directions

Apart from the algorithms proposed in this work, there are also interesting directions that this dissertation could open up to. They are explained in the following section.

7.1.2.1 Ruler length estimation

Herbaria often has reference objects such as rulers and colour checkers. A ruler provides a reference to the actual size of the specimen with regards to the image size especially when digitising with a digital camera. The actual dimensions of the image is complex to estimate, as it depends on the camera lens and individual camera parameters. As it is time-consuming to measure each specimen manually, specimen dimensions are often not included as metadata. Automatically estimating ruler length opens up interesting lines of research pertaining to plant phenotyping and spatio-temporal studies on plant evolution. Therefore it is an interesting direction of research to progress into.

7.1.2.2 Painting enrichment

Digitised paintings needs to be further enriched by estimating the context, namely place, time or seasons. Depth estimation techniques can also be investigated to provide a 3-D perspective to a painting. By synthesising this new information, it would be possible to provide context and enhancement to the original cultural heritage objects. Additionally in combination with style transfer models it might also be possible to simulate environments from paintings which could be an interesting use case for experiencing paintings.

Augmented Reality also presents a unique opportunity to offer visitors of cultural heritage sites an innovative way to enjoy technology guided tours inside museums, exhibits, archaeological areas or simply across the city. Such new lines of research that take advantage of the opportunities brought by artificial intelligence, data, and extended reality to preserve our cultural heritage needs to be explored more.

Bibliography

[Affouard et al., 2017] Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.-C., and Joly, A. (2017). Pl@ntNet app in the era of deep learning. In *ICLR: International Conference on Learning Representations*, Toulon, France.

[Agirre et al., 2012] Agirre, E., Barrena, A., De Lacalle, O. L., Soroa, A., Fernando, S., and Stevenson, M. (2012). Matching cultural heritage items to wikipedia. In *LREC*, pages 1729–1735. Citeseer.

[Albiol et al., 2001] Albiol, A., Torres, L., and Delp, E. J. (2001). Optimum color spaces for skin detection. *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, 1:122–124 vol.1.

[Alhashim and Wonka, 2019] Alhashim, I. and Wonka, P. (2019). High quality monocular depth estimation via transfer learning.

[Altadmri and Ahmed, 2009] Altadmri, A. and Ahmed, A. (2009). Automatic semantic video annotation in wide domain videos based on similarity and commonsense knowledgebases. In *2009 IEEE International Conference on Signal and Image Processing Applications*, pages 74–79.

[Bagdanov et al., 2007] Bagdanov, A. D., Bertini, M., Bimbo, A. D., Serra, G., and Torniai, C. (2007). Semantic annotation and retrieval of video events using multimedia ontologies. In *International Conference on Semantic Computing (ICSC 2007)*, pages 713–720.

[Baraldi et al., 2015] Baraldi, L., Grana, C., and Cucchiara, R. (2015). A deep siamese network for scene detection in broadcast videos. *CoRR*, abs/1510.08893.

[Barré et al., 2017] Barré, P., Stöver, B. C., Müller, K. F., and Steinhage, V. (2017). Leafnet: A computer vision system for automatic plant species identification. *Ecological Informatics*, 40:50 – 56.

[Bochkovskiy et al., 2020] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection.

[Boer et al., 2012] Boer, V., Wielemaker, J., Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., and Schreiber, G. (2012). Supporting linked data production for cultural heritage institutes: The amsterdam museum case study.

[Bras et al., 2017] Bras, G. L., Pignal, M., Jeanson, M. L., Muller, S., Aupic, C., Carre, B., Flament, G., Gaudeul, M., Goncalves, C., Invernon, V. R., and et al. (2017). The french museum national d'histoire naturelle vascular plant herbarium collection dataset. *Scientific Data*, 4(1).

[Bukhari et al., 2012] Bukhari, S. S., Shafait, F., and Breuel, T. M. (2012). Border noise removal of camera-captured document images using page frame detection. In Iwamura, M. and Shafait, F., editors, *Camera-Based Document Analysis and Recognition*, pages 126–137, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Carbon, 2017] Carbon, C.-C. (2017). Art perception in the museum: How we spend time and space in art exhibitions. *i-Perception*, 8(1).

[Carranza-Rojas et al., 2018] Carranza-Rojas, J., Joly, A., Goeau, H., Mata-Montero, E., and Bonnet, P. (2018). Automated identification of herbarium specimens at different taxonomic levels. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, page 151–167.

[Chakraborty and Blumenstein, 2016a] Chakraborty, A. and Blumenstein, M. (2016a). Marginal noise reduction in historical handwritten documents – a survey. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 323–328.

[Chakraborty and Blumenstein, 2016b] Chakraborty, A. and Blumenstein, M. (2016b). Preserving text content from historical handwritten documents. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 329–334.

[Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

[Contours, 2017] Contours (2017). Walking in scotland, england, wales, available at: https://www.contours.co.uk/.

[Cormen, 2001] Cormen, Thomas H.; Leiserson Charles E.; Rivest Ronald L.; Stein, C. (2001). Section 24.3: Dijkstra's algorithm. *Introduction to Algorithms (Second ed.). MIT Press and McGraw–Hill.*, page pp. 595–601.

[Del Fabro and Böszörmenyi, 2013] Del Fabro, M. and Böszörmenyi, L. (2013). State-of-the-art and future challenges in video scene detection: a survey. *Multimedia systems*, 19(5):427–454.

[Diem Markus and Basilis, 2019] Diem Markus, K. F. and Basilis, G. (2019). Icdar 2019 competition on baseline detection (cbad).

[Dijkshoorn et al., 2018] Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., Ter Weele, W., and Wielemaker, J. (2018). The rijksmuseum collection as linked data. *Semantic Web*, 9(2):221–230.

[Dutta et al., 2016] Dutta, A., Gupta, A., and Zissermann, A. (2016). Vgg image annotator (via). *URL: http://www. robots. ox. ac. uk/vgg/software/via*, 2.

[Fan et al., 2002] Fan, K.-C., Wang, Y.-K., and Lay, T.-R. (2002). Marginal noise removal of document images. *Pattern Recognition*, 35(11):2593 – 2611.

[Freire et al., 2019] Freire, N., Voorburg, R., Cornelissen, R., de Valk, S., Meijers, E., and Isaac, A. (2019). Aggregation of linked data in the cultural heritage domain: A case study in the europeana network. *Information*, 10(8):252.

[Gavin Maurice, 2011] Gavin Maurice, Ghosh Bidisha; Pakrashi Vikram; Barton John O'Flynn; Brendan Lawson, A. (2011). A cycle route planner mobile-app for dublin city. *Irish Transport Research Network Annual Conference (ITRN2011)*.

[Ghiasi et al., 2021] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

[Goeau et al., 2017] Goeau, H., Bonnet, P., and Joly, A. (2017). Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017).

[Goëau et al., 2017] Goëau, H., Bonnet, P., and Joly, A. (2017). Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). In *CLEF: Conference and Labs of the Evaluation Forum*, volume CEUR Workshop Proceedings, Dublin, Ireland.

[Goodrich et al., 2009] Goodrich, B., Albrecht, D., and Tischer, P. (2009). Algorithms for the computation of reduced convex hulls. In Nicholson, A. and Li, X., editors, *AI 2009: Advances in Artificial Intelligence*, pages 230–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Goyal et al., 2017] Goyal, P., Hu, Z., Liang, X., Wang, C., and Xing, E. P. (2017). Nonparametric variational auto-encoders for hierarchical representation learning. *CoRR*, abs/1703.07027.

[Goëau et al., 2011] Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J.-F., Birnbaum, P., Mouysset, E., and Picard, M. (2011). The clef 2011 plant images classification task. volume 1177.

[Guttman, 1984] Guttman, A. (1984). A dynamic index structure for spatial searching. *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 47–57.

[Gygli, 2017] Gygli, M. (2017). Ridiculously fast shot boundary detection with fully convolutional neural networks. *CoRR*, abs/1705.08214.

[Han et al., 2018] Han, J., Yang, L., Zhang, D., Chang, X., and Liang, X. (2018). Reinforcement cutting-agent learning for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Hochmair and Fu, 2009] Hochmair, H. H. and Fu, Z. J. (2009). Web based bicycle trip planning for broward county, florida.

[Hrncir et al., 2014] Hrncir, J., Song, Q., Zilecky, P., Nemet, M., and Jakob, M. (2014). Bicycle route planning with route choice preferences. *ECAI 2014*, pages 1149 – 1154.

[Hrncir et al., 2015] Hrncir, J., Zilecky, P., Song, Q., and Jakob, M. (2015). Speedups for Multi-Criteria Urban Bicycle Routing. In Italiano, G. F. and Schmidt, M., editors, *15th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2015)*, volume 48 of *OpenAccess Series in Informatics (OASIcs)*, pages 16–28, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Huang et al., 2016] Huang, G., Liu, Z., and Weinberger, K. Q. (2016). Densely connected convolutional networks. *CoRR*, abs/1608.06993.

[Hussein et al., 2020] Hussein, B. R., Malik, O. A., Ong, W.-H., and Slik, J. W. F. (2020). Semantic segmentation of herbarium specimens using deep learning techniques. In *Computational Science and Technology*, pages 321–330. Springer.

[James, 2013] James, S. P. (2013). Face image retrieval with hsv color space using clustering techniques.

[Ji et al., 0] Ji, H., Hooshyar, D., Kim, K., and Lim, H. (0). A semantic-based video scene segmentation using a deep neural network. *Journal of Information Science*, 0(0):0165551518819964.

[Joly et al., 2014] Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.-F., Boujemaa, N., and Barthélémy, D. (2014). Interactive plant identification based on social image data. *Ecological Informatics*, 23:22 – 34. Special Issue on Multimedia in Ecology and Environment.

[Khan et al., 2018] Khan, N. A., Shafi, S., and Ahangar, H. (2018). Digitization of cultural heritage: Global initiatives, opportunities and challenges. *Journal of Cases on Information Technology (JCIT)*, 20(4):1–16.

[Kumar et al., 2012] Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. B. (2012). Leafsnap: A computer vision system for automatic plant species identification. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision – ECCV 2012*, pages 502–516, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Kwon et al., 2000] Kwon, Y.-M., Song, C.-J., and Kim, I.-J. (2000). A new approach for high level video structuring. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 2, pages 773–776. IEEE.

[Lasseck, 2017] Lasseck, M. (2017). Image-based plant species identification with deep convolutional neural networks. In *CLEF (Working Notes)*.

[Lee et al., 2015] Lee, S. H., Chan, C. S., Wilkin, P., and Remagnino, P. (2015). Deep-plant: Plant identification with convolutional neural networks. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 452–456.

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

[Liu et al., 2016] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

[Lu and Grauman, 2013] Lu, Z. and Grauman, K. (2013). Story-driven summarization for egocentric video. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2714–2721.

[Mahasseni et al., 2017] Mahasseni, B., Lam, M., and Todorovic, S. (2017). Unsupervised video summarization with adversarial lstm networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 2982–2991.

[MapMyRun, 2017] MapMyRun (2017). Plan each stride and learn from every route, available at: https://http://www.mapmyrun.com.

[Meagher, 2007] Meagher, J. (2007). Botanical imagery in european painting.

[Meng et al., 2016] Meng, J., Wang, H., Yuan, J., and Tan, Y. (2016). From keyframes to key objects: Video summarization by representative object proposal selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1039–1048.

[Minerva M. Yeung, 1997] Minerva M. Yeung, B.-L. Y. (1997). Video content characterization and compaction for digital library applications. volume 3022, pages 3022 – 3022 – 14.

[Mischke and Luther, 2005] Mischke, L. and Luther, W. (2005). Document image de-warping based on detection of distorted text lines. ICIAP'05, page 1068–1075, Berlin, Heidelberg. Springer-Verlag.

[Mitchell, 2013] Mitchell, R. (2013). Is physical activity in natural environments better for mental health than physical activity in other environments? *Social Science & Medicine*, 91:130 – 134.

[Mitrović et al., 2010] Mitrović, D., Hartlieb, S., Zeppelzauer, M., and Zaharieva, M. (2010). Scene segmentation in artistic archive documentaries. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 400–410. Springer.

[Nathan Silberman and Fergus, 2012] Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *ECCV*.

[Odobez et al., 2003] Odobez, J.-M., Gatica-Perez, D., and Guillemot, M. (2003). Spectral structuring of home videos. In *International Conference on Image and Video Retrieval*, pages 310–320. Springer.

[OSM, 2018] OSM (2018). Open street map: Open source community maintaining map data all over the world, available at: https://www.openstreetmap.org/.

[OSM-OSMOSIS, 2018] OSM-OSMOSIS (2018). A command line java application for processing osm data, available at: http://wiki.openstreetmap.org/wiki/osmosis.

[Petersohn, 2010] Petersohn, C. (2010). *Temporal video segmentation*. Jörg Vogt Verlag.

[PlotARoute, 2017] PlotARoute (2017). Free route planners for outdoor pursuits, available at: https://www.plotaroute.com/.

[Plummer et al., 2017] Plummer, B. A., Brown, M., and Lazebnik, S. (2017). Enhancing video summarization via vision-language embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1052–1060.

[Protasov et al., 2018] Protasov, S., Khan, A. M., Sozykin, K., and Ahmad, M. (2018). Using deep features for video scene detection and annotation. *Signal, Image and Video Processing*, 12(5):991–999.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

[Redmon and Farhadi, 2018] Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *CoRR*, abs/1804.02767.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

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

[RouteLoops, 2017] RouteLoops (2017). Start travelling in different cities, available at: https://www.routeloops.com/.

[RouteYou, 2017] RouteYou (2017). Enjoy the nicest routes, available at: https://www.routeyou.com/.

[RunKeeper, 2017] RunKeeper (2017). Find the best running routes on runkeeper, available at: https://runkeeper.com/index.

[Runtastic, 2017] Runtastic (2017). Running, cycling and fitness gps tracker, available at: https://www.runtastic.com/.

[Sanchez-Cuevas et al., 2013] Sanchez-Cuevas, M. C., Aguilar-Ponce, R. M., and Tecpanecatl-Xihuitl, J. L. (2013). A comparison of color models for color face segmentation. *Procedia Technology*, 7:134 – 141. 3rd Iberoamerican Conference on Electronics Engineering and Computer Science, CIIECC 2013.

[Shafait and Breuel, 2010a] Shafait, F. and Breuel, T. (2010a). A simple and effective approach for border noise removal from document images. pages 1 – 5.

[Shafait and Breuel, 2010b] Shafait, F. and Breuel, T. (2010b). A simple and effective approach for border noise removal from document images. pages 1 – 5.

[Sidiropoulos et al., 2011] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., and Trancoso, I. (2011). Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

[Song et al., 2014] Song, Q., Zilecky, P., Jakob, M., and Hrncir, J. (2014). Exploring pareto routes in multi-criteria urban bicycle routing. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1781–1787.

[Storandt, 2012] Storandt, S. (2012). Route planning for bicycles-exact constrained shortest paths made practical via contraction hierarchy. In *ICAPS*, volume 4, page 46.

[Stork, 2021] Stork, L. (2021). *Knowledge extraction from archives of natural history collections*. PhD thesis, Leiden University.

[Stroobant, 2016] Stroobant, P. (2016). Automatic generation of minimal-overlapping constrained bicycle routes. Master's thesis, Ghent University.

[Stroobant et al., 2018] Stroobant, P., Audenaert, P., Colle, D., and Pickavet, M. (2018). Generating constrained length personalized bicycle tours. *4OR*.

[Su et al., 2010] Su, J. G., Winters, M., Nunes, M., and Brauer, M. (2010). Designing a route planner to facilitate and promote cycling in metro vancouver, canada. *Transportation Research Part A: Policy and Practice*, 44(7):495 – 505.

[Szegedy et al., 2015] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

[Szeliski, 2010] Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

[Tan and Le, 2019] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

[Tensmeyer et al., 2017] Tensmeyer, C., Davis, B., Wigington, C., Lee, I., and Barrett, B. (2017). Pagenet: Page boundary extraction in historical handwritten documents. pages 59–64.

[Thirukokaranam Chandrasekar et al., 2018] Thirukokaranam Chandrasekar, K. K., Arroubai, R., Dox, G., Nop, S., Stroobant, P., Stragier, J., De Mey, K., and Verstockt, S. (2018). Runamic : dynamic generation of personalized running routes. In *Proceedings of the 6th international congress on sport sciences research and technology support, volume 1 : icSPORTS*, pages 98–105. SCITEPRESS (Science and and Technology Publications).

[Torralba et al., 2003] Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 273–, Washington, DC, USA. IEEE Computer Society.

[Triki et al., 2021] Triki, A., Bouaziz, B., Gaikwad, J., and Mahdi, W. (2021). Deep leaf: Mask r-cnn based leaf detection and segmentation from digitized herbarium specimen images. *Pattern Recognition Letters*, 150:76–83.

[Triki et al., 2020] Triki, A., Bouaziz, B., Mahdi, W., and Gaikwad, J. (2020). Objects detection from digitized herbarium specimen based on improved yolo v3. In *VISIGRAPP (4: VISAPP)*, pages 523–529.

[Turverey et al., 2010] Turverey, R. J., Cheng, D. D., Blair, O. N., Roth, J. T., Lamp, G. M., and Cogill, R. (2010). Charlottesville bike route planner. In *2010 IEEE Systems and Information Engineering Design Symposium*, pages 68–72.

[Ulrich, 1979] Ulrich, R. S. (1979). Visual landscapes and psychological well-being. *Landscape Research*, 4(1):17–23.

[Vendrig and Worring, 2002] Vendrig, J. and Worring, M. (2002). Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia*, 4(4):492–499.

[Virtanen et al., 2019] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright,

J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2019). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, page arXiv:1907.10121.

[White et al., 2020] White, A. E., Dikow, R. B., Baugh, M., Jenkins, A., and Frandsen, P. B. (2020). Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. *Applications in Plant Sciences*, 8(6):e11352.

[Wäldchen et al., 2018] Wäldchen, J., Rzanny, M., Seeland, M., and Mäder, P. (2018). Automated plant species identification—trends and future directions. *PLOS Computational Biology*, 14:e1005993.

[Zhang et al., 2015] Zhang, C., Zhou, P., Li, C., and Liu, L. (2015). A convolutional neural network for leaves recognition using data augmentation. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 2143–2150.

[Zhou et al., 2016] Zhou, B., Khosla, A., A., L., Oliva, A., and Torralba, A. (2016). Learning Deep Features for Discriminative Localization. *CVPR*.

[Zhou et al., 2017] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.