



HAL
open science

Contributions to Deep Learning for Life Science Applications

Pejman Rasti

► **To cite this version:**

Pejman Rasti. Contributions to Deep Learning for Life Science Applications. Computer Science [cs]. Université d'Angers, 2023. tel-04219846

HAL Id: tel-04219846

<https://univ-angers.hal.science/tel-04219846v1>

Submitted on 27 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Document de synthèse présenté pour obtenir L'Habilitation à Diriger des Recherches

L'UNIVERSITÉ D'ANGERS

Laboratoire Angevin de Recherche en Ingénierie des Systèmes - LARIS

Mention : Génie informatique, automatique et traitement du signal (61)

Contributions to Deep Learning for Life Science Applications

Pejman RASTI

Composition du Jury

Christian GERMAIN

Professeur, Bordeaux Sciences Agro, France

Aymeric HISTACE

Professeur, École Nationale Supérieure de l'Électronique
et de ses Applications (ENSEA), France

Daniel SAGE

Professeur, École polytechnique fédérale de Lausanne (EPFL), Switzerland

Anna KRESHUK

Professeur, European Molecular Biology Laboratory (EMBL), Germany

Laure TOUGNE RODET

Professeur, Université Lumière - Lyon 2, France

Rapporteur

Rapporteur

Rapporteur

Examinatrice

Présidente du jury

Parrain

David ROUSSEAU

Professeur, Universités d'Angers, France

Date de soutenance : le 13/07/2023

Abstract

In this manuscript, I present my research accomplishments in various domains, focusing on the development of machine learning and deep learning algorithms for analyzing images and signals. I outline my expertise in creating unique databases that have contributed significantly to research success, such as low-cost seedling growth, the 3D models of natural rosebush plants, and the AgTech data challenge. Furthermore, I detail my involvement in multimodal student behavior monitoring and a multimodal speaker recognition database.

My research has focused on contributing to machine learning and deep learning, specifically in texture-based feature extraction, developing deep learning algorithms, and overcoming challenges related to image annotation. I have explored shallow learning techniques for life science imaging, such as local binary patterns and wavelet scattering transform. In deep learning, I have developed convolutional neural network models for microscopic image analysis and MRI, as well as recurrent neural networks and long short-term memory networks for spatio-temporal images. Additionally, I have examined multimodal CNN models and devised novel techniques to tackle image annotation challenges.

My work has revolved around advancing the field of machine learning and deep learning for examining and interpreting images and signals across various domains, such as life science imaging and biometric analysis. My future research directions include minimizing dependency on manual annotation and developing novel techniques on multimodal generative self-supervised learning to extract meaningful and high-quality features from multimodal data.

Beyond developing new methodologies, my research pursuits aim to foster lifelong training initiatives, introduce new courses, and offer mini-projects and internships for young students. I intend to bridge the gap between academia and industry, promoting the exchange of ideas, resources, and expertise. Ultimately, my research plans strive to inspire and empower the next generation of researchers and innovators by fostering an environment of collaboration, education, and innovation.

Keywords: Multimodal deep learning, Life science imaging, Biometric recognition, Self-supervise learning.

Acknowledgment

I extend my most profound appreciation to everyone who has played a crucial role in my journey, leading to this remarkable achievement. Foremost, my sincere gratitude goes to the jury members: Laure TOUGNE RODET, Anna KRESHUK, Daniel SAGE, Aymeric HISTACE, and Christian GERMAIN for their thorough manuscript and work evaluation.

I am indebted to David ROUSSEAU, my post-doctoral supervisor from 2017 to 2019. His expertise helped me refine my skills and deepen my understanding. Our ongoing collaboration since 2019 has been rewarding, leading to significant advancements in our shared areas of interest. David has become a respected mentor, trusted friend, and colleague, and his family has embraced us as their own.

I also thank Rudolf KIEFER for his Ph.D. mentorship, which laid the foundation for my professional path, and Hasan DEMIREL for supervising my Master's thesis, introducing me to image processing and computer vision.

The work in this document primarily results from collaborations with my Ph.D. students Mouad, Hadhami, Lukman, Sherif, Abderrazzaq, and Mathis. I value the input of the interns and engineers involved in the research.

I am grateful to ESAIP members and directors for their consistent support. The collaborative environment fostered by the ESAIP community has contributed to my growth and provided resources and opportunities to excel. I appreciate the camaraderie and cherish the friendships and professional relationships built within the community.

Lastly, I am profoundly grateful to my family, particularly Salma, and Léo, for their unwavering love and support. Their belief in me has been invaluable in overcoming challenges and staying focused on my goals.

Contents

Introduction	4
1 Activities Digest	7
1.1 Curriculum vitae	7
1.2 Teaching and administrative activities synthesis	8
1.2.1 Teaching duties	8
1.2.2 Administrative duties	9
1.3 Research activities synthesis	10
1.4 Publication list	17
2 Materials and Databases	23
2.1 Databases	24
2.1.1 Low-cost seedling growth monitoring	24
2.1.2 RoseX - 3D models of real rosebush plants	26
2.1.3 AgTech data challenge	27
2.1.4 Multimodal student behavior monitoring	29
2.1.5 Multimodal speaker recognition database	29
3 Methodology	31
3.1 Texture-based features for shallow learning	32
3.2 Deep learning algorithms	37
3.3 Image annotation challenges	43
4 Conclusion and Future directions	49
4.1 Achievements	49
4.2 Future directions	50
4.2.1 Research	50
4.2.2 Pedagogy	55
Bibliography	57
Appendix - collection of publications	64

Introduction

The revolution of deep learning is anchored in the year 2012, a pivotal point in history, when AlexNet [1] emerged victorious in the ImageNet Large Scale Visual Recognition Challenge [2]. This marked the onset of a transformative era in machine learning. In the wake of AlexNet’s groundbreaking success, a succession of established and innovative methodologies surfaced, such as ZF Net [3] and VGGNet [4] in 2014, followed by GoogleNet [5] in 2015, and subsequently ResNet [6] in 2016, all specifically tailored for classification tasks. Concurrently, the landscape of deep learning was being expanded to encompass other tasks such as image segmentation, as exemplified by the introduction of Segnet [7] in 2015, and object detection, as illustrated by the advent of the R-CNN family [8, 9, 10] in 2014, 2015, and 2016. This pivotal period sparked a deluge of research interest, predominantly directed towards natural images and databases, particularly ImageNet [11].

However, the methodologies devised for natural images were not universally transferable to life science applications. By the year 2016, when I redirected my research focus towards adopting deep learning methods for life science applications, the deep learning field had already witnessed substantial progress. However, there were still a several scientific questions within life science applications that remained unexplored due to reliance on traditional methods in this domain. Our goal was to adopting existing deep learning strategies and to innovate new methodologies based on them to tackle these scientific questions. Challenges such as handling time-series or spatial-temporal data, common in life sciences but not typically present in natural image databases, or repurposing models trained on in vitro data for other environments like outdoor settings were some examples of these scientific questions. Another interesting scientific question is how we can leverage different data modalities during the training process to obtain maximum benefit.

On the other hand, amid the rapid growth of deep learning methodologies, a significant bottleneck was the need for annotated data, a particular challenge in the life science domain where data annotation is often required to be done by field experts. This necessity for expert-annotated data became a considerable obstacle in the progression of deep learning

methodologies for life science applications. The annotation process could vary, for instance between genotypes or could even be influenced by the experiential subjectivity of the expert performing the annotation. In my works, I embarked on a mission to address this issue. My aim was to formulate methods that could potentially accelerate this labor-intensive process, or optimally, provide a comprehensive solution for it.

My postdoctoral adventure commenced by developing deep learning models for medical and microscopic imaging, an area I continue to ardently engage in. As an active member of the Imhorphen research team, I primarily concentrate on plant phenotyping through deep learning techniques. Moreover, after joining ESAIP in January 2020, a cybersecurity-focused institution, I have been able to employ my skills in deep learning for biometric applications. My exposure to this wide range of applications has provided me with a distinct edge in my research, as I have been actively involved with various groups tackling diverse challenges in each field. This intentional choice of application domains has enabled me to establish a cohesive and interrelated set of expertise, instead of seeming disorganized or lacking focus.

Working in these interconnected fields has granted me a comprehensive understanding of the development of deep learning algorithms, considering the unique limitations and challenges each application presents. This experience has equipped me to design, manage, and adapt my research to new applications as they arise while maintaining a coherent research trajectory. As I continue to navigate the dynamic landscape of Artificial Intelligence, my carefully curated background will serve as a solid foundation for tackling the challenges and opportunities that lie ahead.

Following the comprehensive account of my research activities since earning my Ph.D., this synthesis report delves deeper into the specifics of my endeavors. The report is structured into three main sections, each dedicated to exploring different facets of my work, seamlessly connecting the overview provided earlier with a more detailed examination of my research contributions.

The first section, contained within Chapter 1, features a curriculum vitae that outlines my academic and professional achievements. Additionally, this section provides a summary of my research and teaching experiences and an exhaustive list of my published works.

The second section, including Chapter 2 and Chapter 3, delves into the theoretical and practical underpinnings of my research, highlighting contributions made by the doctoral students I have supervised and my own research endeavors. My primary research interests lie in the realm of Artificial Intelligence, with a particular focus on Machine Learning and Deep

Learning. These disciplines are categorized under CNU 61 Génie informatique, automatique et traitement du signal and section 27 Informatique. Furthermore, I have explored various application domains, such as image plant phenotyping, medical imaging, and microscopic imaging, biometric recognition, demonstrating the versatility of my research pursuits.

Chapter 2 provides an in-depth examination of the innovative systems that have been created and implemented to compile unique and original databases, especially in fields where data scarcity has been an obstacle to the progression of machine learning and deep learning algorithms.

Chapter 3 delves into the theoretical contributions of our work, tackling machine learning challenges involving various types of data, such as images, videos, voice, and text. Our investigations in medical imaging, plant phenotyping, and microscopic imaging further exemplify the framework. Furthermore, we demonstrate our endeavors in creating state-of-the-art algorithms and tools specifically tailored to streamline the image annotation process, thereby boosting the overall effectiveness of associated tasks.

The report culminates with a third and final section, Chapter 4. This section offers insights into the future directions of this report, as well as a reflection on my academic journey and potential growth opportunities.

Finally, Appendix A features a curated selection of co-authored articles that have been referenced throughout the document, showcasing the breadth and depth of my research collaborations.

Chapter 1

Activities Digest

1.1 Curriculum vitae

Name: Pejman RASTI
Date of birth: 18 September 1985
Family situation: Married, 1 child
Teaching: Department of Computer Science, ESAIP
Research laboratory: CERADE and LARIS
Tel.: +33 2 41 96 65 40
Webpage: <http://perso-laris.univ-angers.fr/~rasti/>
Email: prasti@esaip.org

Education

- 2014 - 2017 Ph.D. from the University of Tartu, Estonia.

Thesis Tittle: Analysis of Remote Sensing Image Super-Resolution using Fluid Lenses.

Supervisor: Prof. Rudolf KIEFER.

Jury members:

A. Enis CETIN, Professor, UC San Diego;
Olev MARTENS, Professor, University of Tallinn;
Väino SAMMELSELG, Professor, University of Tartu;
Alvo AABLOO, Professor, University of Tartu;
Vitaly SKACHEK, Professor, University of Tartu;

- 2012 - 2014 M.Sc. in Electrical and Electronic Engineering with a specialization in image processing and computer vision, Eastern Mediterranean University, Cyprus.

- 2009 - 2012 B.Sc. in Electrical and Electronic Engineering, Azad University, Iran.

Professional activities

- 2020 - now Teacher-Researcher at ESAIP.
- 2018 - now Qualification to the MCF by section 61 of CNU.
- 2017 - 2019 Post-doc at LARIS, Université d'Angers.
- 2015 - 2017 Lecturer at the University of Tartu the same time as my Ph.D.
- 2013 - 2014 Technical assistance in the Cisco laboratory at the Eastern Mediterranean University, Cyprus.
- 2012 - 2013 Technical assistance in the electronic laboratory at the Eastern Mediterranean University, Cyprus.

1.2 Teaching and administrative activities synthesis

1.2.1 Teaching duties

My teaching experiences have been diverse and rewarding, spanning various domains and levels of education. In mathematics, I have taught foundational and advanced courses for students pursuing their Master's degrees and those enrolled in classes préparatoire. These courses provided students with a solid understanding of mathematical concepts and equipped them with the necessary skills to excel in their chosen fields.

In computer science, I have taught various courses for Master's degree and 4th year Engineering, covering topics such as data mining, artificial intelligence, and cloud computing. These courses emphasize practical applications and foster a deep understanding of core concepts in computer science. In addition to my regular courses, I have contributed to long-life training initiatives at the Université d'Angers and EMBL, focusing on deep learning for image analysis. Furthermore, I have had the opportunity to conduct seminars in several countries, expanding my reach and sharing my expertise with a diverse international audience. These experiences have allowed me to refine my teaching techniques and adapt to the unique needs of various student populations. The list of my teaching hours and the level of the participants has been shown in table (1.1).

Table 1.1: The list of the teaching courses

Domain	Level	Subject	Type	Volume
<i>Mathematics Courses</i>				
2018-2019	Master's degree (Bac+4)	Advance Numerical Analysis	Course + TD	36h
2019-2020	Classes préparatoire 1 (Bac+1)	Numerical Analysis	Course + TD	40h
2019-2020	Classes préparatoire 1 (Bac+1)	Algebra 1	Course + TD	36h
2019-2020	Classes préparatoire 2 (Bac+2)	Algebra 2	Course + TD	36h
2019-2020	Classes préparatoire 1 (Bac+1)	Measurement techniques	Course + TD	25h
2020-2021	Classes préparatoire 1 (Bac+1)	Statistics methods	Course + TD	21h
2020-2021	Classes préparatoire 2 (Bac+2)	Algebra 2	Course + TD	18h
<i>Computer Science Courses</i>				
2020-2022	Master's degree et Ing4 (Bac+4)	Data Mining	Course + TP	125h
2020-2022	Master's degree et Ing4 (Bac+4)	Big Data	Course + TP	150h
2020-2022	4th year Engineering (Bac+4)	Business Intelligence	Course + TP	70h
2020-2022	4th year Engineering (Bac+4)	Artificial Intelligence	Course + TP	80h
2020-2022	4th year Engineering (Bac+4)	Computer Vision	Course + TP	28h
2020-2022	4th year Engineering (Bac+4)	Natural Language Processing	Course + TP	75h
2022-2023	4th year Engineering (Bac+4)	Cloud Computing	Course + TP	24h
<i>Long-Life Training (Formation Continue)</i>				
2018	Université d'Angers	Deep Learning for image analysis	Course + TP	20h
2018	Université d'Angers	Deep learning : introduction par la pratique d'applications en traitement d'images	Course + TP	20h
2019	Université d'Angers	Deep learning : introduction par la pratique d'applications en traitement d'images	Course + TP	20h
2019	EMBL, Germany	Deep Learning for image analysis	TP	35h
2020	EMBL, Germany	Deep Learning for image analysis	TP	35h
2021	EMBL, Germany	Deep Learning for image analysis	TP	25h
2022	EMBL, Germany	Deep Learning for image analysis	TP	25h
2022	Université d'Angers (ANF CNRS)	Deep Learning for microscopy image analysis	Course + TP	35h
2023	EMBL, Germany	Deep Learning for image analysis	TP	35h

1.2.2 Administrative duties

Since 2020, my administration has taken on a myriad of responsibilities at ESAIP, demonstrating our commitment to providing top-quality education in the fields of Big Data and Artificial Intelligence. As responsible for the Big Data major within the engineering cycle, my primary focus has been ensuring that our curriculum remains relevant and up-to-date in this rapidly evolving field. In 2021, I expanded my responsibility role to include the Artificial Intelligence major of the engineering cycle, which has allowed me to provide direction and guidance to students interested in pursuing careers in AI, machine learning, and related fields.

In addition to these roles, I have played a pivotal role in developing the

M.Sc. CyberSecurity and Data Science program at ESAIP. I designed the curriculum and syllabus for the courses and devised innovative pedagogical strategies to ensure that the program remains at the cutting edge of research and industry practices. This program is designed to equip students with the skills and knowledge necessary to excel in the increasingly important fields of data protection and information security.

Furthermore, as a member of the computer science department jury for new student recruitments at ESAIP, I have been actively involved in shaping the future of our institution since 2020. This role has allowed me to participate in the selection of exceptional students who demonstrate passion and aptitude for computer science, Big Data, and Artificial Intelligence. Through this collaborative effort, we have fostered an environment where students can thrive academically and professionally.

My administration's responsibilities at ESAIP have focused on developing robust programs and supporting students in pursuing knowledge in Big Data, Artificial Intelligence, CyberSecurity, and Data Science. By maintaining a rigorous, forward-looking curriculum and participating in the student recruitment process, we strive to create an academic environment that prepares students for successful careers in these rapidly growing fields.

1.3 Research activities synthesis

After completing my Ph.D. at the University of Tartu in May 2017, I joined the Imhorphen research group at the Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS) within the University of Angers. Under the guidance of Prof. David Rousseau, I embarked on my postdoctoral research focused on creating and applying machine learning and deep learning algorithms to address challenges in life sciences, such as low-cost plant phenotyping through imaging, as well as medical and microscopy imaging.

After a three-year postdoctoral tenure at the University of Angers, I transitioned to a dual role as a teacher-researcher (*enseignant-chercheur*) and head of the AI and Big Data majors at ESAIP (*École d'Ingénieurs en Informatique*) in January 2020. As a teacher-researcher at ESAIP, I initiated independent research to develop multimodal deep learning algorithms for biometrics and emotion recognition. My research at the ESAIP research center (CERADE), in close collaboration with LARIS and the Imhorphen research group, continued to thrive.

Recently, I have successfully authored and secured funding for three project proposals in this area, receiving support from regional institutions such as PULSAR and Angers Loire Metropole (ALM) and the international TUBITAK in Turkey. The total funding obtained for these projects

amounts to 175,000 euros. Further details about these projects will be provided in the perspective section.

Research topics

Theoretical keywords:

Machine Learning, Deep Learning, Adversarial Learning, Self-Supervised Learning.

Application fields:

Deep Learning applied to: Image Plant Phenotyping, Medical Imaging, Microscopic Images, Biometrics Recognition.

Production

Pulications	Numbers
Journal Articles (Q1 and Q2 of scopus)	18
International peer-reviewed conferences	21
National peer-reviewed conferences	3
Book chapter	2
Book	2

Supervision

supervision	Numbers
Ph.D. (defended)	2
Ph.D. (on going)	4
Master internship (Defended)	5
Master internship (on going)	2
Internship follow-up (tutor of pedagogy)	13

Ph.D. Supervision

- **Defended Thesis:** Mouad ZINE EL ABIDINE (September 2019 – October 2022)

Thesis title: Contributions to computer vision and machine learning for plant variety testing.

Constitution of the jury: J.-P. DA COSTA (Rapporteur),
A. HAFIANE (Rapporteur),
M. ZUDE-SASSE (Examineur),

G. BUCK SORLIN (Examineur),
M.-J. ARANZANA (Examineur).

Supervisors: D. ROUSSEAU (50%), **P. RASTI** (30%), and
H. DUTAGACI (20%).

Funding: Connect Talent.

Publication during the thesis: One journal article [12] and
one international proceeding.

- **Defended Thesis:** Hadhami GARBOUGE (November 2019 –
November 2022)

Thesis title: Deep learning applied to multi-component
imagery for variety testing problems.

Constitution of the jury: C. GERMAIN (Rapporteur),
F. COINTAULT (Rapporteur),
J. BUITINK (Examineur),
P. VERMEULEN (Invited),
P. ROUMET (Invited).

Supervisors: D. ROUSSEAU (50%), **P. RASTI** (30%), and
N. SAPOUKHINA (20%).

Funding: H2020 INVITE European Project.

Publication during the thesis: One journal article [13] and
three proceedings [14, 15, 16].

- **In progress Thesis:** Lukman E. ISMAIL (The defense is scheduled
for July 2023)

Thesis title: Machine learning application in neuroscience for
neurosurgical brain tumor resection procedure.

Supervisors: D. ROUSSEAU (50%), **P. RASTI** (30%), and
J.-M. LEMEE (20%).

Funding: Petroleum Technology Development Fund (PTDF).

Publication during the thesis: One journal article [17] and
three proceedings [18, 19, 20].

- **In progress Thesis:** Sherif HAMDY (beginning of the thesis in
September 2020 - on a part-time basis)

The provisional title of the thesis: Computer vision and artificial intelligence applied to the analysis of the physical quality of seeds.

Supervisors: D. ROUSSEAU (50%), **P. RASTI** (25%), and A. CHARRIER (25%).

Funding: Groupe d'Etude et de contrôle des Variétés Et des Semences (GEVES).

Current situation of the doctoral student: Registered in the third year; one published book chapter [21].

- **In progress Thesis:** Abderrazzaq MOUFIDI (beginning of the thesis in October 2021)

The provisional title of the thesis: Voice biometrics using natural language processing.

Supervisors: D. ROUSSEAU (50%) and **P. RASTI** (50%).

Funding: Angers Loire Metropole.

Current situation of the doctoral student: Registered in the second year, with one journal article [22] and one proceeding [23].

- **In progress Thesis:** Mathis CORDIER (beginning of the thesis in September 2021- CIFRE)

The provisional title of the thesis: Embedded analysis of RGB-Depth images, application to the study of the development of disease symptoms in plants.

Supervisors: D.ROUSSEAU (50%), **P. RASTI** (30%), and Cindy TORRES (20%).

Funding: Vilmorin-Mikado.

Current situation of the doctoral student: Registered in the second year, with one proceeding [24].

Internship in Master 2 (6 months):

- Duy Khong THANH, “Real-Time Engagement Analysis of students based on Classroom Live Videos”, 2022-2023 (INSA Centre Val de Loire). Responsible for the internship: **P. RASTI**.

- Hugo VOYNEAU, “Real-Time Engagement Analysis of students based on Classroom”, 2022-2023 (Université d’Angers). Responsible for the internship: **P. RASTI** and Delphine GUEDAT-BITTIGHOFFER
- Amine MANGACHE, “Emotion Recognition of students based on ClassroomVideos”, 2021-2022 (UCO, Angers). Responsible for the internship: **P. RASTI**.
- Kholoud GHANMI,” Classifications des images de microscopie à balayage de foraminifères benthiques actuels”, 2021-2022(l’université de Rennes 1). Responsible for the internship: **P. RASTI** and E. BICHI.
- Abderrazzaq MOUFIDI,” Voice biometrics using natural language processing. ”, 2020-2021 (ENS Rennes). Responsible for the internship: **P. RASTI** and D. ROUSSEAU.
- Xareni GALINDO,” Image Segmentation and Object Recognition for the Detection of Apples Using Low Cost Image Acquisition Equipment.” , 2018-2019(l’université de Saint-Etienne). Responsible for the internship: D. ROUSSEAU and **P. RASTI**
- Tõnis UIBOUPIN,” Super Resolution and Face Recognition Based People Activity Monitoring Enhancement Using Surveillance Camera.” , 2016-2017(University of Tartu, Estonia). Responsible for the internship: **P. RASTI**

Other academical activities synthesis

In addition to my primary academic pursuits, I have been actively involved in a variety of other academical activities, all of which have contributed to developing my skills and expertise. Among my numerous roles, I have served as a reviewer for various research funding organizations, including **Agence Nationale de la Recherche (ANR)**, **Science Foundation Ireland (SFI)**, and **German Research Foundation (DFG)** projects. This experience has allowed me to evaluate and contribute to the progress of cutting-edge research across multiple disciplines.

Moreover, I am an editor for the Journal of Environmental Science, Frontier, specifically within the AI section. This role further enhances my understanding of the latest advances in AI applied to environmental science and enables me to contribute to disseminating important research in this area.

Furthermore, I have been a reviewer for over 20 journals and conferences, which has expanded my knowledge base and honed my critical thinking and assessment abilities. My involvement in organizing academic events has also been considerable, as I have been a member of the local organizing

committee for the international SampTA 2017 conference held in Estonia. In addition, I have volunteered at five international conferences, including ICGIP (from 2019 to 2021) and IPRIA (2021 and 2023), showcasing my commitment to supporting and fostering collaboration within the academic community.

Moreover, I acted as a trainer on four separate occasions at EMBL for the course "Deep Learning for Image Analysis" between 2018 and 2021. This opportunity has allowed me to share my expertise and inspire others in the field while refining my teaching and communication skills. Lastly, as a member of the Scientific Organisers Committee, I have contributed to the organization of the same course twice at EMBL, in 2022 and 2023. This role has further developed my organizational skills and my ability to collaborate with a diverse team of professionals.

Projects in progress

- Voice biometrics using NLP
Duration: 3 years (15/10/2021 – 15/10/2024)
Funding: 96 000 EUR
Investigator: **Pejman RASTI**
Funder: Angers Loire metropole (ALM)
- Real-Time engagement analysis of students based on live classroom videos
Duration: 2 years (01/09/2021 – 31/08/2023)
Funding: 46 000 EUR
Investigator: **Pejman RASTI**
Funder: PULSAR – Pays de la Loire
- Sentiment analysis using NLP for renewable energies
Duration: 2 years (01/09/2022 – 31/08/2024)
Funding: ~34 000 EUR (700000 Turkish Lira)
Investigator: Ceren ÇUBUKCU (Gebze Technical University, Turkey)
and **Pejman RASTI**
Funder: TUBITAK (National research council of Turkey) via Gebze Technical University
- Competence Development in Collaborative Industrial Internet of Things (CDC-IoT)
Duration: 3 years (01/09/2020 – 31/08/2023)
Funding: 286 242,00 EUR

Investigator: SeAMK University, ESAIP - **Pejman RASTI**,
TalTech University, Riga Technical University, and
Kaunas University of Technology

Funder: Erasmus+

1.4 Publication list

JOURNAL ARTICLES (since 2017)

Q1 (SJR INDICATOR)

1. Abderrazzaq Moufidi, David Rousseau, and **Pejman Rasti**. “Attention based Fusion of Ultra-short Voice Utterances and Depth Videos for Multimodal Person Identification”, *Sensors* (2023) – Under Review.
2. Mouad Zine El Abidine, Helin Dutagaci, **Pejman Rasti**, Maria-Jose Aranzana, Christian Dujak and David Rousseau. “Toward objective variety testing score based on computer vision and unsupervised machine learning Application to Apple Shape”, *Biosystems Engineering* (2023) – Under Review.
3. Mathis Cordier , Torres Cindy, **Pejman Rasti**, and David Rousseau. “On the use of circadian cycles to monitor individual young plants”, *Remote Sensing* (2023) .
4. Hadhami Garbougé, **Pejman Rasti**, and David Rousseau. “Enhancing the tracking of seedling growth Using RGB-Depth Fusion and Deep Learning”, *Sensors*, vol. 21, no. 24, p. 8425, (2021).
5. Salma Samiei, **Pejman Rasti**, Joseph Ly Vu, Julia Buitink, and David Rousseau. “Deep learning-based detection of seedling development”, *Plant Methods*, vol. 16, no. 1, p. 1–11, (2020).
6. Salma Samiei, **Pejman Rasti**, Paul Richard, Gilles Galopin, and David Rousseau. “Toward joint acquisition-annotation of images with egocentric devices for a lower-cost machine learning application to apple detection”, *Sensors*, vol. 10, no. 12, p. 666, (2020).
7. Helin Dutagaci, **Pejman Rasti**, Gilles Galopin, and David Rousseau. “ROSE-X: an annotated data set for evaluation of 3D plant organ segmentation methods”, *Plant methods*, vol. 16, no. 1, p. 1,(2020).
8. Noëlie Debs, **Pejman Rasti**, Tae-Hee Cho, Carole Frindel, and David Rousseau. “Simulated perfusion MRI data to boost training of convolutional neural networks for lesion fate prediction in acute stroke”, *Computers in biology and medicine*, vol. 116, p. 103579, (2020).
9. **Pejman Rasti**, Christian Wolf, Hugo Dorez, Raphael Sablong, Driffa Moussata, Salma Samiei, and David Rousseau. “Machine learning based classification of the health state of mice colon in cancer study from confocal laser endomicroscopy”, *Scientific Report*, vol. 9, p. 20010, (2019).

10. **Pejman Rasti**, Ali Ahmad, Salma Samiei, Etienne Belin, and David Rousseau. “Supervised image classification by scattering transform with application to weed detection in culture crops of high density”, *Remote Sensing*, vol. 11, no. 3, p. 249, (2019).
11. Mathilde Giacalone, **Pejman Rasti**, Noelie Debs, Carole Frindel, Tae-Hee Chol, Emmanuel Grenier, and David Rousseau. “Local spatio-temporal encoding of raw perfusion MRI for the prediction of final lesion in stroke”, *Medical Image Analysis*, vol. 50, p. 117 (2018).
12. Zane Zondaka, Madis Harjo, Mahdi Safaei Khorram, **Pejman Rasti**, Tarmo Tamm, and Rudolf Kiefer. “Polypyrrole / carbide-derived carbon composite in organic electrolyte: Characterization as a linear actuator”, *Reactive and Functional Polymers*, vol. 131, p. 414, (2018).
13. Alo Kivilo, Zane Zondaka, Arco Kesküla, **Pejman Rasti**, Tarmo Tamm, and Rudolf Kiefer. “Electro-chemo-mechanical deformation properties of polypyrrole linear actuators in aqueous and organic electrolyte”, *RSC Advances*, vol. 6, P. 99, (2017).

Q2 (SJR INDICATOR)

14. Lukman E Ismaila, **Pejman Rasti**, Florian Bernard, Mathieu Labriffe, Philippe Menei, Aram Ter Minassian, David Rousseau, and Jean-Michel Lemée. “Transfer learning from healthy to unhealthy patients for the automated classification of functional brain networks in fMRI”, *Applied Sciences*, vol. 12, no. 14, p. 6925, (2022).
15. **Pejman Rasti**, Bolotnikova Anastasia, Daneshmand Morteza, and Ozcinar Cagri. “Optimal image compression via block-based adaptive colour reduction with minimal contour effect”, *Multimedia Tools and Applications* vol. 79, p. 19 (2020).
16. Salma Samiei, **Pejman Rasti**, Hervé Daniel, Etienne Belin, Paul Richard, and David Rousseau. “Toward a computer vision perspective on the visual impact of vegetation in symmetries of urban environments”, *Symmetry*, vol. 10, p. 12vol. 10, no. 12, p. 666, (2018).
17. **Pejman Rasti**, Kamal Nasrollahi, Olga Orlova, Gert Tamberg, and Thomas B Moeslund. “Reducible dictionaries for single image super-resolution based on patch matching and mean shifting”, *Journal of Electronic Imaging*, vol. 12, p. 2 (2017).
18. **Pejman Rasti**, Salma Samiei, Mary Agoyi, and Sergio Escalera. “Robust non-blind color video watermarking using QR decomposition and

entropy analysis”, *Journal of Visual Communication and Image Representation*, vol. 50, p. 127 (2017).

CONFERENCE PROCEEDINGS

COMPUTER SCIENCE CONFERENCES

19. Cerasi, Ceren Cubukcu, Yavuz Selim Balcioglu, Farid Huseynov, and **Pejman Rasti**. ”A Sentiment Analysis to Understand the Role of Twitter Towards Sustainable Consumption.” In: 27th International Conference on Information Technology (IT), IEEE, 2023.
20. Lukman E Ismaila, **Pejman Rasti**, Jean-Michel Lemée, and David Rousseau. ”Toward more frugal models for functional cerebral networks automatic recognition with resting state fMRI” In: 29ème Colloque GRETSI sur Le Traitement du Signal et des Images. 2022.
21. Abderrazzaq Moufidi, David Rousseau, and **Pejman Rasti**. “Wavelet scattering transform depth benefit, an application for speaker identification”. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition. Springer. 2023.
22. Lukman E Ismaila, **Pejman Rasti**, Jean-Michel Lemée, and David Rousseau. “Self-supervised learning for functional brain networks identification in fmri from healthy to unhealthy patients” In: International conference on signal image technology internet based systems. IEEE. 2022.
23. Mathis Cordier, **Pejman Rasti**, Cindy Torres, and David Rousseau. ”Optimisation de l’échelle d’observation pour l’annotation d’images.” In: 28ème Colloque GRETSI sur Le Traitement du Signal et des Images. 2022.
24. Hadhami Garbougé, **Pejman Rasti**, and David Rousseau. “Deep learning-based detection of seedling development from indoor to outdoor”. In: International Conference On Systems, Signals And Image Processing. IEEE. 2021.
25. Mouad Zine El Abidine, Sabine Merdinoglu-Wiedemann, **Pejman Rasti**, Helin Dutagaci, and David Rousseau. “Machine learning-based classification of powdery mildew severity on melon leaves”. In: International Conference on Image and Signal Processing. Springer. 2020.
26. Natalia Sapoukhina, Salma Samiei, **Pejman Rasti**, and David Rousseau. “Data augmentation from RGB to chlorophyll fluorescence imaging application to leaf segmentation of Arabidopsis thaliana from top view

- images”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPR). 2019.
27. Salma Samiei, **Pejman Rasti**, François Chapeau-Blondeau, and David Rousseau. ”Cultivons notre jardin avec Fourier.” In: 27ème Colloque GRETSI sur Le Traitement du Signal et des Images. 2019.
 28. **Pejman Rasti**, Didier Demilly, Landry Benoit, Etienne Belin, Sylvie Ducournau, Francois Chapeau-Blondeau, and David Rousseau. “Low-cost vision machine for high-throughput automated monitoring of heterotrophic seedling growth on wet paper support”. In: British Machine Vision Conference (BMVC). 2018.
 29. Salma Samiei, Ali Ahmad, **Pejman Rasti**, Etienne Belin, and David Rousseau. ”Low-cost image annotation for supervised machine learning. Application to the detection of weeds in dense culture.” In British Machine Vision Conference (BMVC), 2018.
 30. Denis Bujoreanu, **Pejman Rasti**, and David Rousseau. “On the value of graph-based segmentation for the analysis of structural networks in life sciences”. In: Proceedings of the 25th European Signal Processing Conference (EUSIPCO). IEEE. 2017.
 31. **Pejman Rasti**, Olga Orlova, Gert Tamberg, Cagri Ozcinar, Kamal Nasrollahi, and Thomas B Moeslund. “Improved interpolation kernels for super-resolution algorithms”. In: International Conference on Image Processing Theory, Tools and Applications. IEEE. 2016.
 32. **Pejman Rasti**, Gholamreza Anbarjafari, and Hasan Demirel. “Colour image watermarking based on wavelet and QR decomposition”. In: Signal Processing and Communications Applications Conference (SIU). IEEE. 2017.
 33. **Pejman Rasti**, Tõnis Uiboupin, and Sergio Escalera. “Convolutional neural network super Resolution for face recognition in surveillance monitoring”. In: International Conference on Articulated Motion and Deformable Objects. Springer. 2016.
 34. Marco Bellantonio, Kamal Nasrollahi, Sergio Escarela, Thomas B Moeslund, and **Pejman Rasti**. “Spatio-temporal pain recognition in cnn-based super-resolved facial images”. In: International Conference on Pattern Recognition (ICPR). Springer. 2016.
 35. **Pejman Rasti**, Morteza Daneshmand, Fatih Alisinanoglu, and Cagri Ozcinar. “Medical image illumination enhancement and sharpening by using stationary wavelet transform”. In: Signal Processing and Communication Application Conference (SIU). IEEE. 2016.

36. Tõnis Uiboupin, **Pejman Rasti**, and Hasan Demirel. “Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring”. In: Signal Processing and Communication Application Conference (SIU). IEEE. 2016.
37. Christer Loob, **Pejman Rasti**, Iris Lüsi, Julio CS Jacques Junior, Xavier Baró, Sergio Escalera, Tomasz Sapinski, and Dorota Kamin-ska. “Dominant and complementary multi-emotional facial expression recognition using C-support vector Classification”. In: Automatic face and gesture recognition (FG 2017), IEEE, 2017.

LIFE SCIENCE CONFERENCES

38. Hadhami Garbougé, **Pejman Rasti**, and David Rousseau. “Machine learning assisted determination of best acquisition protocols in variety testing”. In: AI for Agriculture and Food Systems. 2021.
39. Mathis Cordier, Hadhami Garbougé, Salma Samiei, **Pejman Rasti**, and David Rousseau. “Growth-data a new tool to characterize spatio-spectral patterns of plant growth”. In: North American Plant Phenotyping Network Annual Conference. Springer. 2021.
40. **Pejman Rasti**, Rosa Huaman, Charlotte Riviere, and David Rousseau. “Supervised machine learning for 3d microscopy without manual annotation: application to spheroids”. In: SPIE Photonics Europe 2018. SPIE. 2018.
41. **Pejman Rasti**, Ali Ahmad, Etienne Belin, and David Rousseau. “Learning on deep network without the hot air by scattering transform application to weed detection in dense culture”. In: International Workshop on Image Analysis Methods in the Plant Sciences (IAMPS). 2018.
42. **Pejman Rasti**, Etienne Belin, and David Rousseau. “A computer vision tool for a high- throughput phenotyping of seedlings during elongation - Application to sugar beet”. In: 76th conference of the International Institute for Research on Sugar Beet (IIRB). 2018.

BOOK CHAPTERS

41. Sherif Hamdy, **Pejman Rasti**, Aurelie Charrier, David Rousseau. “Advances in seed phenotyping and applications to seed testing/monitoring and breeding ; Focus on seed phenotyping with X-Ray imaging”. In: Advances in seed science and technology for more sustainable crop production. Burleigh Dodds Science Publishing. 2022.

42. **Pejman Rasti**, Morteza Daneshmand, and Cagri Ozcinar. “Resolution enhancement based image compression technique using singular value decomposition and wavelet transforms”. In: *Wavelet Transform and Some of Its Real-World Applications*, IntechOpen. 2015.

BOOKs

43. Gholamreza Anbarjafari, **Pejman Rasti**, Fatemeh Noroozi, Jelena Gorbova, and Rain Eric Haamer. “Machine learning for face, emotion, and pain recognition”. International Society for Optics and Photonics, 2018.
44. Gholamreza Anbarjafari and **Pejman Rasti**. “Illumination enhancement: image and video”. LAP Lambert Academic Publishing, 2016.

Chapter 2

Materials and Databases

Over the last five years, my endeavors have extended beyond methodological and applied research to the establishment and execution of innovative systems for amassing unique and original databases. As a researcher grounded in engineering, I have been proactive in transcending the boundaries of merely employing existing online databases for crafting machine learning and deep learning algorithms. This has been particularly relevant in domains where data paucity obstructed the progression of novel machine learning and deep learning algorithms.

To accomplish this, I have actively engaged in collaborations with experts from diverse fields to pinpoint the specific data requisites and demands for each research domain. This cross-disciplinary approach has facilitated a deeper understanding of the challenges encountered by different research areas and subsequently informed the design and creation of data collection systems tailored to address those particular needs. Utilizing state-of-the-art technologies and methodologies, I have succeeded in generating databases that are both exhaustive and precise, laying a robust foundation for the formulation of new machine learning and deep learning algorithms.

The databases I have devised not only cater to my research requirements but also serve as vital assets for numerous Ph.D. students and fellow team members. They have used the data to develop pioneering deep learning techniques and propel their interdisciplinary projects. In this manner, my data collection efforts have substantially contributed to our research group's overall accomplishments and impact.

Furthermore, developing these databases has enabled me to acquire diverse skills related to data procurement, management, and analysis. This expertise has become invaluable in steering the design and implementation of subsequent research initiatives, ensuring the collected data is of the utmost quality and best suited for our research objectives. As we persist in exploring new possibilities for machine learning and deep learning applications, the capacity to efficiently and effectively gather and manage data

will remain an essential component of our ongoing achievements.

2.1 Databases

2.1.1 Low-cost seedling growth monitoring

As a starting point, we created a network of 60 Raspberry Pi devices and cameras in partnership with INRAE Angers, as shown in Fig. (2.1). This infrastructure was designed to automatically gather and transfer pictures of plant growth at different intervals, both during the day and at night [25]. The technology developed within this infrastructure was used to collect data for two Ph.D. studies of Salma Samiei and Hadhami Garbougé. After the success of the initial RGB imaging system (Fig. (2.1-left)), depth cameras were added as an enhancement to enrich the system and enable it to capture images during the night or under different lighting conditions, providing more versatility and improving the overall performance (Fig. (2.1-right))[13].

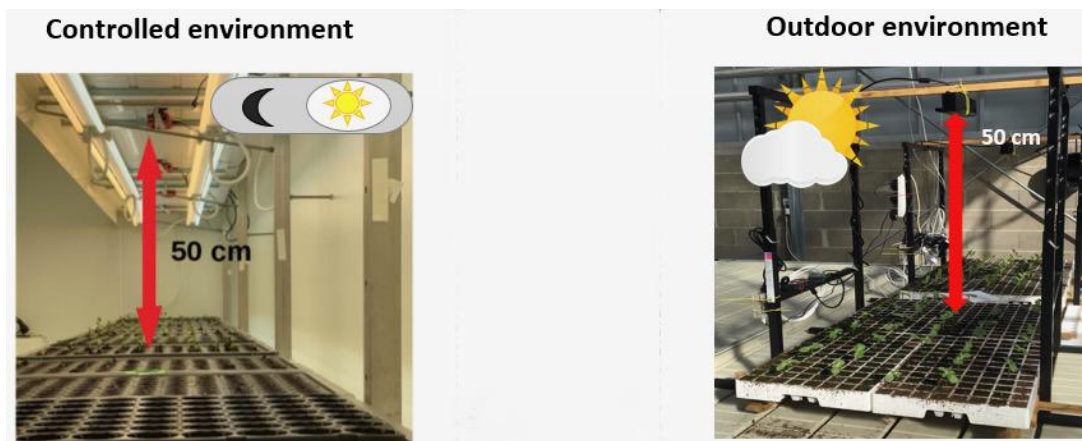


Figure 2.1: Left: RGB imaging system installed in a growth chamber (controlled environment). Right: RGB + depth imaging system installed in a green house (Outdoor environment) [13].

For each experiment, which lasted almost two weeks, images were captured with a time-lapse of around 15 minutes. In total, the database consists of almost 42,000 spatio-temporal sequences of RGB images, where each temporal sequence comprises nearly 768 individual images. This extensive collection of images has been crucial for the development of our deep learning techniques and models. Fig. (2.2) shows an example of collected data by the imaging system.

This innovative imaging system and database, specifically tailored for plant phenotyping, features a diverse range of seedlings at various growth stages and under different environmental conditions. The comprehensive nature of the database allows researchers to train deep learning algorithms

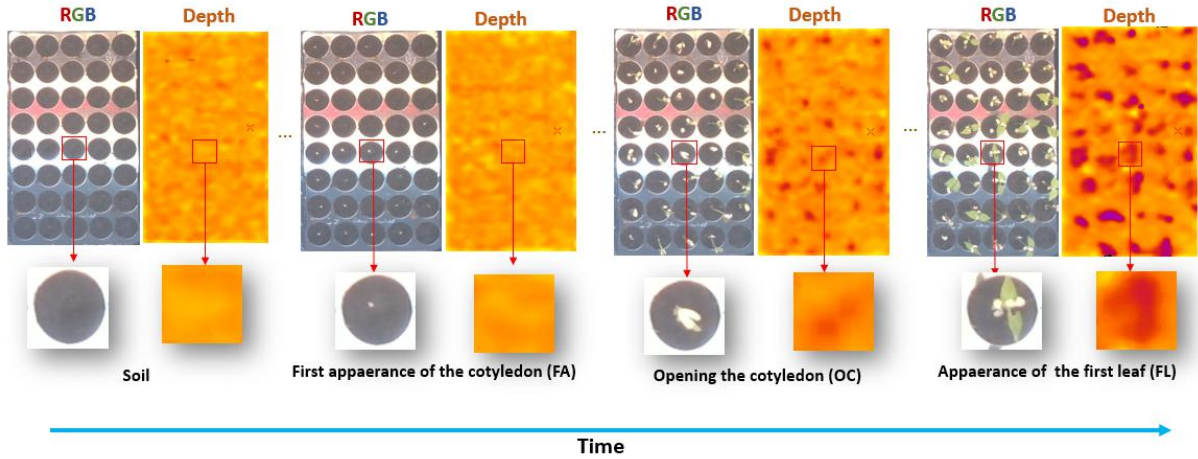


Figure 2.2: An overview of the time-lapse collected for our works. Upper row, view of RGB and depth of a full tray with 40 pots from the top view. Lower row, a zoom on a single pot at each stage of development to be detected from left to right: soil, the first appearance of the cotyledon (FA), opening the cotyledons (OC) and appearance of the first leaf (FL) [13].

on a wide variety of plant images, ultimately improving the accuracy and reliability of the algorithms for detecting and tracking plant growth and development.

The technology is now being used as the focus of my current Ph.D. student’s research, Mathis Cordier, at the company Vilmorin-Mikado for further research and analysis of images. The Raspberry Pi devices and cameras were designed to communicate seamlessly with one another, facilitating the acquisition of images from multiple locations. By having access to this large and diverse dataset, we were able to train our developed deep learning techniques to accurately categorize and scrutinize the images, which holds the potential to enhance our comprehension of plant growth and development.

One of the key advantages of the proposed imaging system and database is its potential for automating the seedling phenotyping process. Traditional phenotyping methods are labor-intensive, time-consuming, and prone to subjective biases. By leveraging the power of deep learning and the high-resolution imaging data collected by the Raspberry Pi network, a system capable of providing rapid, objective, and accurate assessments of seedling growth and development was developed.

Furthermore, the potential for the imaging system and database to be applied to other areas of plant research is highlighted, such as disease detection, stress response analysis, and genotype-phenotype mapping. The flexibility and scalability of the system make it suitable for a wide range of applications in plant science and agriculture, where high-throughput and accurate phenotyping methods are critical for driving advances in crop breeding and management [25, 13].

2.1.2 RoseX - 3D models of real rosebush plants

Within the scope of an internal project by the Imhorphen group, in collaboration with Dr. Helin Dutagaci, we introduced an open-source collection of comprehensive 3D models of real rosebush plants, complete with ground truth annotations at the organ level [26]. These models were obtained using a Siemens 3D X-ray imaging system with a voltage range of 10-450 kV, employing a tungsten transmission target and a 280-mA current. For this research, the system operated at an 80-kV voltage. With 900 projections, each radiograph is an average of three exposures lasting 333 ms to minimize noise. The acquisition time for each plant was 20 minutes, and 11 rosebush plants of varying architectural complexity were imaged.

The acquired data consists of a series of X-ray images with a pixel spacing of 0.9766 mm and a slice spacing of 0.5 mm, forming a 3D voxel space. The intensity of each voxel corresponds to the plant shoot’s material properties at that point. To extract the 3D voxels of the rosebushes and their pots from the raw data, masking and thresholding methods were employed. Manually created masks helped remove unrelated materials from the imaging platform, while thresholding differentiated plant voxels from air.

The remaining voxels were assigned to one of the following categories: (1) stem, (2) leaf, (3) flower, (4) pot, (5) tag. The stem category includes main branches and petioles due to their similar geometric structures and spatial connections. Fig. (2.3) illustrates the thresholded X-ray volume (a), organ-level labels obtained through annotation (b), labels corresponding to the plant shoot (c), and the stem and petiole structure (d) of a sample rosebush model from the dataset.

Manual annotation was performed using ilastik (Interactive Learning and Segmentation Toolkit) [27]. With ilastik’s pixel classification tool, we manually marked several voxels in regions belonging to each class on a rosebush model to train the classifier. We then obtained full-volume predictions for all models generated by ilastik’s trained classifier and manually corrected any inaccurately labeled voxels.

The dataset is available online at [28] and provided in various forms: (1) the raw X-ray image stack, (2) the binary volume mask indicating the voxels of only the plant shoot, tag, and pot, along with corresponding organ-level labels, (3) the binary volume mask indicating the voxels only on the surface of the plant shoot and corresponding organ-level labels, (4) the point cloud consisting of the points of the plant shoot, tag, and pot with colors indicating organ-level labels, (5) the point cloud consisting of the points on the surface of the plant shoot with colors indicating organ-level labels. Additional file 1 contains detailed information on file formats and

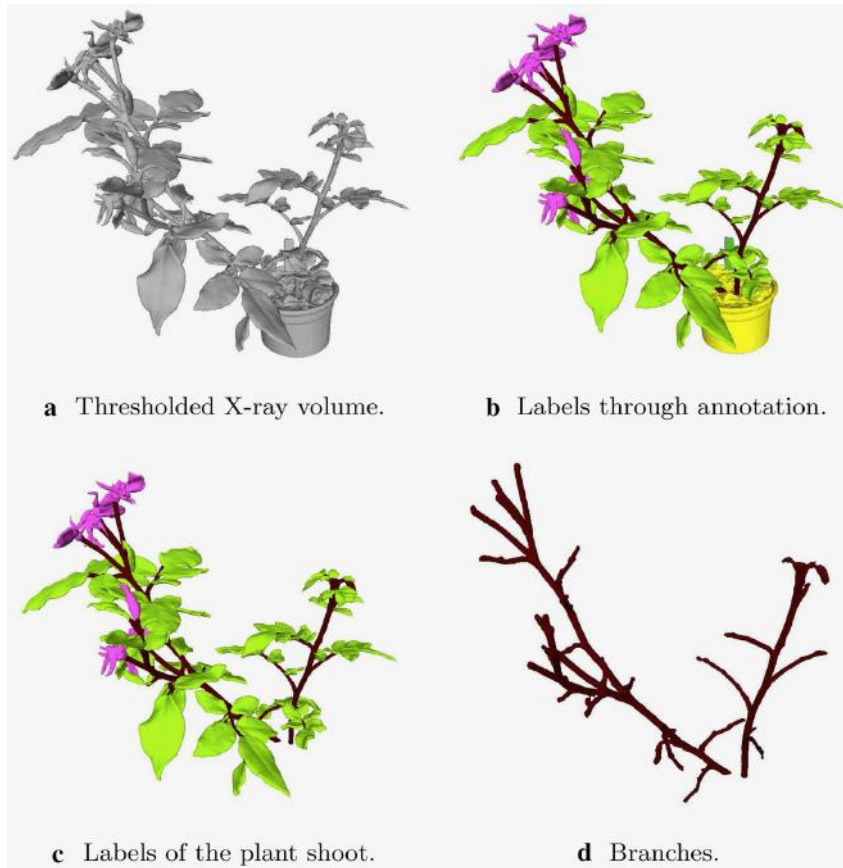


Figure 2.3: A sample rosebush model from the data set. The raw X-ray volume is thresholded and masked to obtain the solid part shown in a. Each voxel in the volume is annotated as leaf, stem, flower, pot, or tag to obtain the ground-truth segmentation as shown in b. In c only the parts corresponding to the plant shoot are shown, excluding the pot and the tag. The voxels corresponding only to stem class are shown in d [26].

label data. Using these resources, it is possible to convert 3D volumetric models into labeled polygon mesh models and obtain 3D point clouds as viewed from any position around the plant through ray casting.

2.1.3 AgTech data challenge

In 2019, we organized a worldwide online data challenge that attracted participation from approximately 200 individuals across the globe. This ambitious project was a collaborative effort between three institutions – the University of Angers, ESEO, and ESA – and was generously supported by companies such as Business and Decision and Credit Agricole bank. The primary objective of this project was to develop machine learning or deep learning algorithms capable of being trained on synthetic data and then detecting weeds in challenging scenarios using real images.

To facilitate this, we created a simulation tool to generate realistic synthetic weeds in densely populated environments where the weeds are surrounded by plants [29] shown in Fig. (2.4). This innovative approach allowed us to create a large and diverse dataset of synthetic images, which

could be used to train deep learning algorithms for the task of weed detection. The use of synthetic data in this context has the potential to overcome some of the limitations of traditional data collection methods, such as labor-intensive manual annotation and the need for large amounts of real-world data.

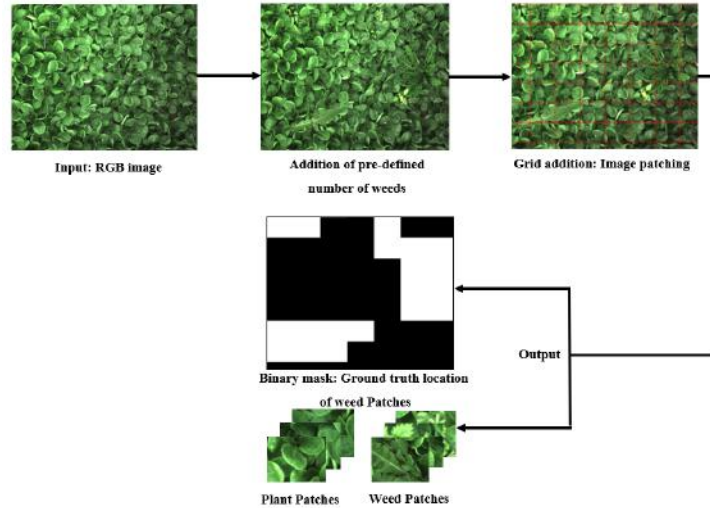


Figure 2.4: Simulation pipeline for the creation of images of weeds in densely populated environments [29].

Throughout the data challenge, we organized online and in-person training sessions to instruct trainers on deep learning methods that can be applied to image classification for these types of images. These sessions provided participants with the necessary knowledge and skills to develop their own deep learning algorithms and adapt them to the specific requirements of the weed detection task. By sharing our expertise and fostering collaboration among participants, we aimed to promote innovation and the development of new techniques for image-based weed detection.

The data challenge served as a platform for researchers and practitioners from various fields to come together, share their ideas, and work towards a common goal. This collaborative environment facilitated the exchange of knowledge and expertise, leading to the development of innovative solutions for weed detection using synthetic data. The resulting deep learning algorithms showcased the potential of synthetic data in training models that can generalize well to real-world scenarios, opening up new avenues for research and applications in the field of agriculture and beyond.

Through the organization of the data challenge and the development of innovative solutions for weed detection, we have demonstrated the power of collaboration and the potential of synthetic data in advancing the state of the art in deep learning and its applications. As we continue to explore new challenges and opportunities in this field, we remain committed to

fostering innovation and promoting the use of cutting-edge techniques to address real-world problems.

2.1.4 Multimodal student behavior monitoring

For my second work, which is a cross-disciplinary project aimed at utilizing computer vision and deep learning for educational purposes, funded by the Pulsar project, we are creating a sophisticated system network made up of cameras as well as oximeters to record the students' status during an in-person course visually and by measuring their heart rate variation (HRV). This innovative approach allows for the collection of a large dataset that can be used to analyze the students' moods and emotions during a single session or multiple sessions of the course, ultimately aiming to improve the course's content and pedagogical strategies.

The integration of cameras and oximeters in the classroom environment provides an opportunity for continuous monitoring of students' facial expressions and physiological signals. The cameras capture high-resolution images and videos of the students, while the oximeters measure their HRV, providing valuable insights into their emotional states. By combining these two modalities, we can obtain a more comprehensive understanding of students' reactions and engagement levels during the course.

The potential applications of this work extend beyond the classroom setting, with potential use in various educational contexts such as online learning, tutoring, and training programs. By gaining insights into students' emotions and engagement levels, educators can tailor their teaching methods and course materials to better meet the needs and preferences of their students, resulting in improved learning outcomes and overall student satisfaction.

2.1.5 Multimodal speaker recognition database

In the context of my current project ALM, we are breaking new ground by creating a database specifically for biometric recognition using voice (speaker recognition) in the presence of both environmental noise and human emotions. Most existing datasets for this purpose only include one or two modalities and do not contain any environmental noise [30, 31]. Our innovative database is created by recording volunteers in three modalities: voice, RGB video, and depth video. This comprehensive approach allows for a more accurate analysis and understanding of speaker recognition under various conditions.

During our data collection process, volunteers read a set of sentences while expressing eight different emotions: anger, anticipation, joy, trust,

fear, surprise, sadness, and disgust. We collect data in two distinct environments, starting with a controlled, noise-free environment, and then moving to an environment with background noise, such as streets, parks, or parties. This approach ensures that our database captures a diverse range of scenarios, enabling the development of more robust speaker recognition systems.

The comprehensive nature of our database has the potential to have a significant impact on biometric recognition systems by including various challenges and enhancing their performance under real-world conditions. This database will be used by Ph.D. student Abderrazzaq Moufidi, who is working on multimodal deep learning algorithms for speaker recognition. By leveraging the rich information contained in our database, Moufidi aims to develop new techniques and models that can accurately identify speakers in the presence of noise and emotional variability.

One of the main challenges we face in this project is ensuring the quality and consistency of the collected data. To address this issue, we have implemented strict protocols for data collection, ensuring that the volunteers follow the same guidelines regarding their emotional expressions and the reading of sentences. Additionally, we are using high-tech equipment for capturing voice, RGB video, and depth video, ensuring that the collected data is of the highest quality and suitable for deep learning algorithms.

The potential applications of our database and the resulting deep learning algorithms extend beyond biometric recognition. The ability to accurately identify speakers in noisy environments and under various emotional states can have a significant impact on industries such as security, customer service, and telecommunication. Furthermore, our multimodal approach can be applied to other fields where understanding human behavior and emotions is essential, including healthcare, education, and human-computer interaction.

Chapter 3

Methodology

My research methodology is primarily guided by a bottom-up strategy, where I commence with a specific application and progress toward the methodology. This approach has proven to be highly effective in developing novel methods that can be employed across a wide range of fields. By initiating with the application and thoroughly understanding the unique requirements and challenges associated with it, I have been able to create and implement algorithms that effectively address these issues and achieve the desired outcomes.

Having established the value of this bottom-up approach, it is worth noting that its success has not only been reflected in my scientific publications but has also opened up new opportunities for me in the realm of teaching. As a result, I have managed to introduce new courses in machine learning and deep learning at ESAIP, broadening the horizons for students in these rapidly advancing fields. Additionally, this approach has allowed me to participate in lifelong training courses as a trainer on deep learning for image analysis at both the national and international levels.

Working in collaboration with renowned research institutes, such as EMBL in Germany, has further enhanced my expertise and allowed me to share my knowledge with a wider audience. These collaborations have fostered a dynamic learning environment, enabling researchers, students, and professionals to gain valuable insights into the latest developments in machine learning and deep learning. Consequently, the bottom-up strategy has not only contributed to the advancement of my research but also enabled me to make a meaningful impact in the field of education by providing cutting-edge knowledge and resources to learners and fellow researchers alike.

In this chapter, I will present a synthesis overview of the various methodological contributions made throughout this research. Initially, I will delve into the realm of shallow learning, specifically focusing on my approach to texture-based feature extraction that has significantly improved performance in this domain. In the second section, I will shift my attention to

the advancements made in deep learning techniques, particularly in the area of multimodal deep learning, which has allowed for more effective fusion of different data modalities. Lastly, the third section will be dedicated to addressing the critical issue of annotation bottlenecks in deep learning. I will discuss my proposed models and approaches that aims to mitigate these challenges, ultimately enhancing the efficiency and applicability of deep learning models across various domains.

3.1 Texture-based features for shallow learning

During the initial phase of my research following the completion of my PhD, I collaborated with researchers at the University of Lyon, in the PhD study of Mathilde GIACALONE and Noelie Debs, to delve deeper into the use of texture-based feature extractors for machine learning algorithms. My research on texture-based features was driven by their importance in medical imaging, as they provide valuable information about the spatial arrangement and patterns of pixels within an image [32]. This information can aid in identifying and classifying various structures or regions of interest within the image [33].

Texture-based features can encompass various structures, whether repetitive, completely random, or transitory, making them particularly useful in different medical applications [34]. Some of these applications include lesion detection in medical images [35], identifying different structures in microscopy images [36], and even in early disease diagnosis [33]. Moreover, texture-based features have demonstrated their robustness in a wide range of situations and can often help improve the performance of machine learning algorithms, especially when combined with other types of features or data [34].

Another key advantage of texture-based features is that they are insensitive to small variations in the brightness or contrast of an image, making them suitable for representing the texture of objects in images under varying lighting conditions [32]. This insensitivity is particularly valuable in medical imaging, where capturing images in controlled lighting conditions is often challenging, and changes in brightness or contrast can significantly impact the visibility of various structures within the image [34].

In addition to their robustness and insensitivity to lighting variations, texture-based features have proven to be highly versatile and adaptable. They can be easily incorporated into a wide range of machine learning algorithms, including both traditional methods and deep learning techniques [33, 37, 38]. As a result, my research on texture-based features has not only provided valuable insights into their utility and applicability in

various medical imaging contexts but has also led to the development of innovative solutions that can address the unique challenges and requirements of these diverse applications.

Local Binary Patterns

In my first contribution, we began by investigating feature extractors for machine learning approaches that can represent the texture of objects in an image for different purposes, such as image classification. Local Binary Patterns (LBP) [39] is a commonly used method that has a wide range of applications. One of the advantages of LBP is its robustness to variations in grayscale. This means that it is not sensitive to small changes in the brightness or contrast of an image, making it a suitable technique for representing the texture of objects in images under varying lighting conditions.

Building upon this understanding, in the work presented in [40], we proposed a new variant of LBP for encoding texture features from spatio-temporal images, specifically focusing on the application of perfusion MRI. The new encoding proposed for perfusion MRI is motivated by the fact that the spatio-temporal signature of each voxel is difficult to observe due to the $3D + \text{time}$ nature of the data structure. We propose to encode this spatio-temporal signature into a discriminant texture that is easily identifiable by the human eye and useful for automatically characterizing the state of tissues in each voxel using simple texture analysis tools from computer vision. In order to do so, we propose to encode the information contained in the Moore neighborhood of order 1 of each voxel, as illustrated in Figs (3.1) and (3.2).

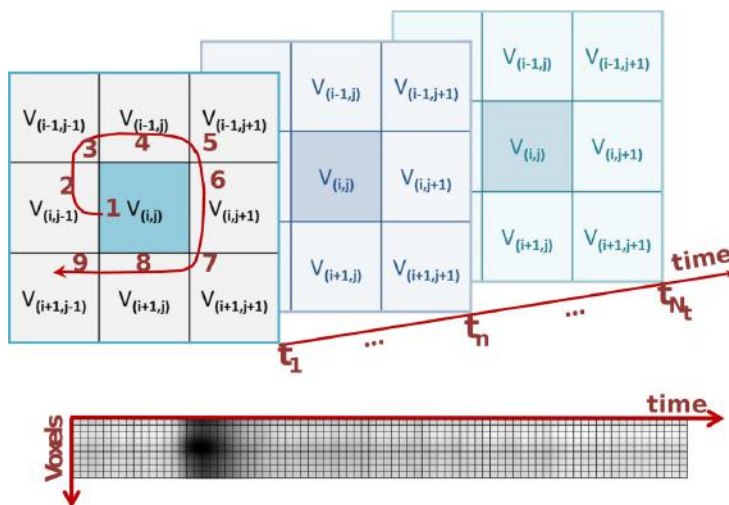


Figure 3.1: Encoding of the spatio-temporal signature of perfusion MRI signals as a patch [40].



Figure 3.2: Illustration of the typical patterns obtained for patches for healthy voxels (left) and pathological voxels (right) [40].

We unfold the temporal signals along a spatial dimension and then pile up, one on top of the other, the temporal signals of the 8 voxels in the Moore neighborhood of order 1 of each voxel of interest, creating a patch of size 9 by N_t , where N_t is the number of temporal acquisition points in the perfusion imaging sequence. This encoding method allows us to create a unique patch for each voxel, which can then be further analyzed using texture analysis and classification tools.

After applying the LBP operator to a patch, the concatenated histograms of the sub-patches separating the labeled patch into contiguous segments can be used as a feature vector to describe the texture in the initial patch, as shown in Fig. (3.3). We use these feature vectors to classify each voxel depending on its final state status using a support vector machine classifier.

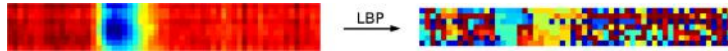


Figure 3.3: Illustration of the LBP labels obtained (right) from a given patch (left) [40].

By evaluating the impact of observation scale and segment width on the precision of tissue fate prediction, we compared the predictive potential of patches and feature vectors of various sizes, aiming to optimize the classification process. This approach demonstrates the potential for texture analysis and classification tools, such as LBP and SVM, to be applied in medical imaging, particularly in perfusion MRI, to aid in tissue fate prediction and other diagnostic tasks.

Although LBP has been widely utilized and proven effective in our works as well as other varieties of applications, it comes with certain limitations. One notable drawback is that LBP is not as invariant to deformations and changes within an image as some other feature extraction methods. This means that when it comes to changes such as scaling, rotation, or translation, LBP might not be as proficient at extracting features that are resilient to these alterations. As a result, LBP's performance may be compromised in tasks where objects within the image could appear at different scales, orientations, or positions like in plant phenotyping applications.

Wavelet Scattering Transform

Alternative feature extraction techniques, such as the Wavelet Scattering Transform (WST) [41], have demonstrated a higher degree of invariance to these types of changes in signals and images. WST is based on the principle of representing an image as a scattering transform, which entails a multi-scale decomposition of the image using wavelets. Wavelets are mathematical functions that can be employed to analyze and represent data, particularly in situations where the signal or image exhibits a hierarchical structure.

The scattering transform effectively captures both the local and global structure of the image, resulting in features that are invariant to deformations like scaling, rotation, and translation. This characteristic of WST makes it particularly well-suited for applications within the life sciences, where the presence of complex and varying structures is common. Furthermore, WST's ability to provide a more comprehensive representation of the image allows it to better adapt to different types of data, thus enhancing its utility in various fields and applications.

A potential challenge when utilizing WST is determining the appropriate number of layers to incorporate into the transform. As the layers increase, the energy of the transformed signal may diminish, potentially leading to the loss of sensitivity to fine details or structures within the image. In the initial WST paper, the authors suggested a maximum of two layers for the transform. Nonetheless, this approach may not always be the most suitable choice, given that the required number of layers can vary based on the specific application and the attributes of the input images. For the first time, we have proposed an automated method for the optimal design of the scatter transform, which is based on energy contrast. This approach effectively tackles the issue of determining the appropriate number of layers for different type of the data [29, 23]. Our method involves examining the energy of the transformed signal at every layer of the scatter transform and determining the number of layers by considering the contrast between the energy at various layers. By taking into account the energy contrast between layers, our method aims to strike a balance between capturing adequate information about the image and preventing the vanishing energy issue that may arise with deeper layers of the transform [29].

In the context of weed detection, our suggested approach in [29] succeeded in enhancing the algorithm's accuracy by selectively choosing the number of layers in the WST, resulting in a superior feature representation of the weeds in the images. In the biometric recognition application, our proposed method in [23] was employed to extract more robust features

from voice signals, ultimately leading to improved performance in terms of recognition accuracy. These outcomes demonstrate the adaptability and versatility of our suggested approach, as it can be applied across a wide array of domains and applications, thus enhancing the performance of the underlying algorithms.

In our research [29], we also emphasized that a key advantage of WST is its capacity to achieve satisfactory performance even when the available data is limited. This is attributed to the fact that WST’s performance does not rely on the size of the training data, and increasing the training data’s size does not result in significant performance improvements. This is particularly valuable in circumstances where acquiring large quantities of training data is challenging or impractical. For instance, in certain medical imaging applications, obtaining vast amounts of labeled data may be difficult or impossible due to ethical and privacy concerns. In such situations, WST’s ability to function effectively with small amounts of data renders it a useful tool.

It is important to mention that although WST can perform well with limited data, employing deep learning algorithms, such as Convolutional Neural Networks (CNNs), can yield even better performance when more data is available as shown in Fig. (3.4). CNNs’ ability to learn from vast amounts of data allows them to achieve superior generalization and performance.

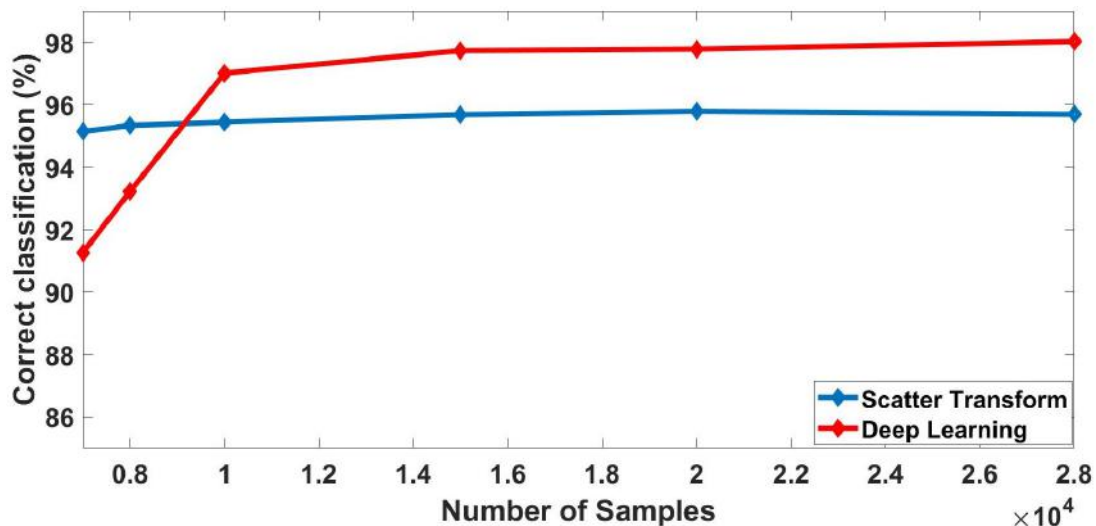


Figure 3.4: Comparison of the recognition accuracy between scatter transform and deep learning when the number of samples increases [29].

3.2 Deep learning algorithms

Upon gaining insights from our study [29], I have chosen to redirect my attention towards utilizing and advancing deep learning algorithms, specifically for applications capable of producing an ample amount of data. This is particularly relevant in areas such as microscopy imaging and plant phenotyping, where the data is abundant in context, expressiveness, and presents numerous challenges.

In order to gain a deeper understanding of this area, I joined forces with the University of Lyon, where we developed CNN models for the analysis of microscopic images [38], as well as magnetic resonance imaging (MRI) [37]. Additionally, these models were utilized in the research carried out by my Ph.D. student Lukman ISMAIL, focusing on MRI [19].

Microscopy imaging provides an in-depth view of internal organs and tissues at a high resolution, but factors such as inconsistent lighting and varying tissue structures can complicate the analysis. By implementing deep learning algorithms, we can develop models that are better equipped to handle these challenges and extract meaningful information from the data.

Similarly, plant phenotyping images present a complex set of challenges that can be addressed through deep learning techniques. Issues such as self-occlusion, the continuous growth of the organism, and variations in lighting, pose, and growth stage make it difficult to analyze these images using traditional methods. By employing advanced CNN models, we can overcome these hurdles and gain a deeper understanding of the underlying processes and patterns.

Building upon the earlier exploration of CNN models for microscopical and MRI images, our research also delved into the challenges presented by spatio-temporal images, which are frequently found in real-life datasets in the life sciences domain. A key limitation of employing CNNs with spatio-temporal images is their primary focus on spatial information, which may hinder their effectiveness in identifying temporal relationships. CNNs excel at processing images by applying convolutional filters to a local region of the image, enabling the detection of local patterns and features. However, they may not efficiently capture dependencies between video frames or time series data. Alternative techniques, such as Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks, may be more appropriate for tasks that necessitate the analysis of spatio-temporal data and modeling of temporal dependencies, given their inherent memory mechanisms. Yet, they might not be as proficient in capturing spatial information as CNNs in computer vision tasks.

In addressing the challenges associated with spatio-temporal data in the plant phenotyping domain, I collaborated with Ph.D. student Salma SAMIEI to develop an innovative model that merges the strengths of CNNs and LSTM networks [25]. This powerful combination, as shown in Fig. (3.5), enhances the classification performance for spatio-temporal images by providing a more comprehensive and effective solution to handle the intricate nature of such data.

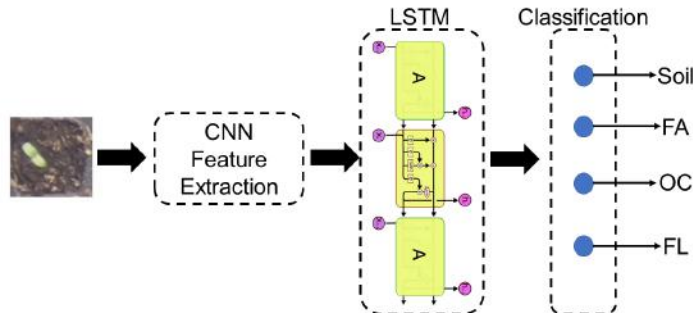


Figure 3.5: CNN-LSTM block [25].

Our initial efforts focused on improving the naive multi-class CNN architecture by incorporating the ontology of plant growth for better discrimination between different growth stages. We developed several ordinal CNN models for binary classification of consecutive developmental stages, training them to detect various growth stage pairs. This approach allowed for the automatic progression through different models as each event was detected during the analysis of a time-lapse sequence.

To further refine the model and introduce memory directly into the CNN architecture, we embedded an LSTM network between the feature extraction and classification blocks. The LSTM architecture, a special RNN structure, has demonstrated stability and effectiveness for long-range dependency modeling in previous studies [42, 43, 44, 45]. The LSTM’s memory cell c^t , acting as an accumulator of state information, is accessed, written, and cleared by several self-parameterized controlling gates. This control of information flow prevents the gradient from vanishing too quickly and is one advantage of the memory cell and gate system [43]. Eq. (3.1) provides the activations for the memory cell and the three gates.

$$\begin{aligned}
 i^t &= \sigma(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i), \\
 f^t &= \sigma(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f), \\
 c^t &= f^t c^{t-1} + i^t \tanh(W_{xc}x^t + W_{hc}h^{t-1} + b_c), \\
 o^t &= \sigma(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^{t-1} + b_o), \\
 h^t &= o^t \tanh(c^t),
 \end{aligned} \tag{3.1}$$

where $\sigma()$ is the sigmoid function, all the matrices W are the connection weights between two units, and $x = (x^0, \dots, x^{T1})$ represents the given input.

The proposed CNN-LSTM model combines the spatial feature extraction capabilities of CNNs with the temporal dependency modeling of LSTMs. This architecture allows for the effective processing of spatio-temporal data and can be applied to various vision tasks involving sequential inputs and outputs. By harnessing the strengths of both CNNs and LSTMs, our innovative model contributes a powerful solution to the challenges posed by spatio-temporal data in plant phenotyping and other domains.

Building on the advancements made with the CNN-LSTM model, my Ph.D. student Hadhami GARBOUGE embarked on a groundbreaking endeavor by introducing and developing a novel vision transformer method for spatio-temporal data. This marked the first application of such a method to spatio-temporal images, as illustrated in Fig. (3.6) [13]. This innovative approach enables the network to concentrate on key events within time-series images, further enhancing our ability to address the inherent challenges associated with the analysis of spatio-temporal data in plant phenotyping and beyond.

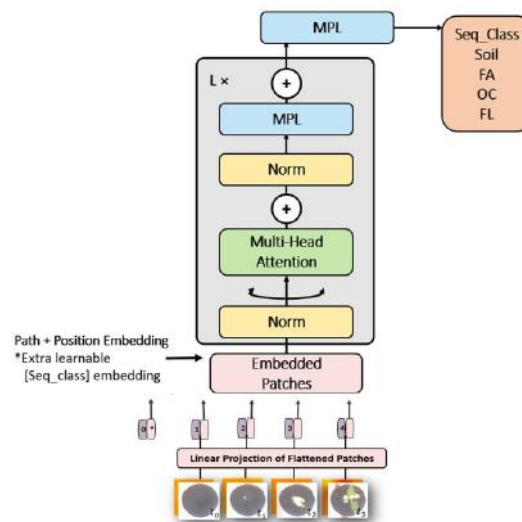


Figure 3.6: Vision transformer method for spatio-temporal data [13].

The fusion of CNNs, LSTM networks, and vision transformer methods provides a robust solution for managing the intricacies of spatio-temporal images in the life sciences domain. By capitalizing on the unique strengths of each approach, we have developed a more adaptable and effective tool for the analysis and interpretation of this essential data. These advancements carry significant potential for future applications and research within the plant phenotyping field and other related areas where spatio-temporal images are of critical importance. Our work in this domain highlights the

promise of deep learning algorithms and emphasizes the need for ongoing exploration into their wide-ranging applications across various disciplines.

Continuing from our earlier work and delving further into the contributions of my Ph.D. student Hadhami GARBOUGE, we identified that one of the challenges when implementing CNNs, CNN-LSTM, transformers models for tasks like classification on spatio-temporal data is that depending exclusively on a single type of modality, such as RGB images, might not provide enough information for accurate predictions. This issue becomes especially pronounced under poor lighting conditions. To address this, we explored methods that take advantage of multimodal CNN models, which combine multiple sources of information to enhance the accuracy and performance of the model.

Multimodal CNNs and transformers can harness various information sources, enabling a more comprehensive and precise representation of the data [13]. When developing multimodal CNN models, it is essential to consider different ways that modalities can be integrated. We examined several fusion approaches, including early fusion, hybrid fusion, feature fusion, and late fusion. Each method has its advantages and drawbacks, and it is crucial to weigh the task requirements and the data characteristics when selecting the most suitable approach. Fig. (3.7) shows an example of the fusion models tested in our work at [13].

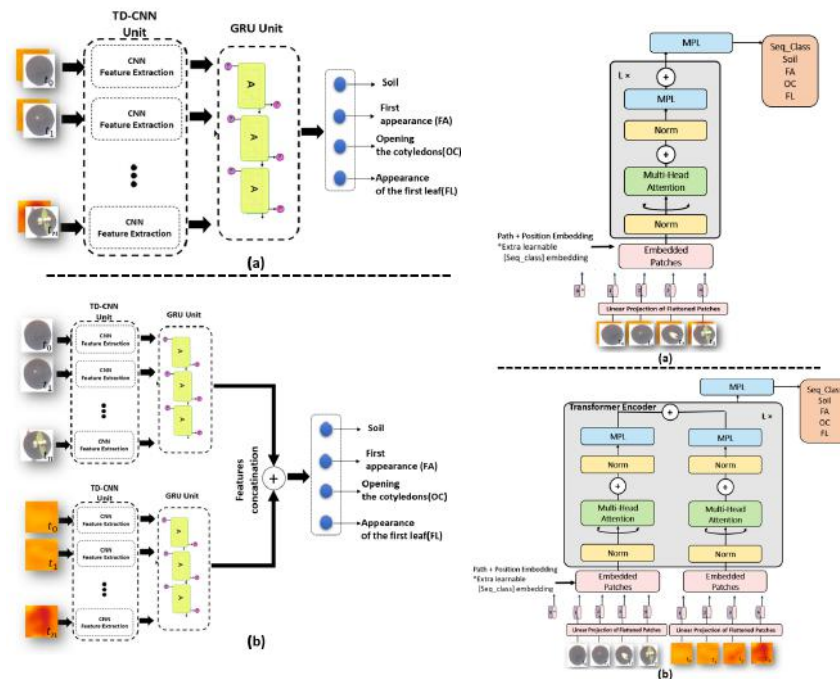


Figure 3.7: (Left) Two fusion techniques for combining RGB-Depth data in the TD-CNN-GRU architecture. (Right) Transformer-based fusion strategy for merging RGB-Depth information in image analysis [13].

Another prime example of effectively merging modalities can be found in the work of my two Ph.D. students, Mathis CORDIER and Abderrazzaq MOUFIDI. They developed multimodal CNN models that incorporate spatio-temporal RGB and depth images, as well as voice signals and depth images. These multimodal models have shown superior performance in classification tasks when compared to single-modality models, such as those using only RGB images or voice signals.

In Abderrazaaq Moufidi’s research, a late fusion architecture was employed, utilizing separate networks designed to handle the distinct information types inherent in each modality. This approach ensures that the unique characteristics of each modality are properly addressed as shown in Fig. (3.8).

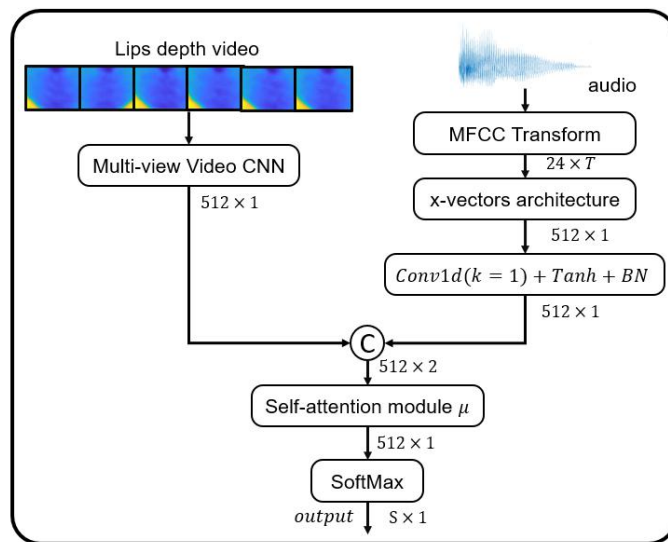


Figure 3.8: Fusion model of depth video and audio signal by using a self-attention model [22].

For the audio modality, it is processed using an Time Delay Neural Network architecture [46], with mean normalization applied to word duration rather than the duration mentioned in the reference article. This step generates a feature vector $X_a \in \mathbb{R}^{512}$. Subsequently, the vector undergoes a 1D-convolution with a kernel size of 1, combined with batch normalization and a $Tanh$ activation function.

For depth video, after mean normalization, the depth video input is fed into the novel architecture outlined in Fig. (3.9). This process results in a feature vector $X_d \in \mathbb{R}^{512}$, which encompasses both visual and dynamic information related to lip movements.

Following the processing of the various modalities, the resulting vector representations are directed through the self-attention module. This specific module calculates the weighted sum of the paired vectors, in accordance with the method detailed in our study [22]. Employing this approach effectively reduces possible redundancy or ambiguity stemming from each

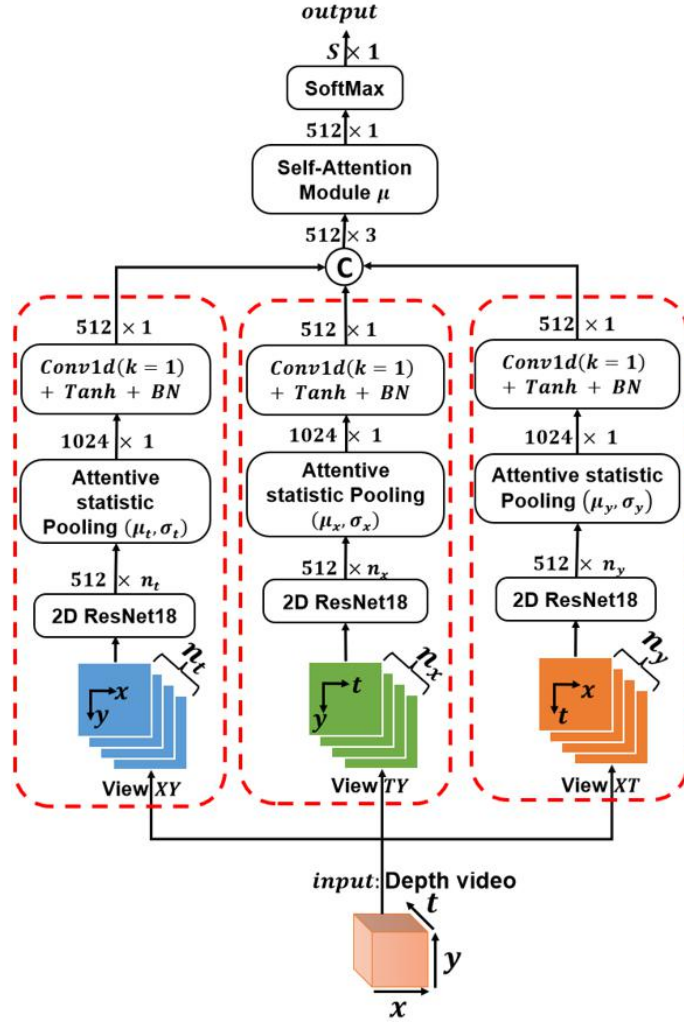


Figure 3.9: Multi-view Video CNN architecture used on lips depth videos (The red dashed line represent the extraction of the features vector from the view projection of the video) [22].

individual modality. Consequently, this leads to the resolution of any contradictions and notably improves the overall performance of the model.

Subsequent to the processing of these modalities, the resulting vectors are channeled through the self-attention module. This module computes the weighted sum of the two vectors, as described in [22]. By the use of this technique, potential redundancy or uncertainty originating from each modality is minimized, thereby mitigating contradictions and significantly enhancing the model’s overall performance.

The integration of multimodal data can significantly enhance the performance of deep learning models, especially when confronted with intricate data sources commonly encountered in life sciences applications. By skillfully merging the advantages of diverse modalities, we can attain a more precise and resilient comprehension of the inherent patterns and associations present in the data. This, in turn, enriches the decision-making process across numerous applications.

As we continue to explore the potential of multimodal data and its implications in deep learning, it is crucial to address the challenges associated with image annotation. The upcoming section will delve into the various obstacles faced in the realm of image annotation and present innovative strategies to overcome these challenges. By focusing on these aspects, we aim to further optimize the efficiency and accuracy of deep learning models, paving the way for more effective utilization of multimodal data in life sciences applications and beyond.

3.3 Image annotation challenges

In our prior research, we have established that deep learning can deliver high-performance outcomes for a variety of issues across multiple domains, provided that extensive annotated databases are accessible. As the development of new deep learning models relies on considerable annotated datasets, image annotation becomes a critical bottleneck. This is attributable to the labor-intensive and time-consuming nature of manual image annotation. Furthermore, the high cost and resources needed for annotation can also impede the acquisition of the necessary volume of data for deep learning models. Consequently, it is essential to discover effective and efficient image annotation methods to overcome this bottleneck and enable the development of novel deep learning models.

In order to tackle the challenges associated with annotation and enhance the performance of deep learning models, we have devised innovative annotation techniques or tools, established more efficient annotation processes, and investigated methods to increase the availability of annotated data. We have begun exploring and proposing new approaches to accelerate the annotation process [47, 48]. For the first time, we examined the utilization of egocentric vision techniques for joint image acquisition and automatic annotation, as opposed to the conventional two-step process of acquisition followed by manual annotation for applications like apple segmentation and counting in the field as shown in Fig. (3.10) [47]. We highlighted the time-saving advantages of employing egocentric vision methods for joint image acquisition and annotation, showcasing a gain of over 10-fold in comparison to the traditional approach [47].

In our study [47], we demonstrated that the most outstanding average performance in apple segmentation accuracy was achieved using eye-tracking-based methods. Examples of challenging images and their resulting annotations with these methods can be seen in Fig. (3.11) for qualitative assessment. Despite significant shift errors, the embedded glasses eye-tracker proved to be highly beneficial, as it enabled simultaneous image

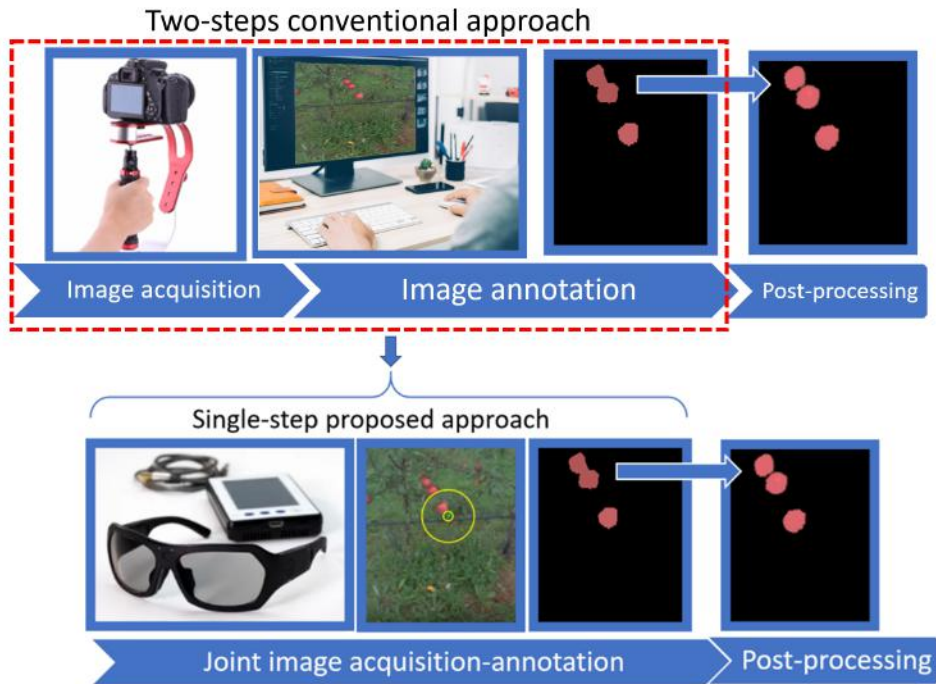


Figure 3.10: The red dotted-line encapsulates the conventional two steps of the acquisition and annotation process. We jointly perform image acquisition and image annotation by the use of a head-mounted egocentric device, which simultaneously captures images and the gaze of the person who wears the device and reaps benefits from both factors to annotate images automatically [47].

acquisition and annotation.

Interestingly, these results were consistent for all three tasks evaluated: segmentation, counting, and localization. This highlights the robustness and usefulness of eye-tracker devices for annotation purposes. Although eye-tracking systems may be considered expensive (typically ranging from 10,000 to 20,000 euros), it is worth noting that the egocentric prior approach can be accessible with any camera embedded on glasses, which could cost between 10 and 100 euros.

Exploring innovative annotation techniques, such as egocentric vision approaches, has paved the way for enhancing the annotation process by automating parts of it, reducing manual intervention, and increasing efficiency. These advancements have made it more feasible to collect the annotated data needed for deep learning models while also potentially extending their impact to various applications and domains. In line with this spirit, we have continued to develop novel solutions for image annotation challenges. Although significant strides have been made to lessen the time and effort required, expert domain knowledge remains essential to guarantee the accuracy of annotations.

Building upon these advancements, we have turned to the application of transfer learning and data augmentation as promising avenues for miti-

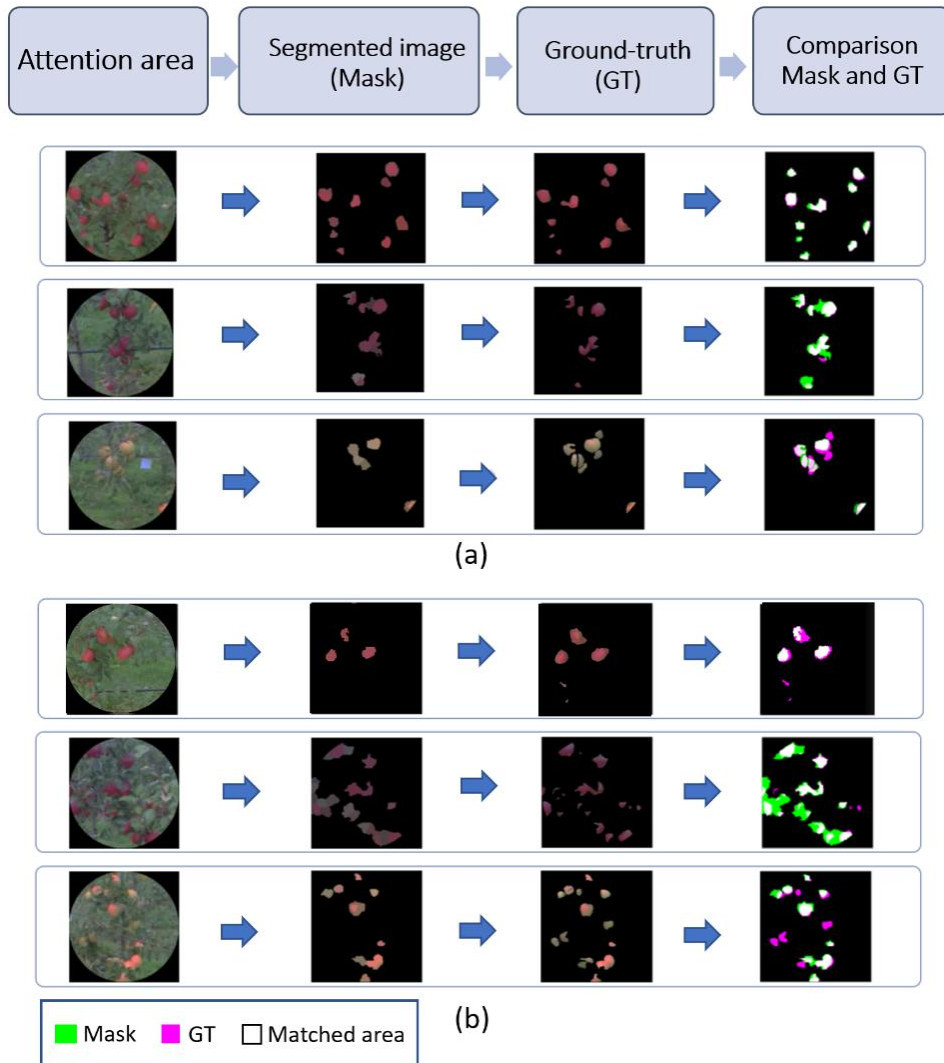


Figure 3.11: Qualitative assessment of results. From left to right, an example of the attention area captured by eye-tracking, automatic annotation obtained from the proposed image processing pipeline, ground-truth manually recorded, and comparison of manual ground-truth and automatic segmentation. **(a)** Examples of good performance; **(b)** Some challenging conditions wherein more errors were found (missed detection, false detection) [47].

gating the difficulties associated with annotation. Specifically, we have investigated the use of transfer learning techniques to facilitate the training of models using synthetic or augmented data [29, 37, 14], thereby leveraging knowledge gained from one modality to improve the performance of models in another modality [17, 14, 49]. Our works has demonstrated the potential of transfer learning approaches to reduce the amount of annotation required while maintaining high levels of accuracy and performance. For instance, in our study [37], we have illustrated the use of hemodynamic signal simulations to increase the amount of data and improve the performance of our CNN model for predicting lesion progression in acute ischemic stroke using DSC-PWI as shown in Fig. (3.12). This novel demonstration of the value of simulation in training machine learning techniques for med-

ical imaging has led to performance levels comparable to those found in the literature for this crucial stroke issue.

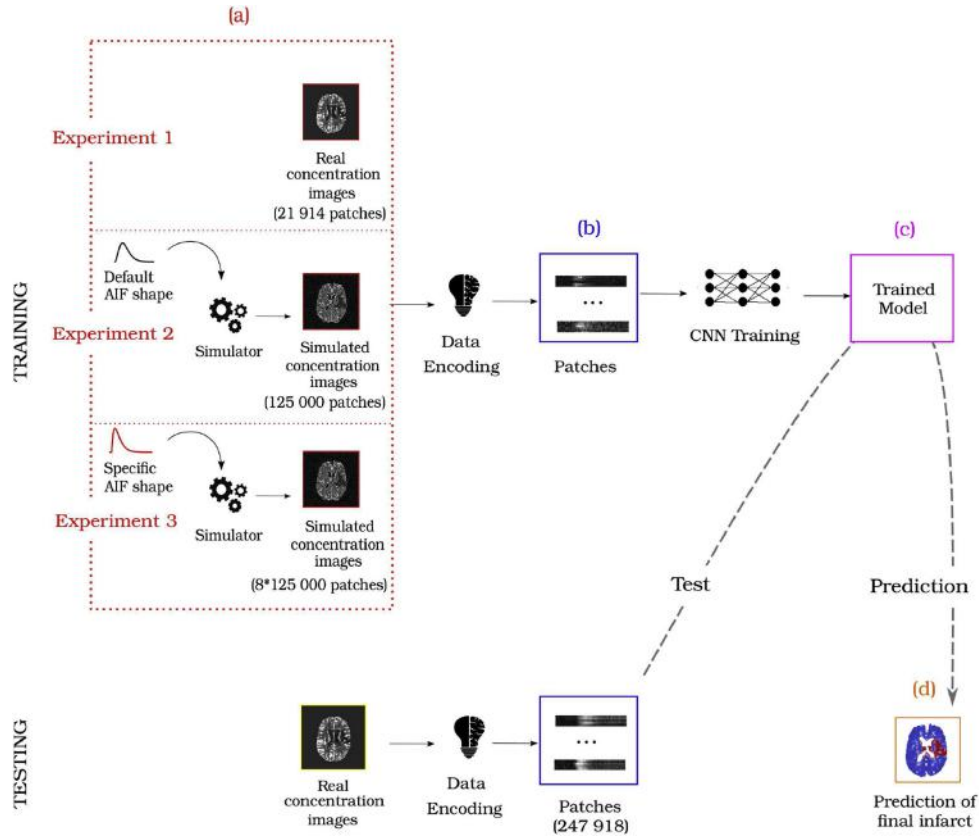


Figure 3.12: (a) The initial images are contrast-agent concentration images. In experiment 1, the training dataset consists in patches from real concentration images, whereas in experiment 2 and 3, the training dataset consists in synthetic patches obtained from the simulator. In experiment 2, AIF input parameter is set to a default value, and in experiment 3, AIF input parameter are the ones of the tested patient. (b) Concentration images are encoded into spatio-temporal patches. (c) A CNN model is trained from patches of the concentration images. (d) Each voxel from the tested concentration images is classified as healthy or infarcted [37].

In another study we conducted [17], we utilized functional MRI (fMRI) images of healthy volunteer subjects to train a model, which was then transferred to another model to detect functional brain networks on fMRI images of individuals with brain tumors.

Another example of transfer learning can be found in our research that demonstrates the use of annotated RGB datasets for segmenting leaves in fluorescence images [49]. By utilizing synthetic datasets with physical modeling of noise in fluorescence and real images in the training process, we achieved good segmentation performance. This instance further highlights the versatility and applicability of transfer learning in addressing the challenges posed by limited data in specific domains, thereby expanding its impact across various fields.

In our another study [14], where we used both transfer knowledge and data augmentation, we expand on our earlier findings presented in [25], in-

investigating the feasibility of transferring knowledge gained from controlled indoor settings (in vitro) to outdoor environments with variable lighting conditions and potential shadows created by the sun or drifting clouds. The goal of this research is to transfer knowledge from a model trained on the initial dataset from a controlled environment to a second dataset from an outdoor environment, as depicted in Fig. (3.13). The challenge in the proposed experiment therefore lay in the presence of shadows which occurs in the green house environment only.

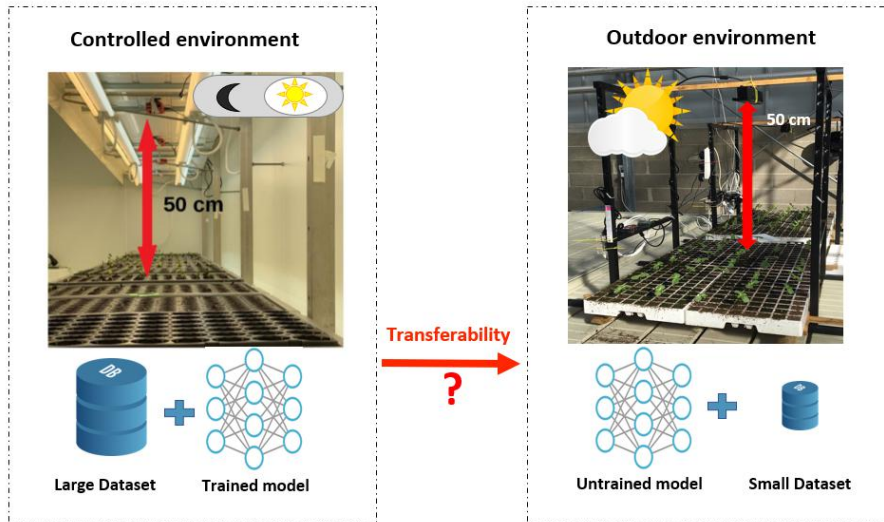


Figure 3.13: Left panel illustrates the imaging system in controlled environment associated with the large database of [25]. Right panel illustrates the imaging system in an outdoor environment with a smaller database. We investigate the possibility of transfer of knowledge from left to right panels [14].

To simulate images acquired in the outdoor environment from indoor images, we propose an automatic shadow generator in [14]. The shadows are randomly positioned by using a thresholded speckle generator [50, 51]. All sizes of shadow can be present in greenhouses. However, only shadows larger than the typical size of seedling organs and smaller than a single plant are expected to impact the detection of seedling development. This information was used to adjust the value of the threshold in the algorithm. Each image in the indoor database is then spatially modulated by the generated shadow with a simple multiplication as shown in Fig. (3.14).

We have explored a range of knowledge transfer methods to facilitate the transition from indoor to outdoor environments [14], encompassing brute transfer (where models trained indoors were directly employed to predict outdoor images), data augmentation (involving the generation of shadows), and model fine-tuning, which incorporated a limited amount of real outdoor data in the training process. Intriguingly, the model's perfor-

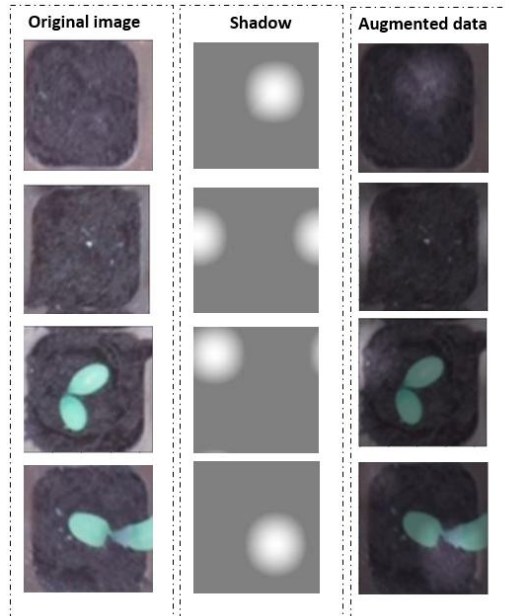


Figure 3.14: Example of original indoor images (left), generated shadows (middle) and, indoor images with simulated shadows (right) [14].

mance saw a marked enhancement when fine-tuning the network trained on data-augmented indoor images with shadows. This approach converged to an elevated level of performance using only a small number of outdoor plants, highlighting the effectiveness of these techniques in adapting to diverse conditions.

Chapter 4

Conclusion and Future directions

4.1 Achievements

Throughout my research career, my primary aim was to investigate and address the scientific questions that persisted within the realm of life science applications, particularly those that could benefit from the integration of deep learning methodologies. The pursuit of this goal led to the proposition of numerous novel methodologies, several of which were first-of-their-kind at the time. Specifically, we investigated the detection of plant growth stages by the use of spatio-temporal data through our proposed CNN-LSTM and spatio-temporal transformers methodologies [25, 13]. Furthermore, the research also addressed the topic of model adaptation and transfer knowledge, focusing on the utilization of models trained on specific data types and repurposing them for different data types or environments. This was particularly evident in our work involving the application of models trained on in vitro data for various environmental settings [13, 14, 19]. Lastly, we explored the potential of exploiting other data modalities to enhance the learning process by our works at [13, 22]

I initially focused on plant phenotyping and medical imaging, which frequently entail multiple visual modalities. Over time, I have consistently broadened the scope of my independent research endeavors, maintaining a bottom-up approach that transitions from practical applications to methodological advancements. This evolution has allowed me to tackle other demanding multimodal fields, including emotion recognition in educational settings, voice signal analysis for biometric identification, and sentiment analysis, all of which are intricate and necessitate the integration of multiple modalities and interdisciplinary elements.

My prior research and project engagements have significantly influenced and shaped my future research objectives and trajectory. This progression is exemplified by my involvement in leading several multi-disciplinary national and international initiatives, such as those funded by the Angers Loire Metropole (ALM) for biometric recognition, PULSAR for emotion

recognition in pedagogical environments, TUBITAC for sentiment analysis in renewable energy, and the Erasmus+ program for machine vision in the realm of industrial Internet of Things (IIoT). These collaborations have been facilitated through partnerships with esteemed institutions, such as the University of Angers' French language center, the pedagogical research group at ESAIP, and Gebze Technical University.

In each of these projects, I am vigorously dedicated to the development of novel and cutting-edge deep learning techniques, aiming to enhance the capabilities of artificial intelligence in solving complex, real-world problems. As I continue to refine and expand my research, my ultimate goal is to contribute to the advancement of machine learning and deep learning methodologies in various domains, fostering innovation, and positively impacting the way we approach challenges in multiple sectors. By cultivating interdisciplinary collaborations and fostering a spirit of discovery, I strive to pave the way for more sophisticated and efficient solutions to problems in fields as diverse as healthcare, education, security, and sustainability.

Building upon my comprehensive experience in multidisciplinary research, I have devised an innovative strategic plan that adheres to my established approach of beginning with practical applications and progressing toward methodological advancements. This plan encompasses the development of multimodal deep learning models for biometric identification, emotion recognition, and sentiment analysis.

As we continue to develop these models, our research has the potential to revolutionize various industries, including cybersecurity, education, and energy. By bridging the gap between applications and methodologies, we hope to foster a more comprehensive understanding of the challenges and opportunities that lie ahead in these fields. Additionally, by embracing the power of multimodal deep learning, we can create more robust and accurate models that account for the complexity and nuance inherent in human behavior and emotions.

4.2 Future directions

4.2.1 Research

Recent breakthroughs in deep learning, such as generative models and multimodal data analysis, have paved the way for the development of robust and accurate models for a wide range of applications. A primary focus of our research is to minimize the dependency on manual annotation, thereby enhancing the efficiency and effectiveness of various applications. To achieve our goal, we are actively investigating innovative techniques that can decrease the reliance on manual annotation within our individual

and internal projects. As we make progress in this area, it opens up opportunities to participate in ANR and European projects in collaboration with academic institutions. Furthermore, we are mentoring new Ph.D. students and postdoctoral researchers to help advance these efforts. Our aim is to improve the performance of projects in areas like life science imaging, emotion recognition, and biometric recognition while reducing the burden of annotation.

A key objective in achieving this goal is the development and refinement of self-supervised learning representation (SSLR) methods to decrease annotation requirements. We have recently proposed a SSLR technique for situations where only limited annotated data is available, as demonstrated in our work on fMRI images of unhealthy subjects within the context of my current Ph.D. student’s work [18]. This approach leverages abundant unannotated data to improve the performance of deep learning models when labeled data is scarce. By utilizing unannotated data for pretext tasks and annotated data for downstream tasks, we can capitalize on the wealth of unannotated data, ultimately enhancing deep learning model performance.

In the literature, various technologies and models, such as generative models and SSLR models, have been developed. Popular SSLR pretext tasks include rotation, jigsaw, and instance discrimination [52, 53]. However, traditional SSLR pretext tasks can be challenging for certain domains, like life sciences, which encompass computational biology, medicine, and digital plant phenotyping. These challenges arise due to the absence of canonical orientation and the textural nature of the problems.

The work proposed in [54] were among the first to explore image generation as an SSLR pretext task with biological images, focusing on the morphological profiling of human cultured cells with fluorescence microscopy. Although they speculated that adversarially learned representations might be superior to autoencoder-based ones, their generative approach was found to be not yet competitive with traditional transfer learning-based methodologies. More recent studies have investigated and improved generative-based SSLR methods [55]. However, these works typically rely on side information to construct their generative pretext task by the use of Generative Adversarial Networks (GANs), tailoring it to specific applications [56].

The concept of utilizing GAN’s discriminator as a feature extractor was initially introduced and employed in various studies [57, 58, 59]. The research conducted in [60] demonstrated that the efficiency and robustness of discriminator features are strongly dependent on preventing mode collapse within the network. Wasserstein GANs, which include the StyleGAN2

family, are known for their notable resistance to mode collapse [59, 61]. However, GANs can be computationally demanding during training and may occasionally produce images with artifacts or unrealistic features. Additionally, achieving precise control over generated outputs can be a challenging task.

In contrast, diffusion models, which are based on denoising score matching and contrastive divergence, have surfaced as a viable alternative for image synthesis. These models offer several advantages, such as a more stable training process, reduced sensitivity to hyperparameter selection, and generally lower requirements for training data and computational resources compared to GANs. In lieu of this, diffusion models hold significant promise for various image synthesis tasks due to their inherent benefits. As advancements in the field continue to unfold, it is expected that the performance of diffusion models will improve, thereby increasing their versatility and applicability across a wide range of domains and applications.

Nonetheless, the majority of these methods concentrate on a single modality, while multimodal models have the potential to perform better. Building on my previous research, I plan to develop a collaborative approach called multimodal generative self-supervised representative learning that integrates self-supervised learning with generative models like diffusion and autoencoder models.

My aim in merging these two model types is to enhance image analysis efficiency and accuracy while minimizing the need for annotation. I intend to employ the self-representations acquired by diffusion models as a replacement for the traditional pretext task in SSLR models. This will guide the generation process in autoencoder models, increasing their robustness to changes in data distribution. Furthermore, integrating multiple modalities or domains into the training process will improve the models' generalization capabilities. Ultimately, I aim to create a method that can effectively transform images between various modalities or domains with minimal human intervention, making it a valuable resource for researchers and practitioners across different fields.

In order to accomplish this, I will develop a novel technique called Multi-Modal Diffusion Self-Supervised Learning (MMDSSL). This method will seek to extract meaningful and high-quality features from multimodal data, taking advantage of the distinct characteristics of each modality to enrich the overall understanding of the data. The conceptual illustration of the MMDSSL model can be seen in Fig. (4.1), which provides a preliminary visual representation of the proposed idea. By concentrating on the development and improvement of self-supervised learning methods, I aspire to reduce manual annotation requirements, leading to more efficient and

effective deep learning models. In turn, this will advance my work in fields such as emotion recognition, sentiment analysis, and biometric recognition, ultimately contributing to the broader domain of artificial intelligence and its real-world applications.

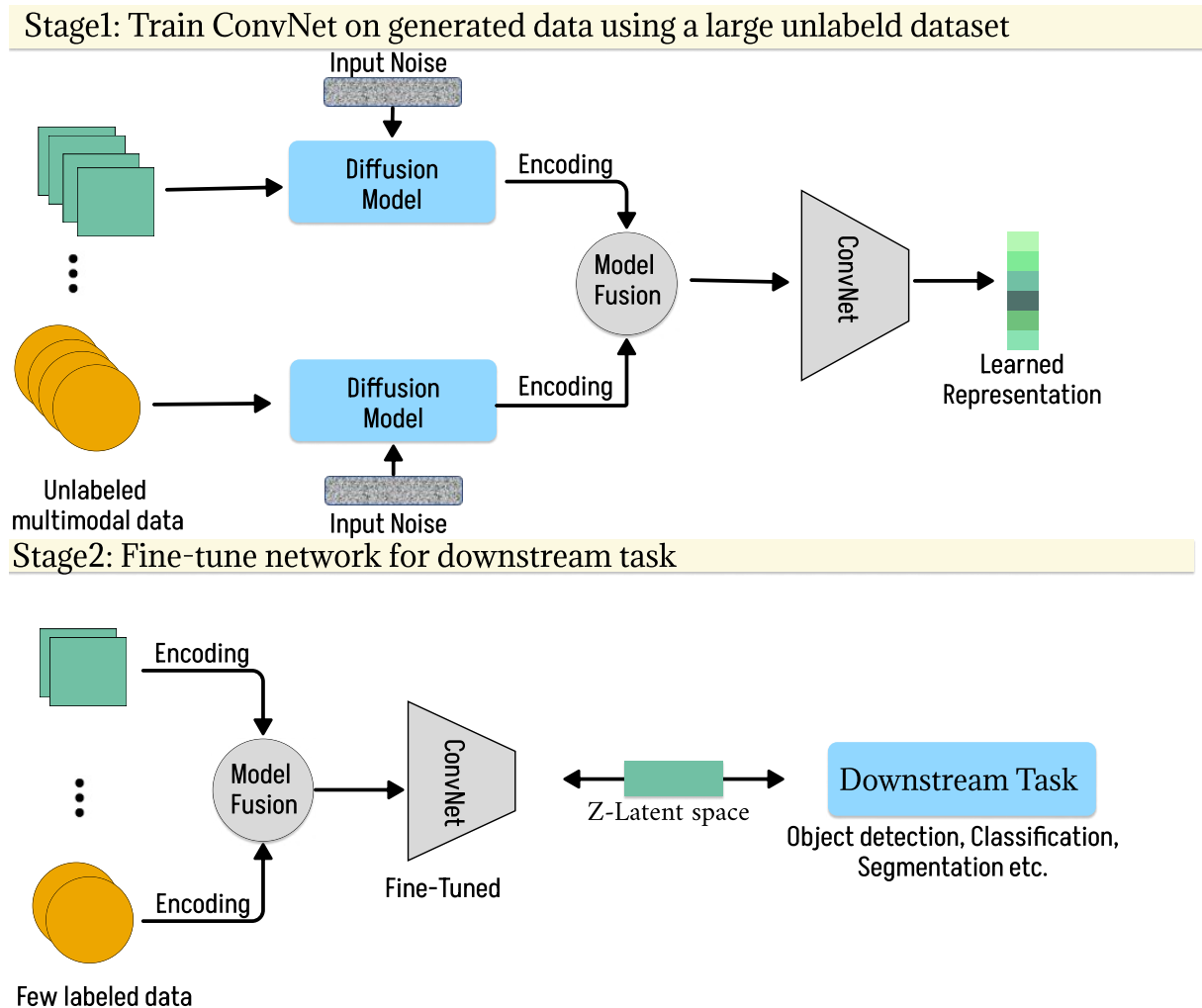


Figure 4.1: Preliminary visual representation of the MMDSSL

In the context of MMDSSL, I plan to represent the multimodal data as a set of observations with multiple modalities. The objective will be to learn a joint feature representation that effectively captures the relationships between the modalities.

The MMDSSL method will involve two main components: a diffusion model and a self-supervised learning task. The diffusion model will be responsible for learning the data representation and generating new samples for each modality in the dataset. This will be achieved by incorporating a noise schedule function and an isotropic Gaussian noise.

In MMDSSL, the self-supervised learning task will aim to learn useful features by using diffusion models that do not require any specific annotation. I will design a loss function to encourage the model to understand the relationships and dependencies between modalities.

By optimizing the joint objective function with a hyperparameter controlling the trade-off between the generative and self-supervised learning objectives, I aim to learn a robust and informative feature representation that effectively captures the underlying structure and relationships in the multimodal data.

To seamlessly incorporate the attention mechanism into the MMDSSL framework, I plan to introduce a transition that highlights its importance in learning complex relationships between modalities. This can be achieved by emphasizing the role of attention in effectively capturing the underlying structure and dependencies across different data modalities.

In order to improve the multimodal feature learning process, I suggest incorporating attention mechanisms into the MMDSSL method. By integrating the attention mechanism into both the encoder and the diffusion model, the models can learn more intricate relationships between modalities, ultimately enhancing the MMDSSL method’s performance across various tasks and applications.

The attention mechanism allows the model to dynamically concentrate on the most relevant portions of the input data across modalities, resulting in more robust and discriminative features. This approach can be particularly beneficial in situations where different modalities provide complementary information for a more comprehensive understanding of the data.

The MMDSSL method has the potential to be applied to various domains in life science and biometric recognition, due to its ability to effectively learn high-quality representations from multimodal data with minimal annotation. Some possible application domains include:

In microscopy imaging, MMDSSL can be applied to study cellular structures and molecular interactions using different imaging techniques, such as fluorescence, phase contrast, and electron microscopy. By learning joint representations from these diverse modalities with minimum annotation data, MMDSSL can facilitate the analysis of complex biological processes and enable researchers to gain a better understanding of cellular and sub-cellular structures.

For medical imaging, MMDSSL can be used to fuse and analyze information from multiple imaging modalities, such as MRI, CT, PET, and ultrasound, with limited annotation requirements. By learning a shared representation of the data, MMDSSL can help improve the diagnosis and treatment of various medical conditions. For example, it can be employed to better identify and localize tumors, assess tissue damage in stroke patients, or monitor the progression of neurodegenerative diseases.

In plant phenotyping, MMDSSL can be applied to analyze multimodal data from various imaging techniques, such as visible, infrared, and hy-

perspectral imaging, to assess plant traits and growth patterns. With minimum annotation data, MMDSSL can help researchers identify critical features related to plant health, stress responses, and yield potential, ultimately contributing to more efficient breeding strategies and improved crop management.

Lastly, in biometric recognition, MMDSSL can be employed to learn robust and discriminative features from multimodal data, such as face, fingerprint, iris, and voice data, with minimal annotation needs. By effectively capturing the relationships between different modalities, MMDSSL can enhance the performance of biometric recognition systems, making them more accurate and reliable for various security applications.

4.2.2 Pedagogy

My research endeavors not only strive to advance machine learning and deep learning methodologies but also hold the potential to pave the way for new lifelong training initiatives and the introduction of new courses at both the national and international levels through collaborations like Erasmus+ projects. These research pursuits also enable me to propose cutting-edge mini-projects and internship opportunities for young students with aspirations to work in the realm of machine learning and deep learning.

Moreover, I intend to present these projects to the industry as services or innovative product technologies, thereby fostering strong connections between academic research and the industrial sector, encouraging knowledge exchange and collaboration. These efforts can lead to the creation of practical technologies with real-world applications that ultimately benefit society as a whole.

By bridging the gap between academia and industry, I believe we can create a dynamic environment where the exchange of ideas, resources, and expertise leads to the rapid development and implementation of advanced technologies in various sectors. These interdisciplinary collaborations can drive innovation and efficiency, promoting economic growth and improving the overall quality of life for individuals and communities.

In addition to offering training programs and courses, I envision creating a comprehensive platform that provides resources, mentorship, and networking opportunities for students and professionals alike. This platform would facilitate knowledge sharing, skill development, and collaboration among individuals and organizations working in machine learning, deep learning, and related fields.

Ultimately, my research plans aim to not only push the boundaries of what is possible in the field of machine learning and deep learning but also to inspire and empower the next generation of researchers and innovators.

By fostering an environment of collaboration, education, and innovation, we can work together to tackle the complex challenges that lie ahead and drive meaningful change in the world.

Bibliography

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [3] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.
- [4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [7] V. Badrinarayanan, A. Handa, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling,” *arXiv preprint arXiv:1505.07293*, 2015.
- [8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [9] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

- [10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [12] M. Zine El Abidine, H. Dutagaci, P. Rasti, M.-J. Aranzana, C. Dujak, and D. Rousseau, “Toward objective variety testing score based on computer vision and unsupervised machine learning application to apple shape,” *Biosystems Engineering*, 2023.
- [13] H. Garbougé, P. Rasti, and D. Rousseau, “Enhancing the tracking of seedling growth using rgb-depth fusion and deep learning,” *Sensors*, vol. 21, no. 24, p. 8425, 2021.
- [14] H. Garbougé, P. Rasti, S. Samiei, and D. Rousseau, “Deep learning-based detection of seedling development from indoor to outdoor,” in *Systems, Signals and Image Processing*. Springer, 2022.
- [15] H. Garbougé, S. Samiei, P. Rasti, and D. Rousseau, “Machine-learning assisted determination of best acquisition protocols in variety testing,” in *AI for Agriculture and Food Systems*, 2021.
- [16] M. Cordier, H. Garbougé, S. Samiei, P. Rasti, and D. Rousseau, “Growth-data a new tool to characterize spatio-spectral patterns of plant growth,” in *North American Plant Phenotyping Network Annual Conference*. Springer, 2021.
- [17] L. E. Ismaila, P. Rasti, F. Bernard, M. Labriffe, P. Menei, A. T. Minasian, D. Rousseau, and J.-M. Lemée, “Transfer learning from healthy to unhealthy patients for the automated classification of functional brain networks in fmri,” *Applied Sciences*, vol. 12, no. 14, p. 6925, 2022.
- [18] L. E. Ismaila, P. Rasti, J.-M. Lemée, and D. Rousseau, “Self-supervised learning for functional brain networks identification in fmri from healthy to unhealthy patients,” in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2022.
- [19] L. Ismael, J.-M. Lemée, D. Rousseau, and P. Rasti, “Deep learning pour la classification automatique de réseaux cérébraux fonctionnels par irmf de repos,” *Journal of Neuroradiology*, vol. 49, no. 2, p. 119, 2022.

- [20] L. E. Ismaila, P. Rasti, J.-M. Lemee, and D. Rousseau, “Toward more frugal models for functional cerebral networks automatic recognition with resting state fmri,” in *29ème Colloque GRETSI sur Le Traitement du Signal et des Images, Grenoble, France*, 2023.
- [21] S. Hamdy, A. Charrier, L. Le Corre, P. Rasti, and D. Rousseau, “Advances in seed phenotyping using x-ray imaging,” 2022.
- [22] A. Moufidi, D. Rousseau, and P. Rasti, “Attention-based fusion of ultra-short voice utterances and depth videos for multi-modal person identification,” *Sensors*, 2021.
- [23] A. Moufidi, P. Rasti, and D. Rousseau, “Wavelet scattering transform depth benefit, an application for speaker identification,” in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2023.
- [24] M. Cordier, P. Rasti, C. Torres, and D. Rousseau, “Optimisation de l’échelle d’observation pour l’annotation d’images,” 2022.
- [25] S. Samiei, P. Rasti, J. L. Vu, J. Buitink, and D. Rousseau, “Deep learning-based detection of seedling development,” *Plant Methods*, vol. 16, no. 1, pp. 1–11, 2020.
- [26] H. Dutagaci, P. Rasti, G. Galopin, and D. Rousseau, “Rose-x: an annotated data set for evaluation of 3d plant organ segmentation methods,” *Plant methods*, vol. 16, no. 1, pp. 1–14, 2020.
- [27] C. Sommer, C. Straehle, U. Koethe, and F. A. Hamprecht, “Ilastik: Interactive learning and segmentation toolkit,” in *2011 IEEE international symposium on biomedical imaging: From nano to macro*. IEEE, 2011.
- [28] “The rose-x dataset,” <https://uabox.univ-angers.fr/index.php/s/rnPm5EHFK6Xym9t>, accessed: 2023-04-24.
- [29] P. Rasti, A. Ahmad, S. Samiei, E. Belin, and D. Rousseau, “Supervised image classification by scattering transform with application to weed detection in culture crops of high density,” *Remote Sensing*, vol. 11, no. 3, p. 249, 2019.
- [30] N. Harte and E. Gillen, “Tcd-timit: An audio-visual corpus of continuous speech,” *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.
- [31] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.

- [32] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.
- [33] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, “Image biomarker standardisation initiative-feature definitions,” *arXiv preprint arXiv:1612.07003*, vol. 10, 2016.
- [34] G. Castellano, L. Bonilha, L. Li, and F. Cendes, “Texture analysis of medical images,” *Clinical radiology*, vol. 59, no. 12, pp. 1061–1069, 2004.
- [35] E. M. Senan and M. E. Jadhav, “Techniques for the detection of skin lesions in ph 2 dermoscopy images using local binary pattern (lbp),” in *3rd International Conference, RTIP2R 2020*. Springer, 2021.
- [36] A. Gorai, K. Cetina, L. Baumela, and A. Ghosh, “A comparative study of local binary pattern descriptors and gabor filter for electron microscopy image segmentation,” in *2014 International Conference on Parallel, Distributed and Grid Computing*. IEEE, 2014.
- [37] N. Debs, P. Rasti, L. Victor, T.-H. Cho, C. Frindel, and D. Rousseau, “Simulated perfusion mri data to boost training of convolutional neural networks for lesion fate prediction in acute stroke,” *Computers in biology and medicine*, vol. 116, p. 103579, 2020.
- [38] P. Rasti, C. Wolf, H. Dorez, R. Sablong, D. Moussata, S. Samiei, and D. Rousseau, “Machine learning based classification of the health state of mice colon in cancer study from confocal laser endomicroscopy,” *Scientific Report*, 2019.
- [39] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [40] M. Giacalone, P. Rasti, N. Debs, C. Frindel, T.-H. Chol, E. Grenier, and D. Rousseau, “Local spatio-temporal encoding of raw perfusion mri for the prediction of final lesion in stroke,” *Medical Image Analysis*, 2018.
- [41] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.
- [42] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013.

- [43] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [44] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *International conference on machine learning*, 2013.
- [45] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014.
- [46] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.
- [47] S. Samiei, P. Rasti, H. Daniel, E. Belin, P. Richard, and D. Rousseau, “Toward a computer vision perspective on the visual impact of vegetation in symmetries of urban environments,” *Symmetry*, vol. 10, no. 12, p. 666, 2018.
- [48] S. Samiei, A. Ahmad, P. Rasti, E. Belin, and D. Rousseau, “Low-cost image annotation for supervised machine learning. application to the detection of weeds in dense culture,” in *British Machine Vision Conference (BMVC), Computer Vision Problems in Plant Phenotyping (CVPPP)*. BMVA Press Newcastle, UK, 2018.
- [49] N. Sapoukhina, S. Samiei, P. Rasti, and D. Rousseau, “Data augmentation from rgb to chlorophyll fluorescence imaging application to leaf segmentation of arabidopsis thaliana from top view images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2019.
- [50] J. W. Goodman, *Speckle phenomena in optics: theory and applications*. Roberts and Company Publishers, 2007.
- [51] C. Douarre, R. Schielein, C. Frindel, S. Gerth, and D. Rousseau, “Transfer learning from synthetic data applied to soil–root segmentation in x-ray tomography images,” *Journal of Imaging*, vol. 4, no. 5, p. 65, 2018.
- [52] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2021.
- [53] B. Wallace and B. Hariharan, “Extending and analyzing self-supervised learning across domains,” in *Computer Vision–ECCV 2020: 16th European Conference*. Springer, 2020.

- [54] P. Goldsborough, N. Pawlowski, J. C. Caicedo, S. Singh, and A. E. Carpenter, “Cytogan: generative modeling of cell images,” *BioRxiv*, p. 227645, 2017.
- [55] A. Mascolini, D. Cardamone, F. Ponzio, S. Di Cataldo, and E. Ficarra, “Exploiting generative self-supervised learning for the assessment of biological images with lack of annotations,” *BMC bioinformatics*, vol. 23, no. 1, pp. 1–17, 2022.
- [56] P. Yang, Z. Hong, X. Yin, C. Zhu, and R. Jiang, “Self-supervised visual representation learning for histopathological images,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*. Springer, 2021.
- [57] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [58] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, “Marta gans: Unsupervised representation learning for remote sensing image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2092–2096, 2017.
- [59] M. Zhang, M. Gong, Y. Mao, J. Li, and Y. Wu, “Unsupervised feature extraction in hyperspectral images based on wasserstein generative adversarial network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2669–2688, 2018.
- [60] X. Mao, Z. Su, P. S. Tan, J. K. Chow, and Y.-H. Wang, “Is discriminator a good feature extractor?” *arXiv preprint arXiv:1912.00789*, 2019.
- [61] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*. PMLR, 2017.
- [62] S. Samiei, P. Rasti, P. Richard, G. Galopin, and D. Rousseau, “Toward joint acquisition-annotation of images with egocentric devices for a lower-cost machine learning application to apple detection,” *Sensors*, vol. 20, no. 15, p. 4173, 2020.

Appendix

In this appendix, I provide a selection of my highlighted published articles based on the various databases and methodology that have been discussed in Chapters 2 and 3. The purpose of including these articles is to showcase the depth and breadth of research conducted in the development and utilization of the described databases and methodologies, and to demonstrate how these resources have enabled the advancement of novel techniques in the realm of machine learning and deep learning. By highlighting the contributions made through the development of these databases and the exploration of innovative methodologies, this appendix serves as a testament to the interdisciplinary nature of the work and its potential for driving innovation and addressing real-world challenges.

The list begins with works conducted on installing low-cost RGB and Depth imaging platforms to monitor seedling growth [25] and [13]. These articles highlight the innovative imaging system and database specifically tailored for plant phenotyping, featuring a diverse range of seedlings at various growth stages and under different environmental conditions. Additionally, I have included the article [26], which outlines the acquisition and reconstruction process of our 3D images of rosebush plants. This work demonstrates the significance of 3D imaging techniques for capturing detailed plant architecture and their potential in advancing the understanding of plant growth and development.

In the data simulation domain, I have added the article [29], which describes the simulation of weeds surrounded by plants and its application in our global data challenge. This article emphasizes the importance of synthetic data in training deep learning models and showcases the potential of such models in addressing real-world challenges in agriculture and beyond.

For the methodology section explained in chapter 3, I have incorporated our works on texture-based feature extraction for medical and plant phenotyping, as seen in the articles [40] and [29]. These articles illustrate how machine learning approaches have been used to solve problems in the life science domain, highlighting the versatility of such techniques.

I also added our works on deep learning methods, where we proposed new models to be used on a variety of problems in life science applications [25, 13, 38]. These articles showcase the power of deep learning techniques

in addressing complex challenges and the potential of these models to revolutionize the field of life sciences.

To complete this section, I added our works on proposing approaches to solve annotation problems, introducing an egocentric vision technique [62] as well as transfer learning approaches [13, 37, 49]. These articles demonstrate methods that facilitate speeding up the image annotation process and utilizing knowledge transfer from one environment or modality to another, which substantially reduces the time and effort needed for manual annotation.

By presenting this selection of my highlighted published articles, I aim to emphasize the substantial contributions made to the field of machine learning and deep learning through the development and application of the described databases and methodologies. These works not only demonstrate the significance of high-quality data in advancing research but also showcase the potential of interdisciplinary collaboration in driving innovation and addressing real-world challenges.

RESEARCH

Open Access



Deep learning-based detection of seedling development

Salma Samiei¹, Pejman Rasti^{1,3}, Joseph Ly Vu², Julia Buitink² and David Rousseau^{1*} 

Abstract

Background: Monitoring the timing of seedling emergence and early development via high-throughput phenotyping with computer vision is a challenging topic of high interest in plant science. While most studies focus on the measurements of leaf area index or detection of specific events such as emergence, little attention has been put on the identification of kinetics of events of early seedling development on a seed to seed basis.

Result: Imaging systems screened the whole seedling growth process from the top view. Precise annotation of emergence out of the soil, cotyledon opening, and appearance of first leaf was conducted. This annotated data set served to train deep neural networks. Various strategies to incorporate in neural networks, the prior knowledge of the order of the developmental stages were investigated. Best results were obtained with a deep neural network followed with a long short term memory cell, which achieves more than 90% accuracy of correct detection.

Conclusion: This work provides a full pipeline of image processing and machine learning to classify three stages of plant growth plus soil on the different accessions of two species of red clover and alfalfa but which could easily be extended to other crops and other stages of development.

Keywords: Seedling development, Deep learning, Kinetic

Background

A specificity of plants is their continuous capability to metamorphose during their lifetime. This process is characterized by the kinetics of ontological development stages, i.e., stages that occur in a definite order. In this article, we focus on some of these connected steps of a plant's life at the seedling level. The period from seed germination in the soil to the development of the first true leaf is crucial for the plant. During this time, the seedling must determine the appropriate mode of action based on its environment to best achieve photosynthetic success and enable the plant to complete its life cycle. Once the seedling emerges out the soil, it initiates photomorphogenesis, a complex sequence of light-induced

developmental and growth events leading to a fully functional leaf. This sequence includes severe reduction of hypocotyl growth, the opening of cotyledons, initiation of photosynthesis, and activation of the meristem at the shoot apex, a reservoir of undifferentiated cells that will lead to the formation of the first leaf [1]. The molecular mechanisms regulating these time-based events involves profound reprogramming of the genome that is challenging to study in field situation because the heterogeneity of the seedling population must be taken into account. It is essential to understand this seedling development process from an agronomic point of view because the seedling establishment is critical to crop yield. Uneven emergence timing, for instance, is associated with lower yields and poor farmer acceptance.

In this context, time-lapse imaging is a valuable tool, accessible at a rather low-cost [2–5], for documenting plant development and can reveal differences that would not be apparent from a sole endpoint analysis. At the

*Correspondence: david.rousseau@univ-angers.fr

¹ Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRAe IRHS, Université d'Angers, Angers, France
Full list of author information is available at the end of the article



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

seedling level where plants have simple architectures, such time-lapse imaging can be done from top view to provide an efficient solution for seedling vigor assessments and monitoring of seedling growth. While some statistical tools transferred from developmental biology exists to perform time-to-event analysis [6], a current bottleneck [7] lay in the automation of the image analysis. A recent revolution occurred in the field of automated image analysis with deep neural networks [8], which have shown their universal capability to address almost any image processing challenges with high accuracy. This revolution also benefits plant imaging [9], and it is currently a timely topic to adapt these tools, which came from the artificial intelligence community to specific topics of interest in plant sciences. In this article, we propose an entire pipeline based on deep learning dedicated to the monitoring of seedling growth.

Seedling growth monitoring with computer vision has received considerable attention in the literature including [10–24]. It is therefore important to locate our proposition with these related works. While each article of this literature deals with the quantification of some aspects of the early stages of plant development, it includes a large variety of approaches behind the word seedling. Several studies consider germination and seedling growth measurements *in vitro*, using plastic boxes or paper towel [10–17, 21], which enable the monitoring of radicle emergence (germination) or organ growth (seedling growth). Others, like in this article, used soil-based sowing systems, where seedling emergence and early developmental events of the aerial part can be determined under more realistic agronomical conditions [19, 22–26]. Reported approaches to monitor seedling from the top view in the soil are effective for a large set of crops, mainly at the emergence level, i.e., seedling counting to determine stand establishment [19, 23–26], or estimating early plant vigor by spectral imaging or measuring the leaf area index of the small plants [19, 22, 26]. As most related work, deep learning has been applied to the problem of seedling detection and segmentation [24]. By contrast with our work, this has been performed at a fixed stage of development. Here we propose to push forward the detection of the early seedling developmental stages to be able to monitor the kinetics of early seedling development in the soil from cotyledon emergence until the development of the first real leaf. We propose to tackle this task of seedling kinetics monitoring, for the first time to the best of our knowledge, with a deep learning-based approach.

Spatio-temporal approaches in deep-learning have been extensively developed in computer vision for video processing [27] but has so far been very rarely applied in plant imaging [28] (for growth prediction). As most

related work in spatio-temporal processing [2] proposed a graph-based method for detection and tracking of tobacco leaves at the late stage of the plant growth from infrared image sequences. This study was not based on deep learning and was applied on later stage of development than seedling. In the last similar approach [20], a feature-based machine learning algorithm distinct from deep learning was developed to detect two stages of heading and flowering of wheat growth.

In this article, we investigate, for the first time to the best of our knowledge in plant imaging, how the existing methods of spatio-temporal deep learning, can incorporate time-dependency in sequences of images to solve the problem of monitoring the developmental kinetics. While the proposed method is of general value for developmental biology, its performance is assessed on the specific use case of seedlings of red clover and alfalfa imaged from top view.

Materials and method

The proposed plant method includes four main items: (i) The imaging system developed to create (ii) the dataset, which needs to benefit from (iii) pre-processing before investigating (iv) various approaches for the detection of developmental stages of seedling growth based on deep learning methods.

Imaging system

A set of minicomputers (as described in [3]) connected to RGB cameras with a spatial resolution of 3280 by 2464 pixels was used to image seedlings from the top view as illustrated in Fig. 1. The distance of 50 cm was chosen to allow the observation of 2 trays of 200 pots per camera.

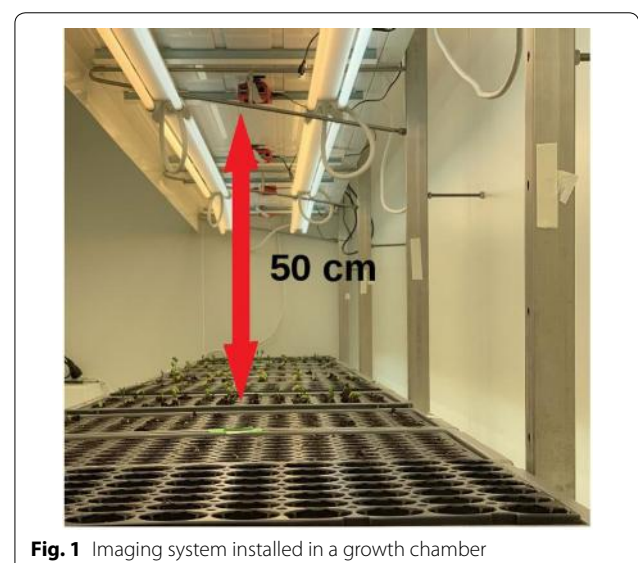


Fig. 1 Imaging system installed in a growth chamber

Dataset

Seedling establishment was recorded for 3 experiments using seed lots from different accessions of red clover (*Trifolium pratense*) (experiment 1) and alfalfa (*Medicago sativa*) (experiments 2 and 3). Each experiment consisted of 70 trays with 200 pots in which 50 seeds of four accessions were sown. Soil pots were hydrated to saturation for 24h after which excess water was removed. After 24h, seeds were sown at a depth of 2 cm, and trays were placed in a growth chamber at 20°C/16°C, with 16 h for photoperiod at 200μMm⁻²s⁻². The soil was kept humid throughout the experiment.

Each experiment took two weeks with a time-lapse of 15 minutes. In total, the database consists of 42000 temporal sequences of RGB images of size 89 × 89 × 3 pixels where each temporal sequence consists of 768 individual images. During day time, images were captured while images during night times were automatically discarded due to the absence of illumination. An example of images from the database is shown in Fig. 2. Among all temporal sequences, images of 3 randomly selected trays were manually annotated by a plant expert from the first experiment (red clover species) and 2 trays from the

second experiment (alfalfa species). This ground-truth annotation consisted of four classes: soil, the first appearance of the cotyledon (FA), the opening of the cotyledon (OC), and the appearance of the first leaf (FL). The algorithms proposed in this article for timing detection of seedling emergence following these four stages of development were trained, validated and tested against this human-annotated ground-truth. In order to avoid cross sampling, we considered images of the trays of the red clover for training (two trays) and validation (one tray) datasets. The testing dataset consisted of images of the remaining two trays from the alfalfa. Table 1 provides a synthetic view of the data set used for training and testing of the models.

Raw images were then sent to pre-processing before being applied to the deep learning method investigated in this study. A filtered variant of the raw images was also created where the soil background was removed from images. This filter was produced by applying a color filter on images in the HSV color domain to keep the green range of images in the Hue channel. This strategy was found robust because the soil used during the experiment was the same, and that lighting was kept constant.

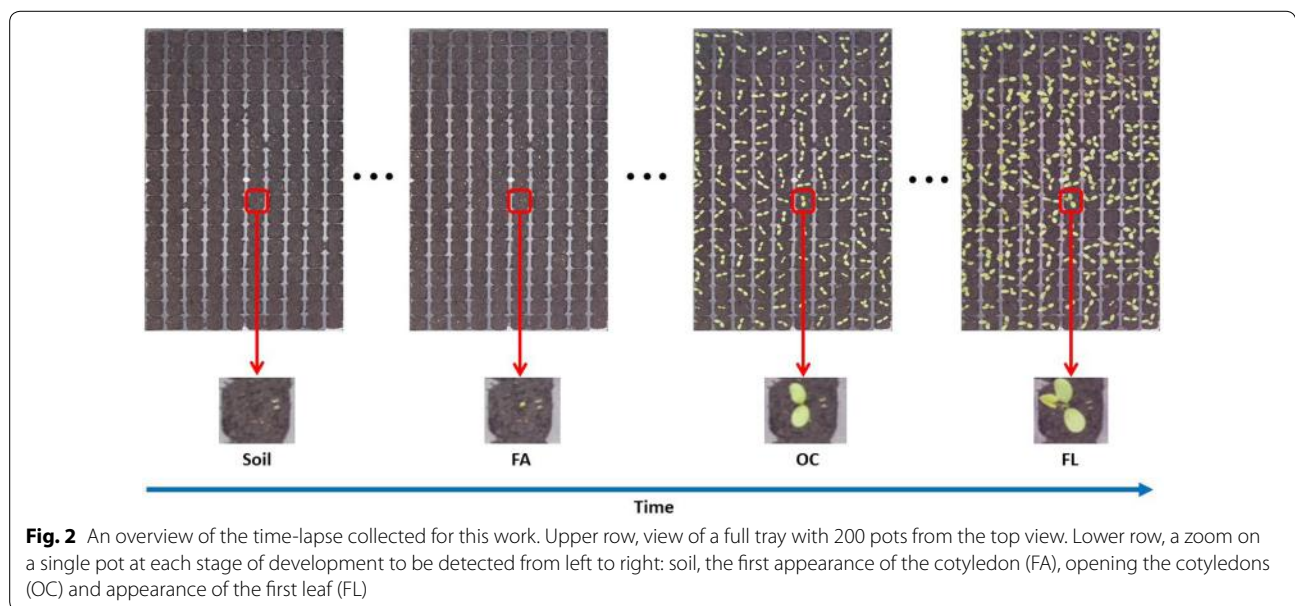


Fig. 2 An overview of the time-lapse collected for this work. Upper row, view of a full tray with 200 pots from the top view. Lower row, a zoom on a single pot at each stage of development to be detected from left to right: soil, the first appearance of the cotyledon (FA), opening the cotyledons (OC) and appearance of the first leaf (FL)

Table 1 Description of the split of the annotated data set for training models

	Species	No. of trays	No. of pots in each tray	No. of temporal sequences	Total No. of images
Training dataset	Red clover	2	200	400	307,200
Validation dataset	Red clover	1	200	200	153,600
Testing dataset	alfaalfa	2	200	400	307,200

Figure 3 shows an example of images with and without background.

Pre-processing

Since deep learning methods have to predict the seedling developmental stage on an individual basis, the raw images of Fig. 2 could not be directly applied to the neural networks. Thus, the first step of pre-processing was to extract produced crops of each pot. In order to extract them, we needed first to detect, extract, and adjust trays; then, pots were extracted from trays. Figure 4 shows a workflow of the pot extraction from trays, which includes three steps described here below.

Landmark detection

In this experiment, trays used included five white landmarks located at the center and four corners of the trays. Because of the constant control of lighting conditions,

these five landmarks were detected with a fixed threshold. Then, the five most prominent objects were kept, and the possible remaining small objects were removed. Among the five significant landmarks, the most central object in the images was considered as the central landmark. At the next steps, the four other landmarks were detected based on their minimum angle corresponding to the central landmark with horizontal and vertical axes.

Tray detection and extraction

In this step, coordinates of the trays were detected using to the landmarks. Then, based on the coordinates of these landmarks, trays could be extracted from the image. Since trays may not be positioned precisely along the axis of the vertical and horizontal axis sensor of the camera, the trays need to be rotated. The orientation of the trays was found after the computation of the angle of the first eigenvector in the principal component analysis of the modulus of the Fourier transform [29]. Finally, a geometric transformation algorithm [30] was implemented to project the rotated trays to make them straight.

Pot extraction

In the last step, all 200 pots of each tray were extracted as an independent temporal sequence of images by using a sliding window with a stride of one pot. The size of these sliding windows was made adjustable by the user to fit with the size of the pot.

This pre-processing pipeline of Fig. 4 has some generic value. Since we did not find something equivalent in the literature for our purpose, we decided to make it

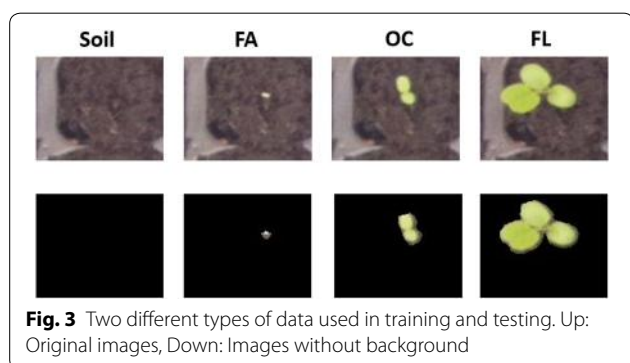


Fig. 3 Two different types of data used in training and testing. Up: Original images, Down: Images without background

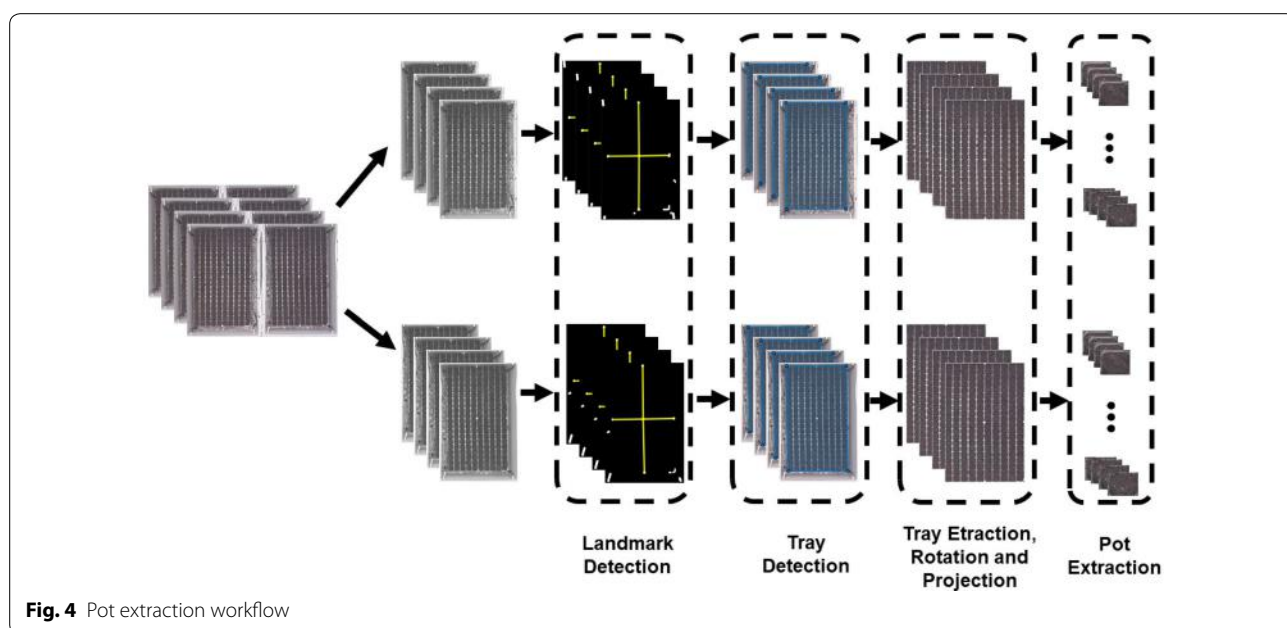


Fig. 4 Pot extraction workflow

available as supplementary material under the form of a free executable (<https://uabox.univ-angers.fr/index.php/s/HJAHp0bhZv1zy1j>). We believe that despite the simplicity of principle this can be used as a useful tool for any imaging of traits.

Deep learning methods

The three plant events plus soil (Soil, FA, OC, and FL) to be detected were expected to occur in a definite order. Different supervised strategies to take benefit from this ontological prior-knowledge on the development were tested against the manually established ground-truth as described in the following subsection.

Baseline multi-class CNN

As a naive baseline approach, we designed a convolutional neural network (CNN) architecture to predict the classes of each event of Soil, FA, OC, and FL of each frame of the time-lapses independently and without any additional information regarding the temporal order in which they should occur. Given a training set including K pairs of images x_i and labels \hat{y}_i , we trained the parameters θ of the network f using stochastic gradient descent to minimize empirical risk

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^K \mathcal{L}(\hat{y}_i, f(x_i, \theta)) \tag{1}$$

where \mathcal{L} denotes the loss function, which was chosen as cross-entropy in our case. The minimization was carried out using the ADAM optimizer [31] with a learning rate of 0.001.

Our proposed architecture $f(\cdot, \cdot)$, shown in Fig. 5, consisted of two main blocks, the feature extraction block, followed by classification block. In a CNN model, the feature extraction block takes care of

extracting features from input images by convolutional layers, and the classification block decides classes. Several CNN architectures have been deployed. First, we designed a small AlexNet [32] like CNN structure to keep the number of parameters to be learned low. This AlexNet like CNN is illustrated in Fig. 5 and reads as follows: four convolutional layers with filters of size 3×3 and respective numbers of filters 64, 128, 256, and 256 each followed by rectified linear unit (ReLU) activations and 2×2 max-pooling; a fully connected layer with 512 units, ReLU activation and dropout ($p = 0.5$) and a fully connected output layer for four classes corresponding to each event with a softmax activation. We also tested some other well-known larger CNN architectures such as VGG16 [33], Resnet50 [34], and DenseNet121 [35] on our data and choose the one with the highest performance as the base line for a naive memoryless multiclass architecture. These proposed CNN architectures have been optimized on a hold-out set.

2-class CNN's

The baseline multi-class CNN architecture of Fig. 5 is naive because it does not incorporate the prior knowledge of the ontology of plant growth to decide between different growth steps of plants plus soil (Soil, FA, OC, and FL). As a first improvement of the previous naive baseline, we implemented a variant of the CNN model of Fig. 5 dedicated to the binary classification of two consecutive stages of development. We thus trained 3 models detecting between M_1 (Soil, FA), M_2 (FA,OC) and M_3 (OC,FL). At the beginning of the analysis of an entire time-lapse sequence M_1 is used. Then when a first FA is detected M_2 is applied, and so on until the first FL detection is reached.

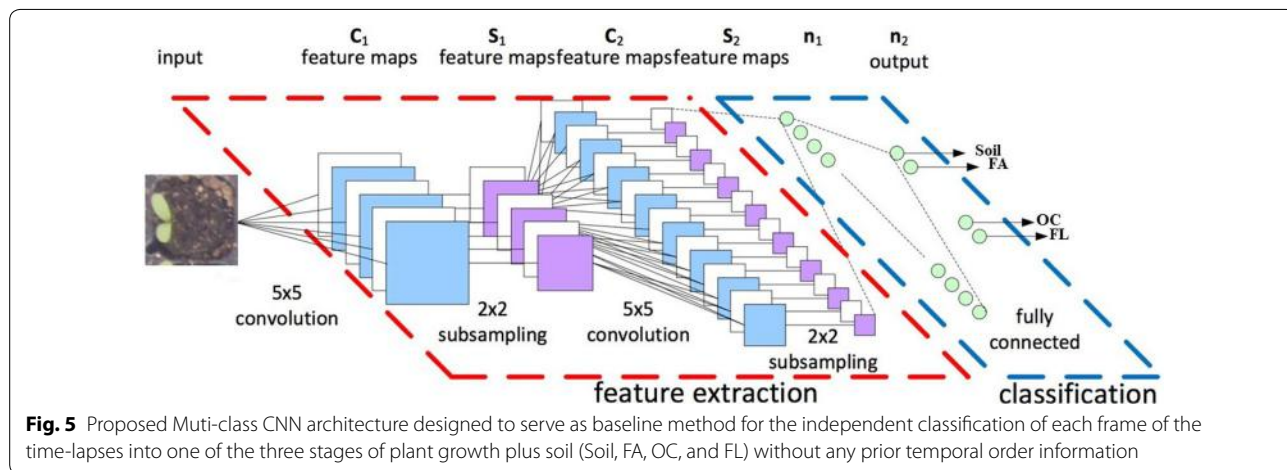


Fig. 5 Proposed Multi-class CNN architecture designed to serve as baseline method for the independent classification of each frame of the time-lapses into one of the three stages of plant growth plus soil (Soil, FA, OC, and FL) without any prior temporal order information

CNN followed by Long short-term memory

The 2-class CNN’s includes the prior knowledge of the ordered development of the seedling along with a given ontology. However, this prior knowledge is added on top of the CNN. In order to bring a memory directly inside the CNN model, the Long-Short Term Memory (LSTM) architecture was embedded between the feature extraction block and the classification block of the proposed CNN model. LSTM has been proposed [36, 37]. LSTM as a special RNN structure has proven stable and powerful for long-range modeling dependencies in various previous studies [37–39]. The major innovation of LSTM is its memory cell c^t , which essentially acts as an accumulator of the state information. The cell is accessed, written, and cleared by several self-parameterized controlling gates. Every time a new input comes, its information will be accumulated to the cell if the input gate i^t is activated. Also, the prior cell status c^{t-1} could be “forgotten” in this process if the forget gate f^t is on. Whether the latest cell output c^t will be propagated to the final state h^t is further controlled by the output gate o^t . One advantage of using the memory cell and gates to control information flow is that the gradient will be trapped in the cell [37] and be prevented from vanishing too quickly. In a multivariate LSTM structure, the input, cell output, and states are all 1D vectors features from the feature extraction block of the proposed CNN model. The activations of the memory cell and three gates are given as

$$\begin{aligned}
 i^t &= \sigma(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i) \\
 f^t &= \sigma(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f) \\
 c^t &= f^t c^{t-1} + i^t \tanh(W_{xc}x^t + W_{hc}h^{t-1} + b_c) \\
 o^t &= \sigma(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^{t-1} + b_o) \\
 h^t &= o^t \tanh(c^t)
 \end{aligned} \tag{2}$$

where $\sigma()$ is the sigmoid function, all the matrices W are the connection weights between two units, and $x = (x^0, \dots, x^{T-1})$ represents the given input.

The CNN-LSTM architecture is an integration of a CNN (Convolutional layers) with an LSTM. First, the CNN part of the model process the data and extract features then the one-dimensional feature vectors feed to an LSTM model to support sequence prediction. CNN-LSTMs are a class of models that is both spatially and temporally deep and has the flexibility to be applied to a variety of vision tasks involving sequential inputs and outputs. Fig. 6 shows a schematic of a CNN-LSTM model.

The proposed CNN-LSTM model consisted of the same convolutional layers as the multi-class CNN model of Fig.4 and an LSTM layer with 128 units.

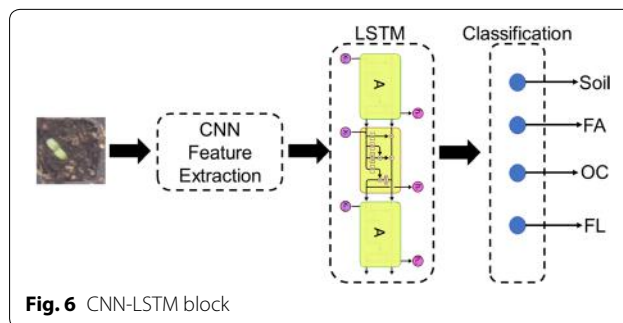


Fig. 6 CNN-LSTM block

Convolutional LSTM (ConvLSTM)

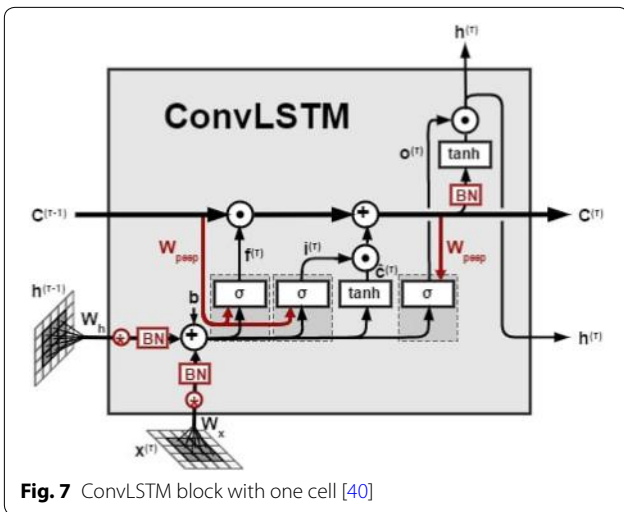
As an alternative to CNN-LSTM, we use ConvLSTM [40] which has convolutional structures in both the input-to-state and state-to-state transitions. In ConvLSTM all the inputs $X^1; \dots; X^t$, cell outputs $C^1; \dots; C^t$, hidden states $H^1; \dots; H^t$, and gates $i^t; f^t; o^t$ of the ConvLSTM are 3D tensors whose last two dimensions are spatial dimensions (rows and columns). The ConvLSTM determines the future state of a certain cell in the grid by the inputs and past states of its local neighbors. This can easily be achieved by using a convolution operator in the state-to-state and input-to-state transitions. The key equations of ConvLSTM are shown in 3 below, where ‘ \otimes ’ denotes the convolution operator.

$$\begin{aligned}
 i^t &= \sigma(W_{xi} \otimes x^t + W_{hi} \otimes h^{t-1} + W_{ci}c^{t-1} + b_i) \\
 f^t &= \sigma(W_{xf} \otimes x^t + W_{hf} \otimes h^{t-1} + W_{cf}c^{t-1} + b_f) \\
 c^t &= f^t c^{t-1} + i^t \tanh(W_{xc} \otimes x^t + W_{hc} \otimes h^{t-1} + b_c) \\
 o^t &= \sigma(W_{xo} \otimes x^t + W_{ho} \otimes h^{t-1} + W_{co}c^{t-1} + b_o) \\
 h^t &= o^t \tanh(c^t)
 \end{aligned} \tag{3}$$

Figure 7 shows a schematic of the ConvLSTM method adopted for our purposes.

Post-processing

The passing from one developmental stage to another can consist of very tiny details. This was, for instance, the case for FA and FL in our case. To address this problem, a post-processing smoothing filter can be designed to reduce the fluctuations that may appear when the seedling shift from one developmental stage to another. Also, post-processing can be of help when the first leaf moves out of the frame after a period of time and just cotyledons remain in the frame in each individual pot. In this case, the model just sees cotyledons and without post-processing would predict a label corresponding to the OC stage. Post-processing can be designed to prevent some switches forbidden by the developmental ontology and in this case keep the stage of the growth at FL.



The designed post-processing smoothing filter illustrated in Fig. 8 was based on a sliding window computing a majority voting by finding the median of classes (4)

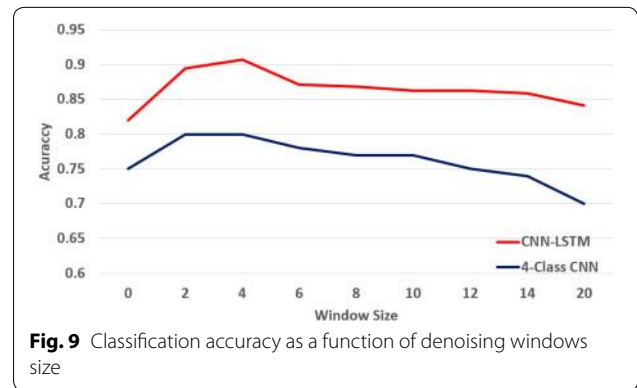
$$c = \lfloor * \rfloor \left\{ \left(\frac{n+1}{2} \right) \right\}^{th} \tag{4}$$

where c and n represent predicted class and window size, respectively. Additionally, this window replaced the current stage of all neighbors to all labels that detected as the previous stage.

The size of the sliding window was optimized on the CNN-LSTM and multi-class CNN architecture. As shown in Fig. 9, performances were found optimal for both architectures on the training data set for a size of 4 frames, corresponding to an observation of 1 hour in our case.

Results and discussion

First, we compared the performance of the tested CNN multi-class structures as shown in Table 2. As expected the performance of deeper architectures like ResNet50



and DenseNet121 is less than smaller deep models such as our proposed model or VGG16. Indeed, increasing parameters in a CNN model lead to over-fitting due to low image dimensions and limited variability in the database [41]. For the following, we keep the best multi-class structure (our proposed CNN of Fig. 5) as baseline model to be compared with other architectures including temporal information.

The proposed deep learning methods multi-class CNN, 2-class CNN's, CNN-LSTM, and ConvLSTM were applied to the dataset produced by our imaging system after pre-processing and post-processing as described in the previous section. We now present and discuss the associated results. The performances of the different deep learning methods tested on our dataset were assessed with classical metrics such as accuracy, error, sensitivity, specificity, precision, and false alarm positive rate. They are provided in Tables 3 and 4, respectively, for images with and without soil background.

Tables 3 and 4 show that all methods performed better than the naive multi-class CNN architecture, which was processing the temporal frames independently of any prior knowledge on the order of the ontological development of seedling. The best strategy to incorporate this knowledge among the ones tested was found to

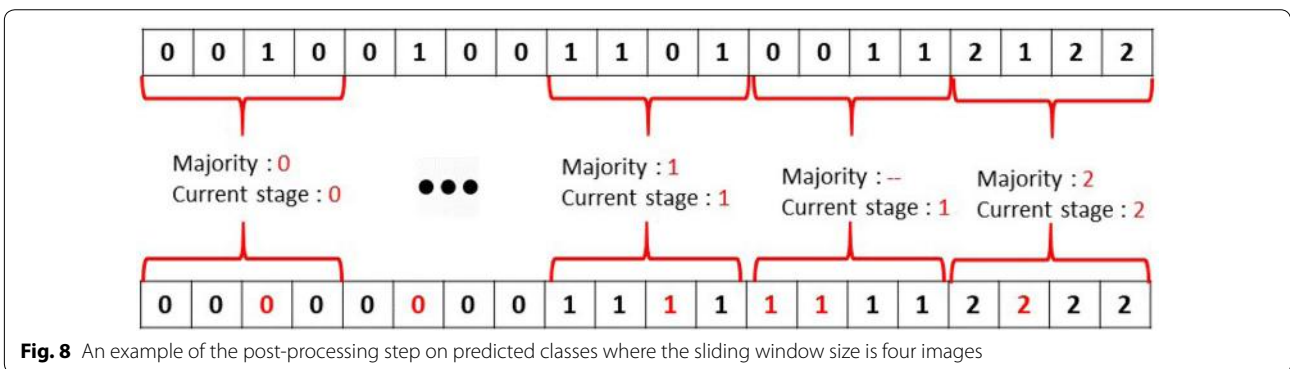


Table 2 The average performance of baseline multi-class CNN models with different evaluation metrics on images without soil background

Model	Accuracy	Error	Sensitivity	Specificity	Precision	False positive rate
Proposed CNN	0.80 ± 0.19	0.20 ± 0.19	0.85 ± 0.13	0.93 ± 0.07	0.85 ± 0.14	0.07 ± 0.07
VGG16	0.80 ± 0.24	0.2 ± 0.24	0.84 ± 0.18	0.93 ± 0.12	0.85 ± 0.07	0.07 ± 0.11
ResNet50	0.78 ± 0.18	0.22 ± 0.18	0.77 ± 0.21	0.89 ± 0.09	0.85 ± 0.11	0.08 ± 0.05
DenseNet121	0.79 ± 0.09	0.21 ± 0.09	0.78 ± 0.08	0.90 ± 0.14	0.86 ± 0.09	0.07 ± 0.10

Table 3 The average performance of models with different evaluation metrics on images with soil background

Model	Accuracy	Error	Sensitivity	Specificity	Precision	False positive rate
Multi-class CNN	0.63 ± 0.20	0.37 ± 0.20	0.63 ± 0.2	0.94 ± 0.05	0.88 ± 0.1	0.06 ± 0.05
2-class CNN's	0.72 ± 0.25	0.28 ± 0.26	0.72 ± 0.24	0.95 ± 0.06	0.90 ± 0.11	0.08 ± 0.05
CNN-LSTM	0.83 ± 0.10	0.15 ± 0.10	0.82 ± 0.10	0.93 ± 0.06	0.85 ± 0.10	0.06 ± 0.06
ConvLSTM	0.62 ± 0.2	0.33 ± 0.2	0.68 ± 0.2	0.93 ± 0.07	0.84 ± 0.1	0.06 ± 0.06

Table 4 Average performance of models on images without soil background

Model	Accuracy	Error	Sensitivity	Specificity	Precision	False positive rate
Multi-class CNN	0.80 ± 0.19	0.20 ± 0.19	0.85 ± 0.13	0.93 ± 0.07	0.85 ± 0.14	0.07 ± 0.07
2-class CNN's	0.88 ± 0.18	0.12 ± 0.18	0.86 ± 0.10	0.95 ± 0.05	0.86 ± 0.11	0.05 ± 0.05
CNN-LSTM	0.90 ± 0.08	0.10 ± 0.07	0.87 ± 0.11	0.96 ± 0.03	0.88 ± 0.15	0.04 ± 0.04
ConvLSTM	0.81 ± 0.11	0.21 ± 0.09	0.85 ± 0.03	0.92 ± 0.09	0.85 ± 0.12	0.07 ± 0.10

Table 5 Average performance of the baseline multi-class CNN and best trained models (CNN-LSTM) on test data before and after post-processing step

Model	Accuracy	Error	Sensitivity	Specificity	Precision	False positive rate
Multi-class CNN (Before)	0.72 ± 0.29	0.28 ± 0.29	0.73 ± 0.19	0.94 ± 0.21	0.91 ± 0.13	0.8 ± 0.08
Multi-class CNN (After)	0.80 ± 0.19	0.20 ± 0.19	0.85 ± 0.13	0.93 ± 0.07	0.85 ± 0.14	0.07 ± 0.07
CNN-LSTM (Before)	0.84 ± 0.04	0.16 ± 0.04	0.83 ± 0.05	0.93 ± 0.06	0.86 ± 0.09	0.05 ± 0.05
CNN-LSTM (After)	0.90 ± 0.08	0.10 ± 0.07	0.87 ± 0.11	0.96 ± 0.03	0.88 ± 0.15	0.04 ± 0.04

be the CNN-LSTM architecture, which outperforms all other models for all tested metrics. Removing the soil numerically, clearly improves all methods while keeping the CNN-LSTM architecture as the best approach.

Our experimental results show that a reasonable recognition rate of plant growth stages detection (approximately 90%) can be achievable by the CNN-LSTM model. Additionally, we measured the performance of our best model (CNN-LSTM) and on worst model (multi-class CNN) on test data before and after post-processing. Table 5 shows that the metrics of performance are systematically improved by a significant 5 to 8%.

It is possible to have a more in-depth analysis of the remaining errors by looking at the confusion matrix of this CNN-LSTM model, as given in Table 6. This confusion matrix shows that most of the errors, almost 98%, happen between the most complicated classes of OC and FL while the remaining 2% of errors appear on borders of the first two classes of soil and FA. The confusion matrix helps us to analyse the performance of the trained model on each individual class. The F1-score of Eq. (5) is considered as one of the common metrics to analyse confusion matrices for each class by calculating the harmonic mean of precision and recall (Table 6 right) where TP, FP, and FN stands for True Positive,

Table 6 Confusion matrix and F1-score of cross-subject performance where the best deep learning method, the CNN-LSTM architecture is used

True classes	Predicted					F1-Score
	Soil	FA	OC	FL		
Soil	97531	0	0	0	Soil	0.98
FA	2591	26855	2915	0	FA	0.90
OC	0	0	58668	19556	OC	0.79
FL	0	0	8219	90610	FL	0.87

Table 7 Average performance of the trained models on images of new genotype of red cloves as well as the species of alfalfa

Model	Accuracy	Error	Sensitivity	Specificity	Precision	False positive rate
CNN-LSTM(red cloves)	0.91 ± 0.01	0.09 ± 0.01	0.88 ± 0.05	0.96 ± 0.02	0.86 ± 0.08	0.04 ± 0.03
CNN-LSTM(alfalfa)	0.90 ± 0.08	0.10 ± 0.07	0.87 ± 0.11	0.96 ± 0.03	0.88 ± 0.15	0.04 ± 0.04

False Positive, and False Negative respectively. It shows that the trained model can perform better on the first two classes of Soil and FA with the highest scores of 0.98 and 0.90 on predicted data while the class of OC is the most challenging class.

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (5)$$

In order to evaluate the robustness and transferability of the best trained model (CNN-LSTM), an additional test was done on images of 50 pots of another genotypes of the red clove species which were captured from a new experiment. Table 7 shows that the average classification accuracy on the new genotype are very close to the one obtained with alfalfa. This confirms the transferability and robustness of the model from one genotype to another.

One may wonder where the classification errors in this experiment can come from. In our error analyses, we found four different sources of errors in the experiment. The first source of errors can come from the different cotyledons and leaf sizes of the two species, as the cotyledons and leaf size of a species can be much bigger or smaller compared with other species. Usually, this type of error happens in the borders of two classes of OC and FL. Figure 10 shows an example of these differences in the size of two plant species. Data augmentation with a variation on the zoom could be a solution to help with these errors.

The second source of errors can be due to the circadian cycle of plants during the growth. The circadian

**Fig. 10** A sample of images from two plant species used for training (left) and testing (right) dataset

cycle of plants makes some movements on cotyledon and leaves during day and nights [42]. This type of error can happen at the border of FA and OC, where these movements make a delay for the detection of fully opening cotyledon. Also, this type of error can happen at the border of two classes of OC and FL, where the circadian cycle does not allow the system to recognize the appearance of the first leaf from the middle of the cotyledon. The third source of errors happens due to the overlapping of plants in a tray. Plants grow at different speeds and directions in a tray, and it makes overlapping on plants of neighbor pots at some points. This type of error usually happens in the last two classes of OC and FL. The last source of the errors can come from annotation errors. In general, the annotation of plant growth stages is challenging since plants grow continuously; it means there are no striking events of growth. In this case, a class represents a period of growth. For instance, the FA class is assigned to images which were captured in the period of the first appearance of the cotyledon till the time of the fully opening of the

cotyledon. In this case of annotation, different annotators may define the ending of a stage period with an approximate delay of 15 images. Also, there is a period of formation of the first leaf before its unfolding during plant growth. This period is considered to be a part of the FL class in this experiment. This consideration may bring an additional error for annotation of stages as different annotators may recognize the beginning of the leaf formation with a delay.

Conclusion and perspectives

In this paper, we have presented a complete image processing and machine learning pipeline to classify three stages of plantlet growth plus soil on the different accessions of two species of red clover and alfalfa.

Different strategies were compared in order to incorporate the prior information of the order in which the different stages of the development occur. The best classification performance on these types of images was found with our proposed CNN-LSTM model, which achieved 90% accuracy of detection with the help of a denoising algorithm incorporating the ontological order in the development stages.

In our experiments all models were trained and tested on several genotypes of two species of red clover and alfalfa. Presented results shows that trained model is robust on some genotypes but it does not guaranty the robustness of the model an all genotypes or other species. In order to increase the robustness of models one could either add more real data from several genotypes or use data augmentation to synthetically increase the data variability in the training database [43–45] based on possible priors on the expected morphological plasticity of the species.

These results can now be extended in various directions. It will be interesting to extend the approach to a range of species of agricultural interest in order to provide a library of trained networks. From this perspective, it could be interesting to investigate quantitatively how, by their similarity in shape, the knowledge learned on some species could be transferred to others via transfer learning, domain adaptation, or hierarchical multi-label classification [46]. More events of the development of plants could also be added to extend the investigation of seedling kinetics. This includes for instance the instant where cotyledons are out of soil fully or rise of the first leaf before unfolding. These extensions could be tested easily following the global methodology presented in this article to assess the deep learning models. For even more advanced stages of development and yet still accessible from top view, the issue of plant overlapping each other would arise and become a limitation. Solving this would require to switch to tracking algorithms in order follow

and label the trajectory of each plant despite ambiguity created by partial occlusion and overlapping. Other deep learning architectures would have to be tested in this perspective [47]. As another possible direction, in this study, since we used classical standard RGB images, plants were not measured during nights, and some missed events could shift the estimation of the developmental stages of the seedlings. Lidar cameras, accessible at low-cost [48], could be used to access to night events. Also, Bayesian approaches [6], such as Gaussian processes, could be used to estimate the time for the possibly missing information.

Acknowledgements

Imaging systems and biological samples benefit from funding from the European Union's Horizon 2020 project EUCLEG under Grant Agreement No. 727312. Salma Samiei gratefully acknowledges Région des Pays de la Loire for the funding of her Ph.D.

Additional materials

The pre-processing method for extracting the individual pots from raw images is provided at <https://uabox.univ-angers.fr/index.php/s/HJAHp0bhZv1zy1j>.

Authors' contributions

SS, PR, DR, JLV, and JB conceived and designed this work. PR and JLV carried out the acquisitions. SS, PR, DR, and JB conceived and interpreted the whole data. SS, PR, DR developed the image processing algorithms and associated annotation tools. JLV and JB carried out image annotation. SS, PR, DR, and JB wrote and revised the manuscript. PR, DR, and JB supervised the work. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Data will be available after acceptance upon reasonable request

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRAE IRHS, Université d'Angers, Angers, France. ²Institut de Recherche en Horticulture et Semences-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, Beaucouzé, France. ³Department of Big Data and Data Science, École d'ingénieur Informatique et Environnement (ESAIP), Angers, France.

Received: 1 April 2020 Accepted: 23 July 2020

Published online: 30 July 2020

References

1. Arsovski AA, Galstyan A, Guseman JM, Nemhauser JL. Photomorphogenesis. *The Arabidopsis Book/American Society of Plant Biologists*. 2012;10:
2. Dellen B, Scharr H, Torras C. Growth signatures of rosette plants from time-lapse video. *IEEE/ACM Trans Comput Biol Bioinf*. 2015;12(6):1470–8.

3. Minervini M, Giuffrida MV, Perata P, Tsafaris SA. Phenotiki: An open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *Plant J.* 2017;90(1):204–16.
4. Tovar JC, Hoyer JS, Lin A, Tielking A, Callen ST, Elizabeth Castillo S, Miller M, Tessman M, Fahlgren N, Carrington JC. Others: Raspberry Pi-powered imaging for plant phenotyping. *Appl Plant Sci.* 2018;6(3):1031.
5. Choudhury SD, Samal A, Awada T. Leveraging image analysis for high-throughput plant phenotyping. *Front Plant Sci.* 2019;10:508.
6. Humplik JF, Dostál J, Ugena L, Spíchal L, De Diego N, Vencálek O, Fürst T. Bayesian approach for analysis of time-to-event data in plant biology. *Plant Methods.* 2020;16(1):14.
7. Minervini M, Scharf J, Tsafaris SA. Image analysis: the new bottleneck in plant phenotyping [applications corner]. *IEEE Signal Process Mag.* 2015;32(4):126–31.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–44.
9. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. *Comput Electron Agric.* 2018;147:70–90.
10. McCormac AC, Keefe PD, Draper SR. Others: automated vigour testing of field vegetables using image analysis. *Seed Sci Technol.* 1990;18(1):103–12.
11. Sako Y, McDonald MB, Fujimura K, Evans AF, Bennett MA. A system for automated seed vigour assessment. *Seed Sci Technol.* 2001;29(3):625–36.
12. Hoffmaster AL, Fujimura K, McDonald MB, Bennett MA. An automated system for vigor testing three-day-old soybean seedlings. *Seed Sci Technol.* 2003;31(3):701–13.
13. Marcos-Filho J, Bennett M, McDONALD M, Evans A, Grassbaugh E. Assessment of melon seed vigour by an automated computer imaging system compared to traditional procedures. *Seed Sci Technol.* 2006;34(2):485–97.
14. Marcos Filho J, Kikuti ALP, de Lima LB. Procedures for evaluation of soybean seed vigor, including an automated computer imaging system. *Revista Brasileira de Sementes.* 2009;31(1):102–12.
15. Joosen RVL, Kodde J, Willems LAJ, Ligterink W, van der Plas LHW, Hilhorst HWM. germinator: a software package for high-throughput scoring and curve fitting of arabidopsis seed germination. *Plant J.* 2010;62(1):148–59.
16. Belin É, Rousseau D, Rojas-Varela J, Demilly D, Wagner M-H, Cathala M-H, Dürr C. Thermography as non invasive functional imaging for monitoring seedling growth. *Comput Electron Agric.* 2011;79(2):236–40.
17. Benoit L, Belin É, Dürr C, Chapeau-Blondeau F, Demilly D, Ducournau S, Rousseau D. Computer vision under inactinic light for hypocotyl-radicle separation with a generic gravitropism-based criterion. *Comput Electron Agric.* 2015;111:12–7.
18. Marcos Filho J. Seed vigor testing: an overview of the past, present and future perspective. *Scientia Agricola.* 2015;72(4):363–74.
19. Gnädinger F, Schmidhalter U. Digital counts of maize plants by unmanned aerial vehicles (uavs). *Remote sens.* 2017;9(6):544.
20. Sadeghi-Tehran P, Sabermanesh K, Virlet N, Hawkesford MJ. Automated method to determine two critical growth stages of wheat: heading and flowering. *Front Plant Sci.* 2017;8:252.
21. Rasti P, Demilly D, Benoit L, Belin É, Ducournau S, Chapeau-Blondeau F, Rousseau D. Low-cost vision machine for high-throughput automated monitoring of heterotrophic seedling growth on wet paper support. In: *BMVC*; 2018. p. 323.
22. Chen R, Chu T, Landivar JA, Yang C, Maeda MM. Monitoring cotton (*Gossypium hirsutum* L.) germination using ultrahigh-resolution uas images. *Prec Agric.* 2018;19(1):161–77.
23. Zhao B, Zhang J, Yang C, Zhou G, Ding Y, Shi Y, Zhang D, Xie J, Liao Q. Rapeseed seedling stand counting and seeding performance evaluation at two early growth stages based on unmanned aerial vehicle imagery. *Front Plant Sci.* 2018;9:1362.
24. Jiang Y, Li C, Paterson AH, Robertson JS. Deepseedling: deep convolutional network and Kalman filter for plant seedling detection and counting in the field. *Plant Methods.* 2019;15(1):141.
25. Kipp S, Mistele B, Baresel P, Schmidhalter U. High-throughput phenotyping early plant vigour of winter wheat. *Eur J Agron.* 2014;52:271–8.
26. Sankaran S, Khot LR, Carter AH. Field-based crop phenotyping: multispectral aerial imaging for evaluation of winter wheat emergence and spring stand. *Comput Electron Agric.* 2015;118:372–9.
27. Suresha M, Kuppa S, Raghukumar D. A study on deep learning spatiotemporal models and feature extraction techniques for video understanding. *Int J Multimedia Inf Retr.* 2020;1–21:
28. Sakurai S, Uchiyama H, Shimada A, Taniguchi R-i. Plant growth prediction using convolutional lstm. In: 14th International Conference on Computer Vision Theory and Applications, VISAPP 2019-Part of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2019; 2019. p. 105–113. SciTePress
29. Gonzalez RC, Woods RE, Masters BR. *Digital Image Processing.* 3rd ed.; 2009.
30. Szeliski R. *Computer Vision. Texts in Computer Science.* London: Springer; 2011.
31. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In: *ICML*; 2015.
32. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*; 2012. p. 1097–105.
33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition; 2014. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
34. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 770–778.
35. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2017. p. 4700–8.
36. Graves A, Mohamed A-r, Hinton G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing; 2013. p. 6645–9. IEEE
37. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–80.
38. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: *International conference on machine learning*; 2013. p. 1310–8.
39. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*; 2014. p. 3104–12.
40. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-k, Woo W-c. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In: *Advances in Neural Information Processing Systems*; 2015. p. 68–80. [arXiv:1506.04214](https://arxiv.org/abs/1506.04214)
41. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imag.* 2018;9(4):611–29.
42. Samiei S, Rasti P, Chapeau-Blondeau F, Rousseau D. Cultivons notre jardin avec Fourier. In: 27ème Colloque GRETSI sur Le Traitement du Signal et des Images, Lille, France; 2019.
43. Harisubramanyabalaji SP, ur Réhman S, Nyberg M, Gustavsson J. Improving image classification robustness using predictive data augmentation. In: *International conference on computer safety, reliability, and security.* Springer; 2018. p. 548–61.
44. Zheng S, Song Y, Leung T, Goodfellow I. Improving the robustness of deep neural networks via stability training. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 4480–8.
45. Fawzi A, Samulowitz H, Turaga D, Frossard P. Adaptive data augmentation for image classification. In: 2016 IEEE international conference on image processing (ICIP); 2016. p. 3688–92.
46. Dyrmann M, Skovsen S, Jørgensen RN. Hierarchical multi-label classification of plant images using convolutional neural network
47. Jin J, Dundar A, Bates J, Farabet C, Culurciello E. Tracking with deep neural networks. In: 2013 47th annual conference on information sciences and systems (CISS); 2013. p. 1–5. IEEE.
48. Chéné Y, Rousseau D, Lucidarme P, Bertheloot J, Caffier V, Morel P, Belin É, Chapeau-Blondeau F. On the use of depth camera for 3d phenotyping of entire plants. *Comput Electron Agric.* 2012;82:122–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Article

Enhancing the Tracking of Seedling Growth Using RGB-Depth Fusion and Deep Learning

Hadhami Garbougé ¹, Pejman Rasti ^{1,2} and David Rousseau ^{1,*}

¹ Université d'Angers, Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRAE IRHS, 62 Avenue Notre Dame du Lac, 49000 Angers, France; hadhami.garbougé@univ-angers.fr (H.G.); pejman.rasti@univ-angers.fr (P.R.)

² Centre d'Etudes et de Recherche pour l'Aide à la Décision (CERADE), École D'ingénieur Informatique et Environnement (ESAIP), 49124 Angers, France

* Correspondence: david.rousseau@univ-angers.fr

Abstract: The use of high-throughput phenotyping with imaging and machine learning to monitor seedling growth is a tough yet intriguing subject in plant research. This has been recently addressed with low-cost RGB imaging sensors and deep learning during day time. RGB-Depth imaging devices are also accessible at low-cost and this opens opportunities to extend the monitoring of seedling during days and nights. In this article, we investigate the added value to fuse RGB imaging with depth imaging for this task of seedling growth stage monitoring. We propose a deep learning architecture along with RGB-Depth fusion to categorize the three first stages of seedling growth. Results show an average performance improvement of 5% correct recognition rate by comparison with the sole use of RGB images during the day. The best performances are obtained with the early fusion of RGB and Depth. Also, Depth is shown to enable the detection of growth stage in the absence of the light.



Citation: Garbougé, H.; Rasti, P.; Rousseau, D. Enhancing the Tracking of Seedling Growth Using RGB-Depth Fusion and Deep Learning. *Sensors* **2021**, *21*, 8425. <https://doi.org/10.3390/s21248425>

Academic Editor: Francesco Buonomici

Received: 28 October 2021
Accepted: 14 December 2021
Published: 17 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: deep learning; plant growth; CNN; RGB-Depth; image fusion; feature fusion; transformers

1. Introduction

The detection of the seedling growth stages is a fundamental problem in plant science. This covers the emergence of seedling from the soil, the opening of cotyledons and appearance of the first leave which correspond to the earliest stages of development of plant. The success or failure of these developmental stages and their kinetics have a huge impact on the evolution of the future plant. Recently, seedling growth monitoring has received attention from the computer vision community [1–16]. Among these works, the state-of-the-art approach based on deep learning proposed in [16] has shown the possibility to automatically classify the stages of development of seedling with RGB sequences of images from top view with an accuracy higher than 90%.

One of the limitations of the work proposed in [16] is that the monitoring was done only during daylight with RGB images. Consequently, any events happening during the night would be missed and/or possibly estimated with a temporal bias. In this article, we propose an extension of the work of [16] and investigate the possibility to push forward the monitoring of the seedling growth during the day and the night. To this purpose, RGB-Depth camera were used. These technologies have been demonstrated of wide value in plant phenotyping [17–24]. The depth images are computed by an active LIDAR camera operating in infrared (IR). This camera can be activated during day and night without impact on the development of the plants. As in [16] we selected low-cost versions of these RGB-Depth cameras. These low-cost constraints are specially important in plant phenotyping [25] when moving the plants or the camera is not an option and that replication of cohorts of cameras is to be chosen to monitor large populations of plants. Low-cost RGB-Depth cameras are also logically coming with artifacts and noise. Such artifacts and

metrological limitations of low-cost RGB-Depth cameras have been extensively studied (see [26] for a recent survey). In our case, we rather work at an informational level. We focus on a classification task, i.e., a nonlinear decision, which is by nature more robust to noise since it does not have to provide a high-fidelity, metrological, linear estimation. The hypothesis investigated in this article is that these low-cost RGB-Depth sensors despite their limited spatial resolution and the presence of artifacts may be of enough value to enhance the tracking of seedling growth during day and night.

We demonstrate, for the first time, to the best of our knowledge the value of these RGB-Depth images to monitor the early stages of seedling growth. We investigate fusion strategies between RGB and depth with several neural networks architecture. The underlying motivation to use multimodal data is that complementary information give a richer representation that may be utilized to create better results than a single modality. The multimodal fusion research community has made significant progress in the past decade [27]. Different fusion strategies have been reviewed [28,29]. Specifically for RGB and Depth with deep learning architectures, fusion has been extensively studied in the literature [30–41]. Mainly two types of fusion can be distinguished. First, images can be stacked at the input: this is the early fusion [30–34], that we call image fusion. Second, deep features can be independently extracted and then fused before a classification stage: this is the feature fusion [35–38]. In this work, we investigate these fusions scenarios that we applied to the important problem of seedling growth stage monitoring. Since we process sequences of images we considered time-dependent neural network architectures. As in [16], we included a base line convolutional neural network (CNN) and LSTM [42]. We also added TD-CNN GRU [43] and transformer [44] which were not included in [16].

2. Materials and Methods

2.1. Imaging System and Data Set

We have conducted similar experiments as the ones described in detail in [16] and shortly recalled here. A set of minicomputers, connected to RGB-Depth cameras [45], was used to image seedlings from the top view as illustrated in Figure 1. We used, instead of the RGB cameras of [16], Intel real sense cameras [46] (model D435) which natively produces registered RGB-Depth pairs of images and calibrated Depth maps. We installed 8 of these RGB-Depth cameras in a growth chamber where cameras followed the growth of seedlings from top view. During experiment, soil pots were hydrated to saturation for 24 h after which excess water was removed. After 24 h, seeds were sown at a depth of 2 cm, and trays were placed in a growth chamber at 20 °C/16 °C, with 16 h for photoperiod at 200 $\mu\text{Mm}^{-2} \text{s}^{-2}$. The soil was kept wet throughout the experiments. Each experiment took one week with a frame rate of 15 min. The time lapse program (made in Python) was implemented on a central minicomputer controlling, via ethernet wires, the 8 minicomputers connected to the RGB-Depth cameras.

Concerning the biological material, seedling growth was recorded for 2 experiments using seed lots from different accessions of beans such as Flavert, Red Hawk, Linex, Caprice, Deezer and Vanilla. Each experiment consisted of 3 trays with 40 pots in which 120 seeds of accessions were sown. There is a similarity between the species in this experiment and the two species which were used in [16] as all of them consist in dicotyledon species. The main difference between them comes from the number of varieties in this experiment which is three times higher than the one in [16].

In total, the database consists of 72 temporal sequences of RGB and depth images of size 66×66 pixels where each temporal sequence consists of 616 individual images. Example of images from the database is shown in Figure 1. RGB-Depth temporal sequences acquired during daylight were annotated by expert in biology while looking at RGB images. This ground-truth annotation consisted of four classes: soil, first appearance of the cotyledon (FA), opening of the cotyledon (OC), and appearance of the first leaf (FL). The algorithms presented in this paper for seedling emergence identification following these four phases of growth were trained, validated, and tested against this human-annotated

ground-truth. In order to train robust models, we used the cross-validation approach by considering image sequences of bean varieties in three split of train, validation, and test dataset. Table 1 provides a synthetic view of the data set used for training and testing of the models. For the training dataset, we applied data augmentation using a simple horizontal flip on each temporal sequence.

Table 1. Description of the RGB-Depth dataset used in this study.

	Species	No. of Temporal Sequences	Totale No. of Images during Days	Totale No. of Images during Nights	Totale No. of All Images
Training dataset	Flavert	10	4240	1920	36,960
	Red Hawk	10	4240	1920	
	Linex	10	4240	1920	
	Caprice	10	4240	1920	
	Deezer	10	4240	1920	
	Vanilla	10	4240	1920	
Validation dataset	Flavert	1	424	192	3696
	Red Hawk	1	424	192	
	Linex	1	424	192	
	Caprice	1	424	192	
	Deezer	1	424	192	
	Vanilla	1	424	192	
Testing dataset	Flavert	1	424	192	3696
	Red Hawk	1	424	192	
	Linex	1	424	192	
	Caprice	1	424	192	
	Deezer	1	424	192	
	Vanilla	1	424	192	

Depth images can contain artifacts with missing values. This can happen on part of the scene where not enough light is reflected or for objects that are too close or too far from the camera. While neural networks should be able to cope with such noise, it is better to correct them to use the capability of these networks on clean data. In order to correct these artifacts, we applied a classical inpainting technique [47] of depth images to reduce the noise.

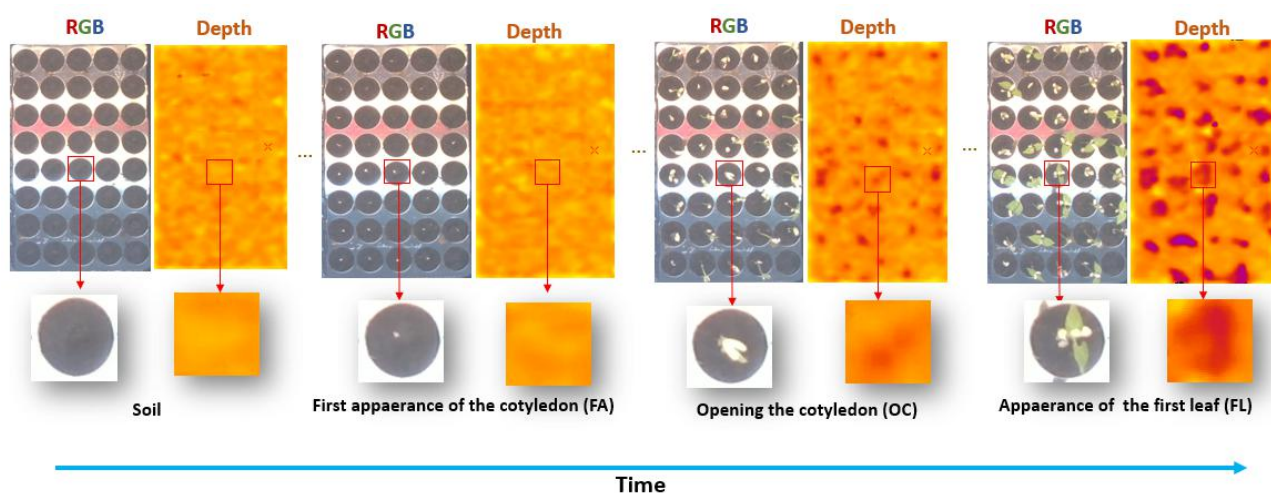


Figure 1. Overview of the time-lapse collected for this work. Upper row, view of a full tray with 72 pots from top view. Lower row, a zoom on a single pot at each stage of development to be detected from left to right: soil, first appearance of the cotyledon (FA), opening the cotyledons (OC) and appearance of the first leaf (FL).

2.2. RGB-Depth Deep Learning Fusion Strategies

We describe here the different neural network architectures tested in this study to fuse the RGB and Depth for the classification of seedling growth stages as depicted in Figure 2.

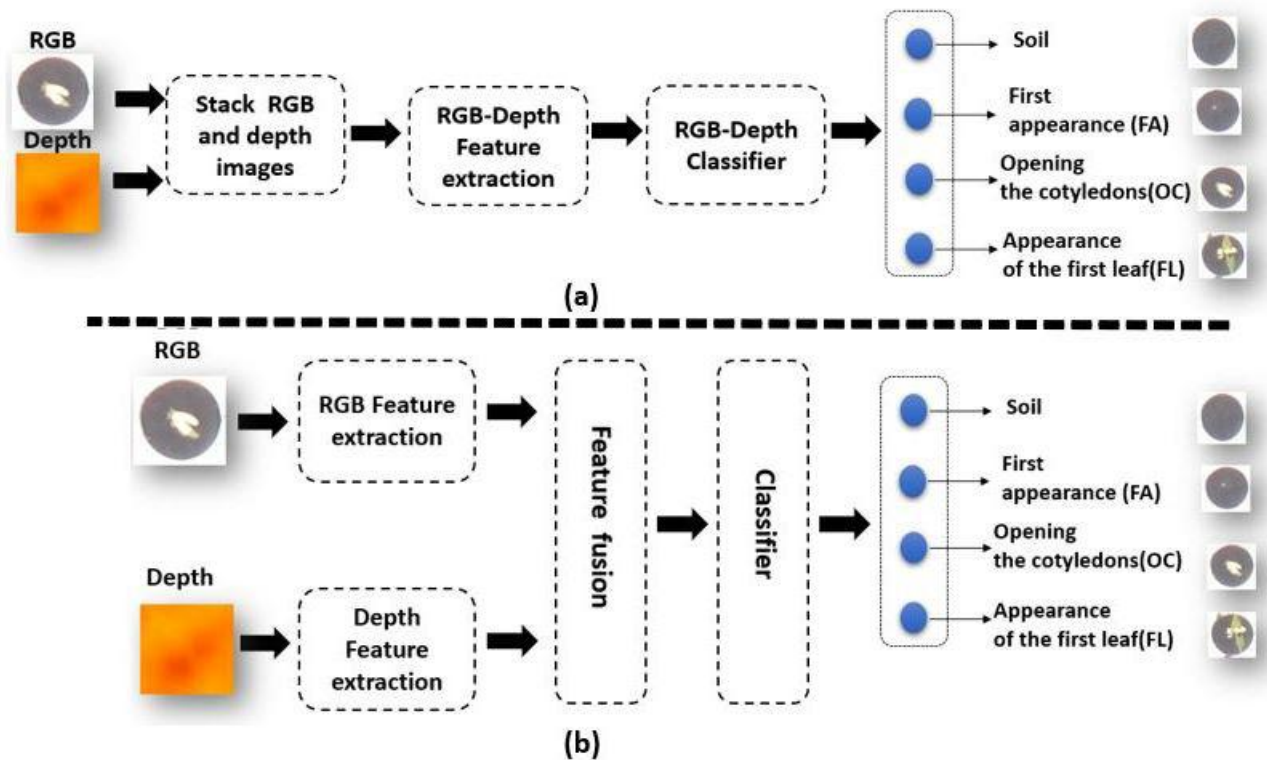


Figure 2. Different types of RGB-Depth fusion architectures tested in this article for image classification. (a) Image-based RGB-Depth fusion, (b) Feature-based RGB-Depth fusion.

2.2.1. CNN-Based Image Early Fusion Learning Structure

We first integrated, as in [48], RGB and Depth data stacked in a four-channel as input to a CNN (see Figure 3a). The feature extraction block from four-channel input images is followed by the classification block (shown in Figure 3a). The CNN architecture is the one of [16,43] that we shortly recall. The feature extraction block of a CNN model is responsible for extracting features from input images using convolutional layers, whereas the classification block determines classes. To keep the amount of train parameters low, we created an AlexNet [49] like CNN structure. This architecture reads as follows: four convolutional layers with filters of size 3×3 and respective numbers of filters 64, 128, 256, and 256 each followed by rectified linear unit (ReLU) activations and 2×2 max-pooling; a fully connected layer with 512 units, ReLU activation and dropout ($p = 0.5$) and a fully connected output layer for four classes corresponding to each event with a softmax activation. This proposed CNN architecture has been optimized on a hold-out set and was demonstrated in [16] to be optimal by comparison with other standard classical architectures (VGG16, ResNet, DenseNet). The network was trained from scratch since the size of the input tensor (4 channels and small spatial resolution) was different from existing pre-trained networks on large RGB data sets.

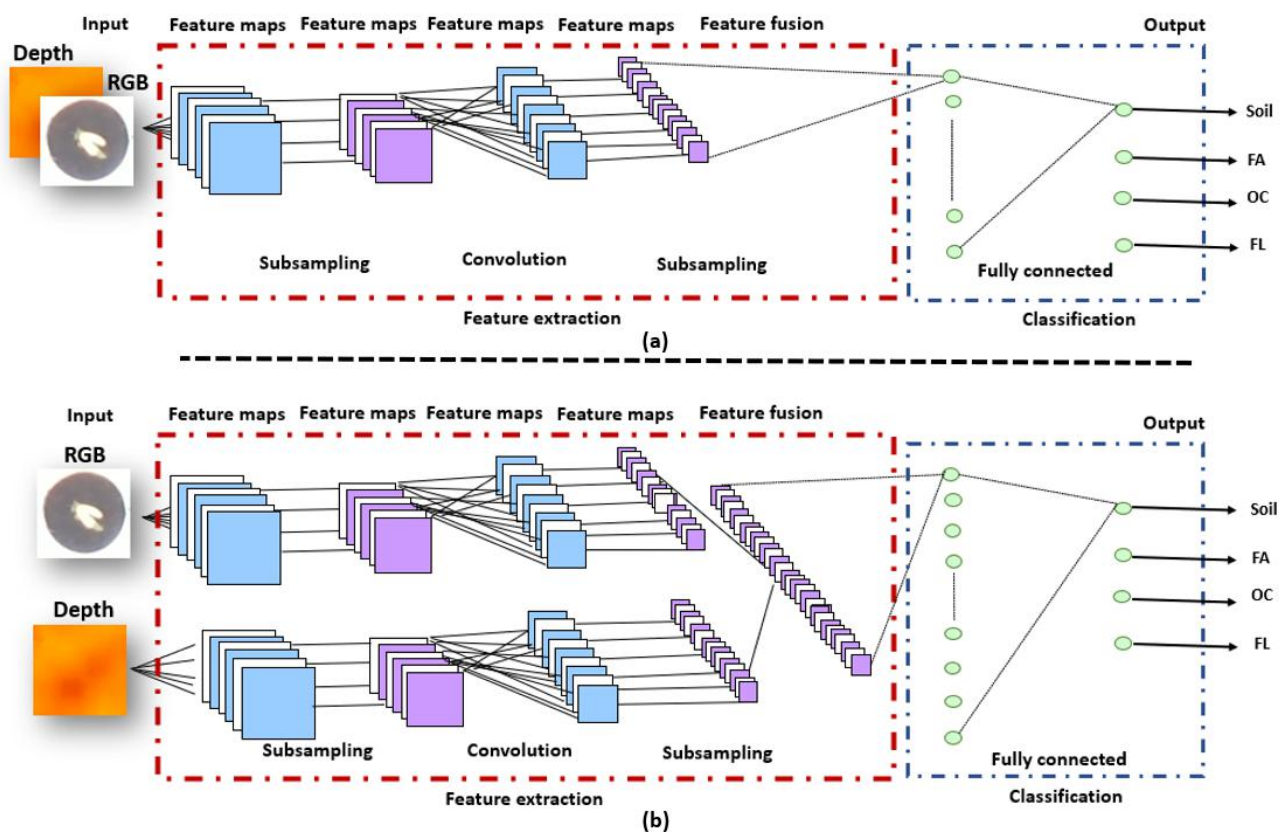


Figure 3. (a) CNN architecture of image fusion for RGB-Depth, (b) CNN architecture of features fusion for RGB-Depth.

2.2.2. CNN-Based Feature Fusion Learning Structure

Our architecture, shown in Figure 3b, is made up of two convolutional network streams that operate on RGB and Depth data, respectively. The same structure of image fusion CNN has been developed for each stream of the feature fusion CNN. The feature extractor part of the CNN architectures of RGB and Depth images consists of four convolutional layers which have 64, 128, 256, and 256 filters, respectively (similar to the AlexNet like structure of the previous subsection). The ReLU activation function is considered for each convolutional layer followed by a max-pooling layer. On the classification part of the CNN architectures, a fully connected layer with 512 units, and an output layer with four neurons corresponding to each event with a softmax activation function.

2.2.3. TD-CNN-GRU-Based Image and Feature Fusion Learning Structure

We demonstrated in [16,43] the possible added value to embed in controlled environment a memory in the process of the sequence of images. We demonstrated in [43], the superiority of Time dependent CNN with gated recurrent units (TD-CNN-GRU) by comparison with other memory based methods such as long short term memory (LSTM) and CNN-LSTM architectures. GRU uses two gates: the update gate and the reset gate while there are three gates in LSTM. This difference makes GRU faster to train and with better performance than LSTMs on less training data [50]. The same CNN architecture of our model in [16] was embedded in our TD-CNN-GRU model where the optimal duration of the memory was found to be 4 images in [16,43] corresponding to 1 hour of recording. Figure 4 shows a schematic view of the proposed TD-CNN-GRU for images and feature fusion respectively.

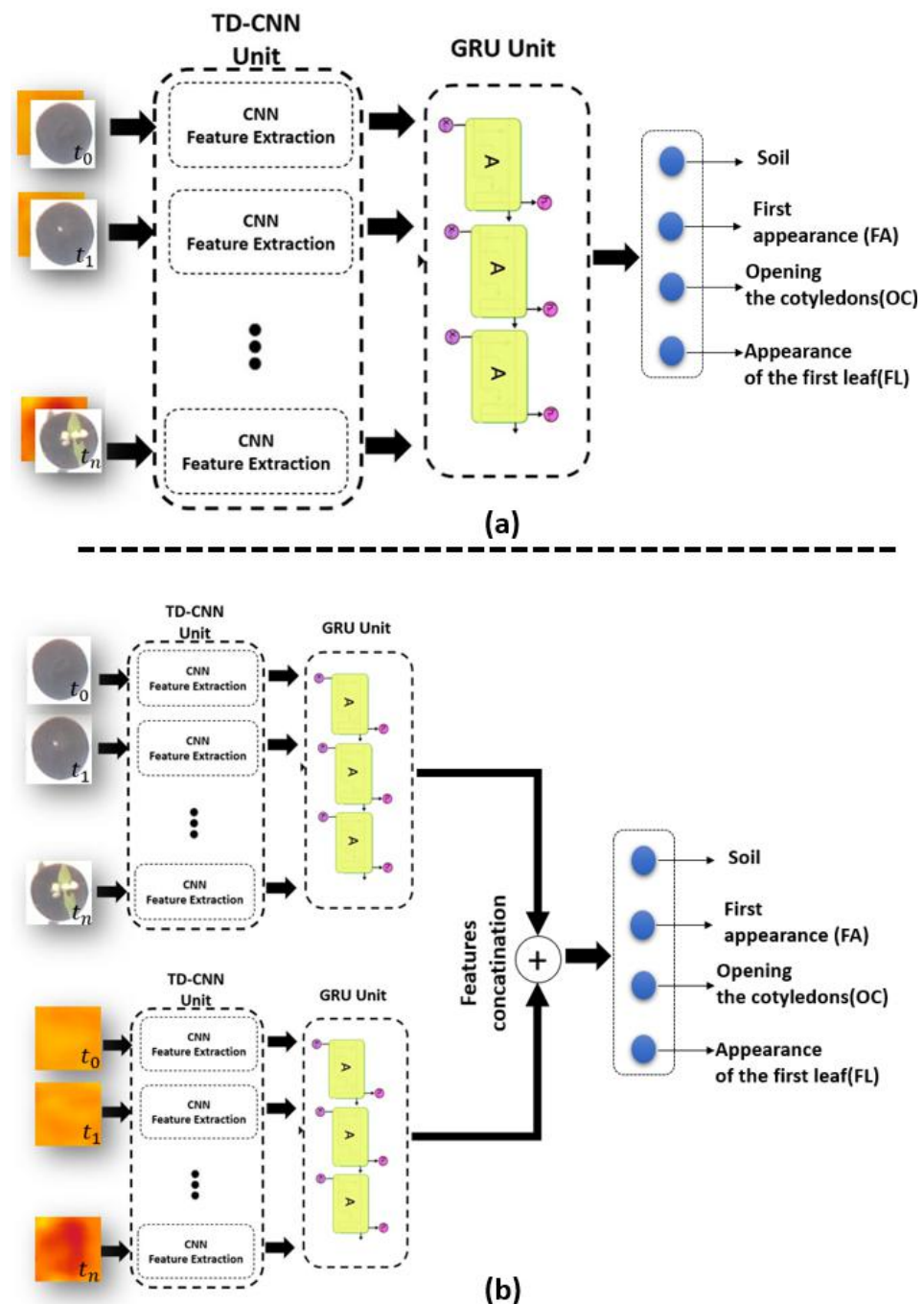


Figure 4. (a) TD-CNN-GRU architecture of image fusion for RGB-Depth, (b) TD-CNN-GRU architecture of features fusion for RGB-Depth.

2.2.4. Transformers-Based Image and Feature Fusion Learning Structure

A last class of neural network dedicated to time series are the transformers. Since their introduction in [44] they have been shown to outperform recurrent neural networks such as LSTM and GRU specially in the field of natural language processing as they do not require that the sequential data be processed in order. Transformers have been shown suitable to process temporal information carried by single pixels in satellite images time series [51–53]. Transformers have recently been extended to the process of images [54] where images were analysed as a mosaic of subparts of the original images creating artificial time series. In our case, we directly have meaningful original images which corresponds to the field of view of the pots. We, therefore, provide the transformer of [54] with time

series of consecutive images of the same pot (we used the same time slot as in the other spatio-temporal methods). We used 32 transformer layers with batch size 64, feed forward layer as classification head layer and the size of our patch size was equal to 66×66 pixels for both architectures of Figure 5.

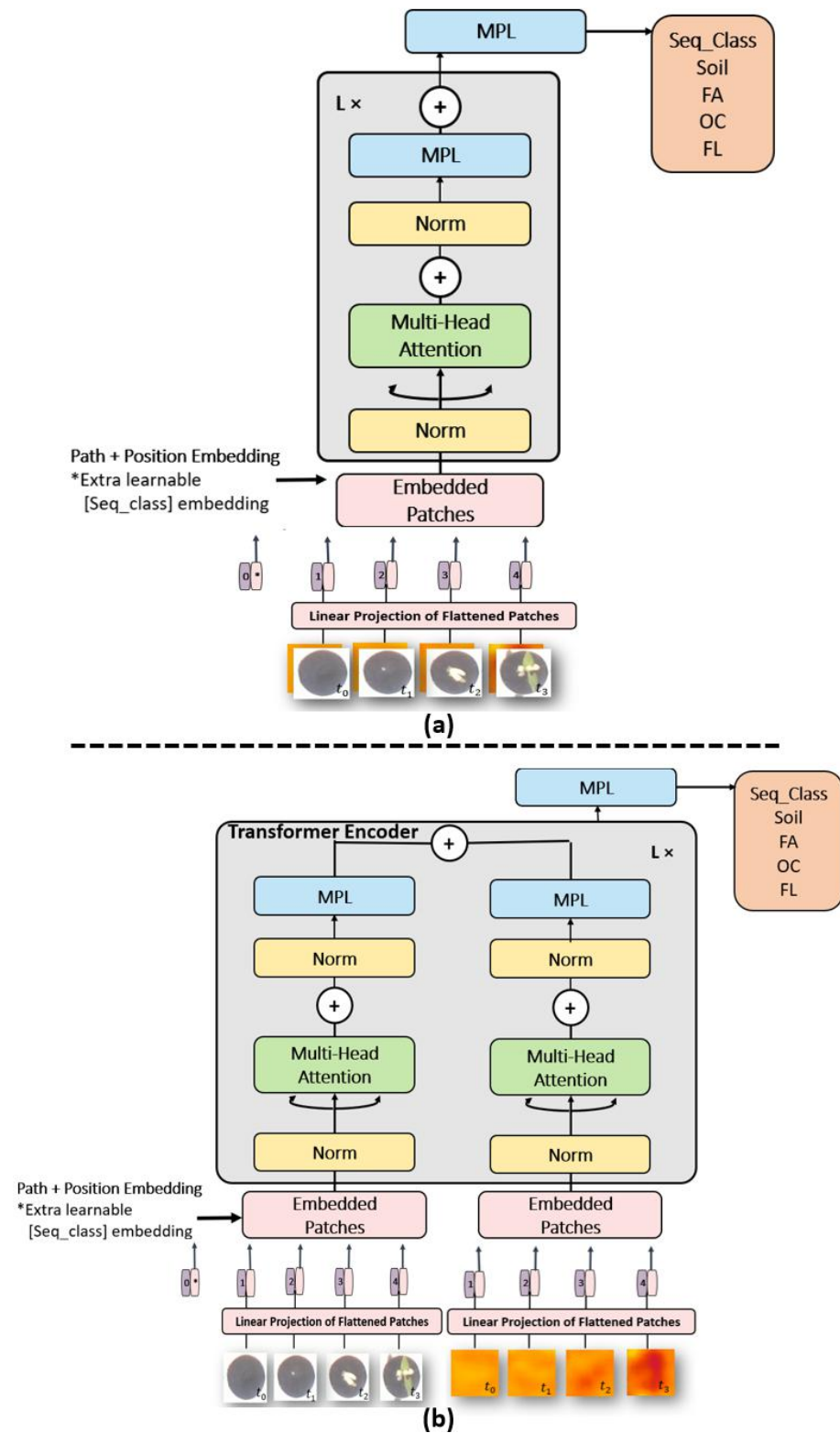


Figure 5. (a) Transformer architecture of image fusion for RGB-Depth, (b) Transformer architecture of features fusion for RGB-Depth.

For all our training, we used the NVIDIA DGX station. This station is composed of 4 GPUs and each one of them have a RAM memory of 32 Gb. We used Python version 3.7.8, Tensor-flow version 2.7.0 and Keras library version 2.3.1.

2.3. Accuracy

The performances of the different fusion strategies tested on our dataset were classically assessed with Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP, TN, FP, and FN stands for true positive, true negative, false positive, and false negative).

3. Results

3.1. Fusion Strategies

The proposed deep learning methods CNN, TD-CNN-GRU, and Transformers with image or feature RGB-Depth fusion were applied to the produced dataset as described in the Section 2. The performances are provided in Tables 2–4 and Figure 6.

Table 2. Seedling growth stage classification average accuracy and standard deviation when performed over 10 repetitions of CNN model.

	Training	Validation	Test
RGB	0.95 ± 0.02	0.91 ± 0.03	0.88 ± 0.05
Image fusion RGB-Depth	0.97 ± 0.02	0.95 ± 0.02	0.94 ± 0.04
Features fusion RGB-Depth	0.97 ± 0.01	0.96 ± 0.01	0.94 ± 0.01

Table 3. Seedling growth stage classification average accuracy and standard deviation when performed over 10 repetitions of TD-CNN-GRU model.

	Training	Validation	Test
RGB	0.87 ± 0.02	0.85 ± 0.01	0.80 ± 0.01
Image fusion RGB-Depth	0.91 ± 0.01	0.87 ± 0.02	0.82 ± 0.01
Features fusion RGB-Depth	0.90 ± 0.01	0.86 ± 0.02	0.81 ± 0.01

Table 4. Seedling growth stage classification average accuracy and standard deviation when performed over 10 repetitions of transformer model.

	Training	Validation	Test
RGB	0.90 ± 0.02	0.86 ± 0.01	0.82 ± 0.01
Image fusion RGB-Depth	0.96 ± 0.02	0.91 ± 0.01	0.88 ± 0.03
Features fusion RGB-Depth	0.92 ± 0.03	0.89 ± 0.02	0.84 ± 0.01

Tables 2–4 show that all methods performed better when RGB and Depth data are fused by comparison with the sole use of RGB data. This improvement is obtained both with image fusion and with feature fusion. This demonstrate the value of RGB-Depth fusion with a gain of 5% (on average) compared to the use of the sole RGB images. This is obtained at a reasonable training time of around 1 to 3 h as detailed in Table 5. The best results are obtained with the CNN method, i.e., the spatial method by comparison with the spatio-temporal method. This CNN is showing the best absolute performance, the smallest training time and also minimum decrease of performance between training, validation and test. This is in agreement with our previous results found in [16,43], where spatio-temporal

methods outperformed memoryless spatial ones only when the kinetic of growth were homogeneous among the dataset. This was not the case in this study.

Table 5. Training time of the different deep learning architectures.

	Model	Training Time
RGB	CNN	1 h 00 min
	Transformer	1 h 30 min
	TD-CNN-GRU	3 h 00 min
Image fusion RGB-Depth	CNN	1 h 15 min
	Transformer	1 h 35 min
	TD-CNN-GRU	3 h 30 min
Features fusion RGB-Depth	CNN	1 h 20 min
	Transformer	1 h 30 min
	TD-CNN-GRU	3 h 20 min

The confusion matrix of the CNN method is displayed in Figure 6 for RGB images and RGB-Depth images. Interestingly errors with both RGB and RGB-Depth only occur on adjacent classes along the developmental order. These are situations where even the human eye can have uncertainty to decide the exact time of switching from one class to the next one. Remaining errors can thus be considered as reasonable errors. The confusion matrices also clearly demonstrate that the main gain brought by the Depth channel is on the stage of opening the cotyledons for which the error are divided by a factor two. First appearance out of the soil, or the appearance of the first leave produce very limited variations on the depth. By contrast, the opening of the cotyledons produces an abrupt variation of the Depth. Therefore, the impact of Depth on the improvement of the performance of classification on this developmental stage is consistent with this rationale. Following also this rationale, one can notice that the errors on opening the cotyledon slightly increase when Depth is added but the overall impact of Depth is on average beneficial to the global accuracy.

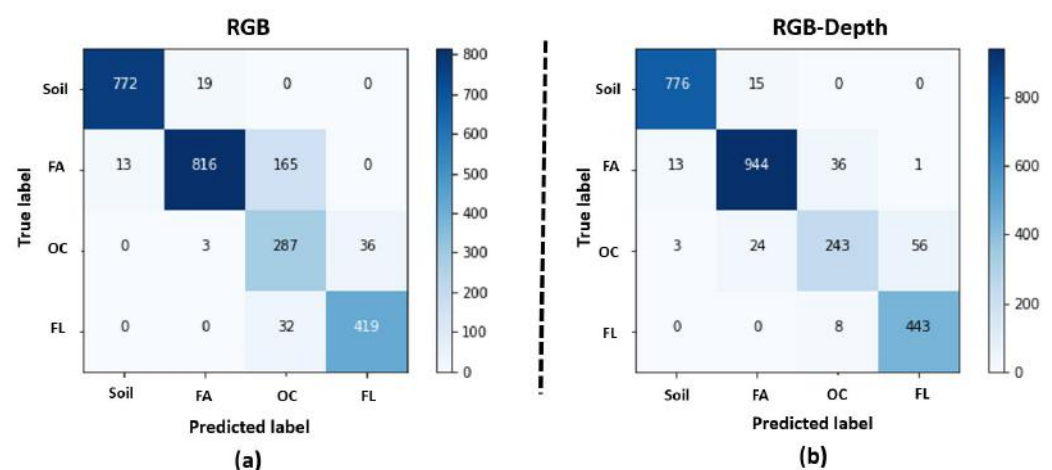


Figure 6. Confusion matrices for the best method found in Table 2, i.e., CNN. (a) for the RGB images and (b) for the RGB-Depth images.

3.2. Detection of Event Changes at Night Using Depth Information

The advantage of using the depth is not limited to enhance the performance during the day as shown in the previous subsection. Depth is also expected to be specifically useful during the night since the RGB cameras are then non operating while the Depth

images can still be acquired. If the growth stage switches during the night the RGB imaging devices detect the switch only on the first frame of the next day time as illustrated in Figure 7. It is possible to screen for Depth alone during these nights and observe the start of a growth pattern actually occurring before the beginning of the day. We demonstrate in this subsection how to take benefit quantitatively of the sole Depth channel during these nights.

We analyzed the number of switches from one growth stage to another happening on the first image acquired during the day in the data set of [16] and found out that it represented 35 percent of the events (see Figure 8). This is similar to what we found with the dataset of this article where we had 100 sets of pots from different varieties. In these frames, we have 115 switches of growth stages with 43 happening during night time. While some could be triggered by the action of light others could also happen earlier during the night. To detect a possible change during the night, we quantitatively used Depth. We designed Algorithm 1 which acts as follows. We first detects nights where a switch between a growth stage to another growth stage is found in RGB images. During these nights, the algorithm then detects the depth frame on which the switch is the most likely to occur. In short, this is obtained by choosing the time where the average spatial depth is permanently (computed over a sliding window of 4 images = 1 h) closer to the average spatial depth of the next growth stage.

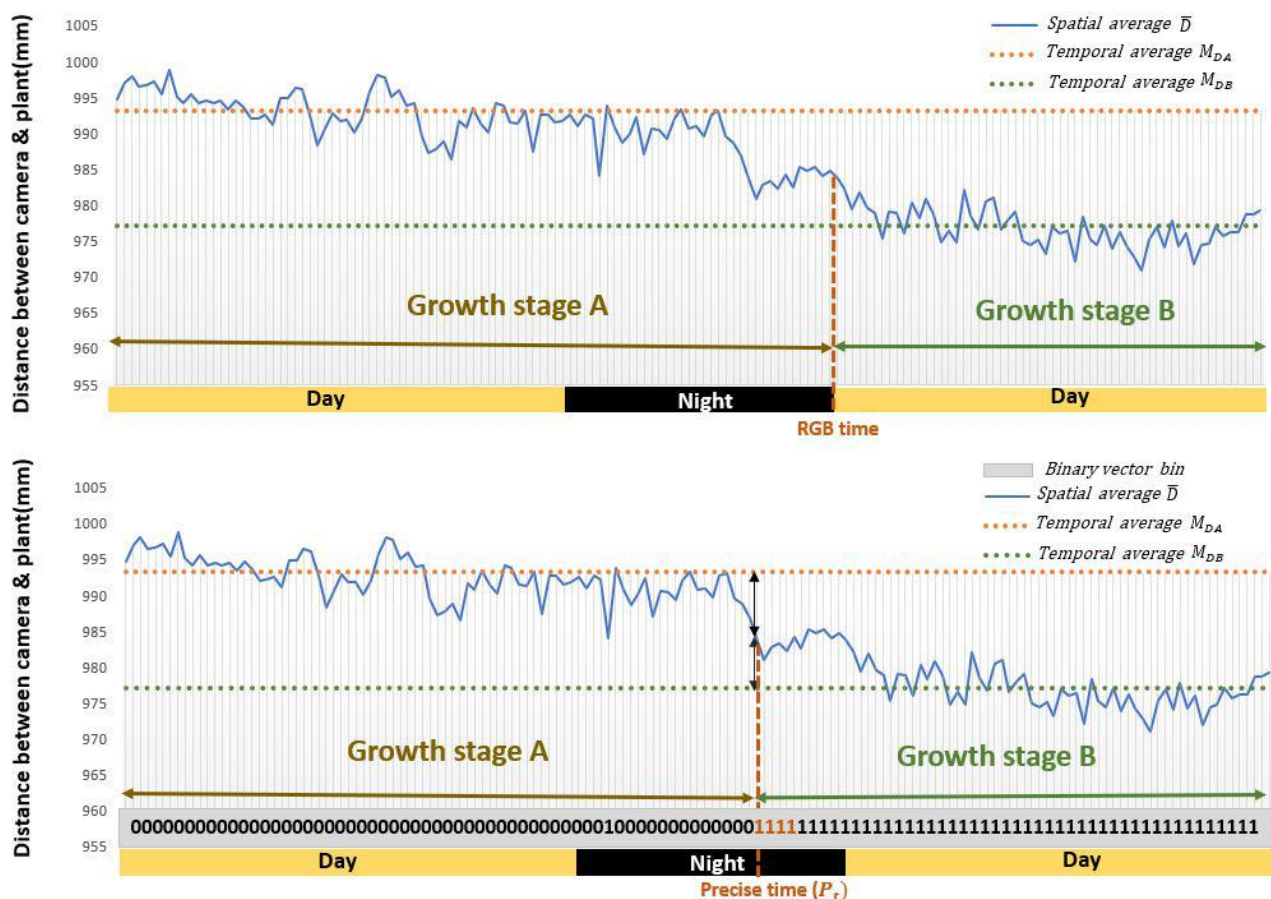


Figure 7. First row: the detection of switch from growth stage A to growth stage B using only daytime RGB images. Second row: the more precise detection of switch from growth stage A to growth stage B using the Depth pattern during the night time as proposed by Algorithm 1.

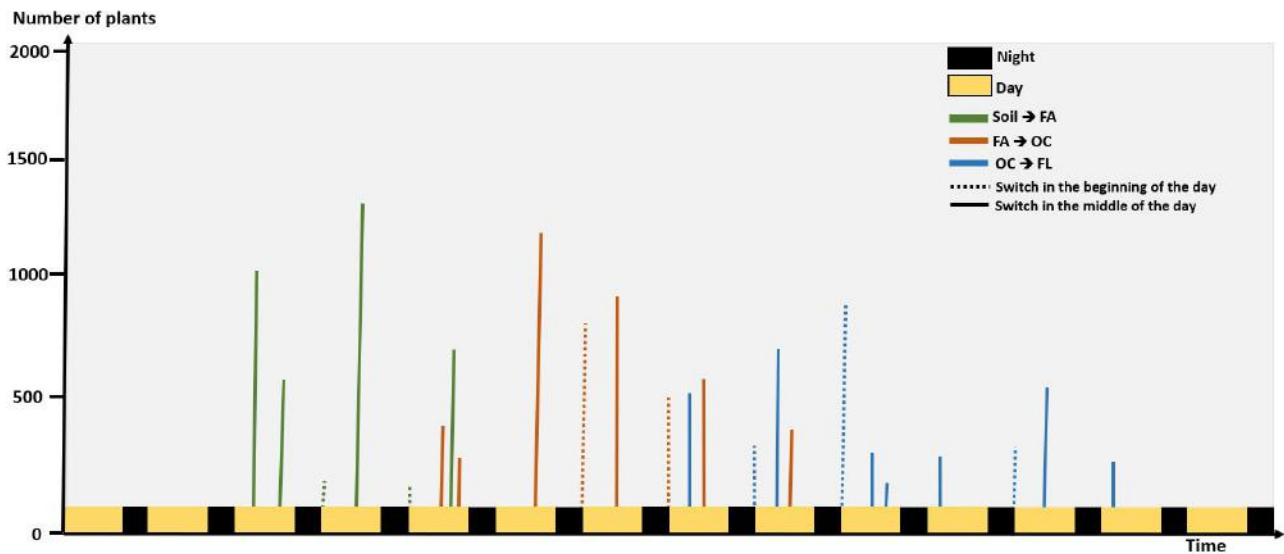


Figure 8. Histogram of detection of growth stage change during day and night from 4000 plants.

Algorithm 1: Detection of night events using depth information.

Input:

S^{night} = Sequences of depth images of a night during which a switch a growth stage is observed in RGB images.

S^a = Sequences of depth images from the last day before the switch of growth stage A to B.

S^b = Sequences of depth images from the first day after the switch of growth stage A to B.

Output: P_t = Precise time of switch of growth stage.

- 1 $\overline{DA} \leftarrow \text{mean}(S^a);$ ▷ Spatial average of S^a
 - 2 $\overline{DB} \leftarrow \text{mean}(S^b);$ ▷ Spatial average of S^b
 - 3 $\overline{DN}_k \leftarrow \text{mean}(S^{night});$ ▷ Spatial average of S^{night}
 - 4 $\langle M_{DA} \rangle \leftarrow \text{mean}(\overline{DA});$ ▷ Temporal average of \overline{DA}
 - 5 $\langle M_{DB} \rangle \leftarrow \text{mean}(\overline{DB});$ ▷ Temporal average of \overline{DB}
 - 6 $GA \leftarrow \overline{DN} - \langle M_{DA} \rangle;$ ▷ Difference between \overline{DN} and $\langle M_{DA} \rangle$
 - 7 $GB \leftarrow \overline{DN} - \langle M_{DB} \rangle;$ ▷ Difference between \overline{DN} and $\langle M_{DB} \rangle$
 - 8 $bin \leftarrow \text{sign}(GA - GB);$ ▷ Binary vector of the sign for the difference between GA and GB
 - 9 $Idx \leftarrow \text{find}(bin == 1111);$ ▷ Get the index of first pattern (1111) in the binary vector.
 - 10 $P_t \leftarrow \text{Length}(S^a) + Idx;$ ▷ Add the length of S^a to the index of the first pattern (1111) to get the precise time
-

To validate Algorithm 1, we could not establish ground truth during the night. As a workaround, we used daylight events and applied the depth channel only to the Algorithm 1. Then, we used the annotated ground truth obtained from the RGB images to compute the performance of Algorithm 1. We found 80% of these 115 switches with a shift of less than 4 frames on average (standard deviation of 2 frames) by comparison with the manually annotated ground truth. This corresponds to an uncertainty (bias here) of 1 h which is very reasonable and much lower than the error duration of the night itself (8 h) if no Depth were used.

4. Discussion

We analyzed the remaining errors of the proposed algorithms and discuss them in this section together with some open perspectives of the work.

Two main sources of errors can be attached to the acquisition protocol and instrumentation itself. These are illustrated in Figure 9. First, some seedlings growth so fast that their leaves or cotyledons go out of the observation window (Figure 9a). This causes drop in depth and change in the RGB pattern. With our current approach, we do focus on individual pots. For such seedlings growing at early stages outside of their pot, we would need to either use larger pots or develop tracking algorithms. This falls outside of the scope of this study which focused on the added value of Depth when fused to RGB for the detection of early growth stages of seedlings. Another source of errors happens due to noise on the Depth channel (Figure 9b). Such noises were observed when too much or too low amount of IR light was reflected on pots. This happens for instance when the plastic material of the pots has a high reflectance or when some remaining water (absorbing IR) is present. These noises can be reduced by carefully choosing the material used for the pot and the watering process. Another type of error comes from the inherent large heterogeneity of shapes and sizes of the bean varieties considered in this study and illustrated in Figure 10. This affects specially the detection of growth stage which shows the tiniest changes, i.e., the opening of the cotyledons. To solve these errors, one could simply add more data or use more advanced data augmentation techniques such as zoom, stretch, color jitter, ... We wanted to provide basic results here which already happen to be of rather high quality without the use of such approach to robustify the model since the main goal was the fusion of the RGB and Depth for seedling growth monitoring.

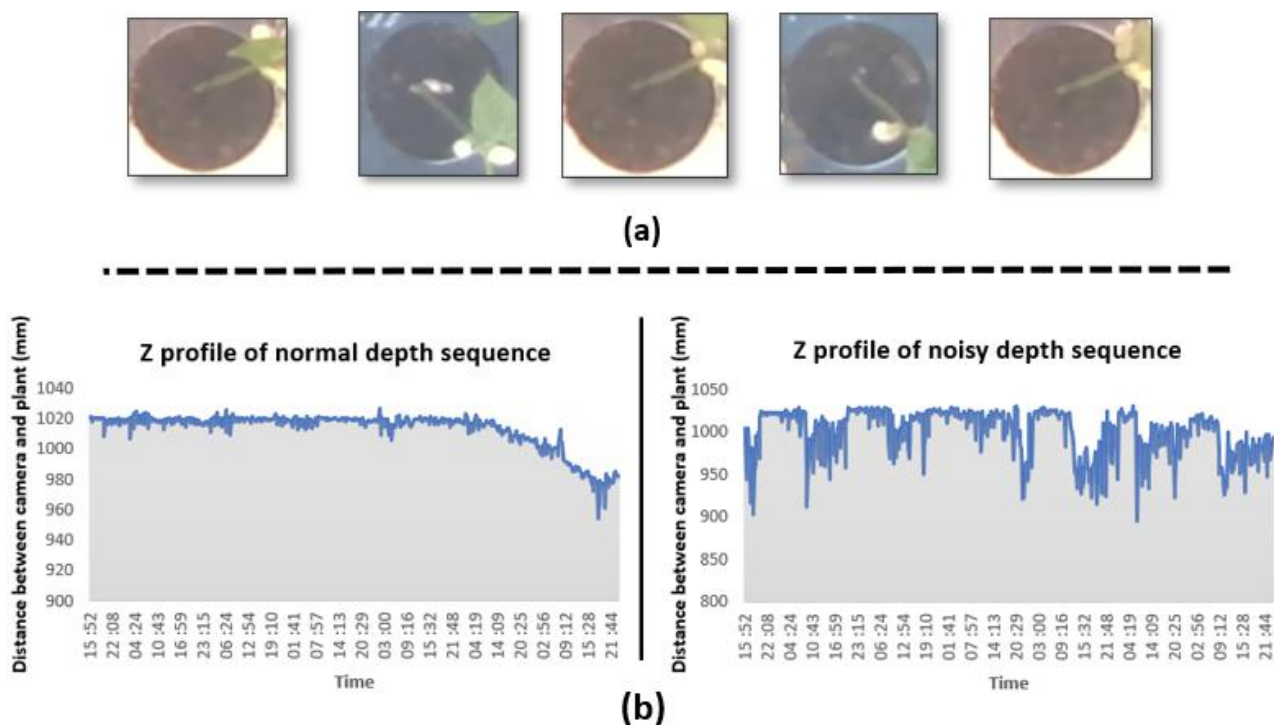


Figure 9. Sources of errors due to the acquisition protocol (a) and instrumentation (b).













Varieties	Flavert	Red Hawk	Linex	Caprice	Deezer	Vanilla
Cotyledons shape						
Cotyledons size (pixels)	576	710	132	165	221	256
First leaf shape						
First leaf size (pixels)	1482	2280	743	853	736	1764

Figure 10. Heterogeneity of shape and size in the two events OC and FL for the different bean varieties used in the training.

One may wonder about the robustness of the model proposed given the relatively small size of the plant population considered. First, the overfit measured with the best method was found to be limited together with the difference of performance between cultivars. It is important to recall here that the point of the work is to quantify the added value of RGB-Depth images by comparison with sole RGB. This is what we do on the same data sets. Interestingly, the performance with RGB images obtained with only 72 samples are similar to the larger data set used in [16] (90% against 88% here). However, we cannot ensure a perfect robustness to large change of phenotypic shapes. If such variability in scale was expected, larger data sets would have to be constituted. The comparison between RGB and RGB-Depth would remain unchanged.

In this work, we focused on early fusion and feature fusion of RGB and Depth. One may also consider decision fusion where the classification from the RGB image and the Depth image would be made. We performed this analysis and found a pure random decision when the classification was made on Depth alone. Therefore, at the decision level, no added value of Depth was to be expected on average. Fusion between RGB and Depth for such small images and low-cost sensors as the one considered in this study is found to be beneficial on average at earlier stages of processing (image or features). However, after analysing the confusion matrix in detail, one could imagine to selectively using the added value of Depth at the stages of growth where it is expected to be the most significant. This was found to be between the FA and OC in our case and more generally when large contrast in Depth happens. On the contrary, one could discard the use of Depth when the growth process is estimated to lay at stages where no contrast in Depth is expected (between Soil and FA in our case).

This work could be developed in several other future directions. First, we could revisit this study with higher resolution Depth sensors [26] to investigate how the reduction of noise and improvement of resolution in Depth could help to further improve the classification results. More advanced stages of development yet still accessible from the top view, could be investigated without targeting 3D reconstruction [55]. An issue comes with the possible overlapping between plants. One solution would be to decrease the density of plants but this would come with a lower throughput for the experiments. Another solution would be to investigate the possibility to track leaves during their growth in order to decipher partial occlusions. Here again, RGB depth sensors coupled with advanced machine learning approaches could be tested to further extend the capability to monitor seedling growth [56]. Last but not least, we can now directly apply the developed algorithms to analyze biologically in detail the statistical distribution of seedling growth events at night

on large datasets. This may unravel new knowledge on the physiological impact of light on these growth kinetics in addition to their links with circadian rhythms [57].

5. Conclusions

In this article, we have demonstrated the added value of Depth when fused with RGB images for the important problem of detection of seedling growth stage development. During day time, Depth was shown to improve by 5% the classification performances on average. Also Depth was shown of value to refine the estimation of switch of growth stage during the night period. These results were established on different fusion strategies including CNN, TD-CNN-GRU and transformers. These methods were compared in order to incorporate the prior information of the order in which the different stages of the development occur. The best classification performance on these types of images was found with our optimized CNN, which achieved 94% accuracy of detection. In our experiments all models and fusion strategies were trained and tested on several genotypes of beans.

Author Contributions: Conceptualization, D.R., P.R. and H.G.; methodology, D.R. and P.R.; software, H.G.; validation, D.R., P.R. and H.G.; formal analysis, H.G.; investigation, H.G.; resources, P.R.; data curation, H.G.; writing—original draft preparation, P.R.; writing—review and editing, D.R.; visualization, H.G.; supervision, D.R.; project administration, D.R.; funding acquisition, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by PPR SUCSEED project, RFI OSMOSE project and by the Horizon 2020 Framework Program of the European Union under grant agreement No 817970 (INVITE: <https://www.h2020-invite.eu/>, accessed on 10 December 2021).

Data Availability Statement: Data are available upon reasonable request.

Acknowledgments: The authors thank Marie Simonin, Arthur Coste from INRAe Angers for management of plants and Daniel Sochard, Remi Gardet for management of growth chamber part of PHENOTIC platform node of the french phenotyping national infrastructure PHENOME.

Conflicts of Interest: The authors declare no conflict of interest.

References

- McCormac, A.C.; Keefe, P.D.; Draper, S.R. Automated vigour testing of field vegetables using image analysis. *Seed Sci. Technol.* **1990**, *18*, 103–112.
- Sako, Y.; McDonald, M.B.; Fujimura, K.; Evans, A.F.; Bennett, M.A. A system for automated seed vigour assessment. *Seed Sci. Technol.* **2001**, *29*, 625–636.
- Hoffmaster, A.L.; Fujimura, K.; McDonald, M.B.; Bennett, M.A. An automated system for vigor testing three-day-old soybean seedlings. *Seed Sci. Technol.* **2003**, *31*, 701–713. [[CrossRef](#)]
- Marcos-Filho, J.; Bennett, M.; McDonald, M.; Evans, A.; Grassbaugh, E. Assessment of melon seed vigour by an automated computer imaging system compared to traditional procedures. *Seed Sci. Technol.* **2006**, *34*, 485–497. [[CrossRef](#)]
- Marcos Filho, J.; Kikuti, A.L.P.; de Lima, L.B. Procedures for evaluation of soybean seed vigor, including an automated computer imaging system. *Rev. Bras. Sementes* **2009**, *31*, 102–112. [[CrossRef](#)]
- Joosen, R.V.L.; Kodde, J.; Willems, L.A.J.; Ligterink, W.; van der Plas, L.H.W.; Hilhorst, H.W. germinator: A software package for high-throughput scoring and curve fitting of Arabidopsis seed germination. *Plant J.* **2010**, *62*, 148–159. [[CrossRef](#)] [[PubMed](#)]
- Belin, É.; Rousseau, D.; Rojas-Varela, J.; Demilly, D.; Wagner, M.H.; Cathala, M.H.; Dürr, C. Thermography as non invasive functional imaging for monitoring seedling growth. *Comput. Electron. Agric.* **2011**, *79*, 236–240. [[CrossRef](#)]
- Benoit, L.; Belin, É.; Dürr, C.; Chapeau-Blondeau, F.; Demilly, D.; Ducournau, S.; Rousseau, D. Computer vision under inactinic light for hypocotyl–radicle separation with a generic gravitropism-based criterion. *Comput. Electron. Agric.* **2015**, *111*, 12–17. [[CrossRef](#)]
- Marcos Filho, J. Seed vigor testing: An overview of the past, present and future perspective. *Sci. Agric.* **2015**, *72*, 363–374. [[CrossRef](#)]
- Gnädinger, F.; Schmidhalter, U. Digital counts of maize plants by unmanned aerial vehicles (UAVs). *Remote Sens.* **2017**, *9*, 544. [[CrossRef](#)]
- Sadeghi-Tehran, P.; Sabermanesh, K.; Virlet, N.; Hawkesford, M.J. Automated method to determine two critical growth stages of wheat: Heading and flowering. *Front. Plant Sci.* **2017**, *8*, 252. [[CrossRef](#)] [[PubMed](#)]
- Rasti, P.; Demilly, D.; Benoit, L.; Belin, E.; Ducournau, S.; Chapeau-Blondeau, F.; Rousseau, D. Low-cost vision machine for high-throughput automated monitoring of heterotrophic seedling growth on wet paper support. In Proceedings of the British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, 3–6 September 2018; p. 323.

13. Chen, R.; Chu, T.; Landivar, J.A.; Yang, C.; Maeda, M.M. Monitoring cotton (*Gossypium hirsutum* L.) germination using ultrahigh-resolution UAS images. *Precis. Agric.* **2018**, *19*, 161–177. [[CrossRef](#)]
14. Zhao, B.; Zhang, J.; Yang, C.; Zhou, G.; Ding, Y.; Shi, Y.; Zhang, D.; Xie, J.; Liao, Q. Rapeseed seedling stand counting and seeding performance evaluation at two early growth stages based on unmanned aerial vehicle imagery. *Front. Plant Sci.* **2018**, *9*, 1362. [[CrossRef](#)]
15. Jiang, Y.; Li, C.; Paterson, A.H.; Robertson, J.S. DeepSeedling: Deep convolutional network and Kalman filter for plant seedling detection and counting in the field. *Plant Methods* **2019**, *15*, 141. [[CrossRef](#)] [[PubMed](#)]
16. Samiei, S.; Rasti, P.; Vu, J.L.; Buitink, J.; Rousseau, D. Deep learning-based detection of seedling development. *Plant Methods* **2020**, *16*, 103. [[CrossRef](#)]
17. Chéné, Y.; Rousseau, D.; Lucidarme, P.; Bertheloot, J.; Caffier, V.; Morel, P.; Belin, É.; Chapeau-Blondeau, F. On the use of depth camera for 3D phenotyping of entire plants. *Comput. Electron. Agric.* **2012**, *82*, 122–127. [[CrossRef](#)]
18. Nock, C.; Taugourdeau, O.; Delagrangé, S.; Messier, C. Assessing the potential of low-cost 3D cameras for the rapid measurement of plant woody structure. *Sensors* **2013**, *13*, 16216–16233. [[CrossRef](#)] [[PubMed](#)]
19. Paulus, S.; Behmann, J.; Mahlein, A.K.; Plümer, L.; Kuhlmann, H. Low-cost 3D systems: Suitable tools for plant phenotyping. *Sensors* **2014**, *14*, 3001–3018. [[CrossRef](#)] [[PubMed](#)]
20. Rousseau, D.; Chéné, Y.; Belin, E.; Semaan, G.; Trigui, G.; Boudehri, K.; Franconi, F.; Chapeau-Blondeau, F. Multiscale imaging of plants: Current approaches and challenges. *Plant Methods* **2015**, *11*, 6. [[CrossRef](#)]
21. Rosell-Polo, J.R.; Gregorio, E.; Gené, J.; Llorens, J.; Torrent, X.; Arnó, J.; Escola, A. Kinect v2 sensor-based mobile terrestrial laser scanner for agricultural outdoor applications. *IEEE/ASME Trans. Mechatron.* **2017**, *22*, 2420–2427. [[CrossRef](#)]
22. Vit, A.; Shani, G. Comparing rgb-d sensors for close range outdoor agricultural phenotyping. *Sensors* **2018**, *18*, 4413. [[CrossRef](#)] [[PubMed](#)]
23. Perez, R.M.; Cheein, F.A.; Rosell-Polo, J.R. Flexible system of multiple RGB-D sensors for measuring and classifying fruits in agri-food Industry. *Comput. Electron. Agric.* **2017**, *139*, 231–242. [[CrossRef](#)]
24. Martínez-Guanter, J.; Ribeiro, Á.; Peteinatos, G.G.; Pérez-Ruiz, M.; Gerhards, R.; Bengochea-Guevara, J.M.; Machleb, J.; Andújar, D. Low-cost three-dimensional modeling of crop plants. *Sensors* **2019**, *19*, 2883. [[CrossRef](#)]
25. Reynolds, D.; Baret, F.; Welcker, C.; Bostrom, A.; Ball, J.; Cellini, F.; Lorence, A.; Chawade, A.; Khafif, M.; Noshita, K.; et al. What is cost-efficient phenotyping? Optimizing costs for different scenarios. *Plant Sci.* **2019**, *282*, 14–22. [[CrossRef](#)] [[PubMed](#)]
26. Servi, M.; Mussi, E.; Profili, A.; Furferi, R.; Volpe, Y.; Governi, L.; Buonamici, F. Metrological Characterization and Comparison of D415, D455, L515 RealSense Devices in the Close Range. *Sensors* **2021**, *21*, 7770. [[CrossRef](#)] [[PubMed](#)]
27. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [[CrossRef](#)] [[PubMed](#)]
28. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [[CrossRef](#)]
29. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [[CrossRef](#)]
30. Valada, A.; Oliveira, G.L.; Brox, T.; Burgard, W. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In Proceedings of the International Symposium on Experimental Robotics, Nagasaki, Japan, 3–8 October 2016.
31. Andreas, E.; Jost, T.S.; Luciano, S.; Martin, R.; Wolfram, B. Multimodal deep learning for robust RGB-D object recognition. In Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015.
32. Jordi, S.R.; Kai-Lung, H.; Yuan-Sheng, H.; Tekoing, L.; Shintami, C.; Wen-Huang, C. A comparative study of data fusion for RGB-D based visual recognition. *Pattern Recognit. Lett.* **2016**, *73*, 1–6.
33. Wang, A.; Lu, J.; Cai, J.; Cham, T.J.; Wang, G. Large-margin multimodal deep learning for RGB-D object recognition. *IEEE Trans. Multimed.* **2015**, *17*, 1887–1898.
34. Bezen, R.; Edan, Y.; Halachmi, I. Computer vision system for measuring individual cow feed intake using RGB-D camera and deep learning algorithms. *Comput. Electron. Agric.* **2020**, *172*, 105345.
35. Srivastava, N.; Salakhutdinov, R. Learning representations for multimodal data with deep belief nets. In Proceedings of the 29th International Conference Machine Learning (Workshop), Edinburgh, UK, 26 June–1 July 2012.
36. Yu, C.; Shawn, S.; Jianbiao, H.; Degui, X.; Cui, T.; Ping, C.; Henning, M. Medical image retrieval: A multimodal approach. *Cancer Inform.* **2014**, *13*, 125.
37. Lenz, I.; Lee, H.; Saxena, A. Deep learning for detecting robotic grasps. *Int. J. Robot. Res.* **2015**, *34*, 705–724.
38. Ashesh, J.; Avi, S.; Hema, S.K.; Shane, S.; Ashutosh, S. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016.
39. Xinhang, S.; Shuqiang, J.; Luis, H.; Chengpeng, C. Learning effective RGB-D representations for scene recognition. *IEEE Trans. Image Process.* **2019**, *28*, 980–993.

40. Cheng, Y.; Zhao, X.; Cai, R.; Li, Z.; Huang, K.; Rui, Y. Semi-supervised multimodal deep learning for RGB-D object recognition. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), New York, NY, USA, 9–15 July 2016.
41. Li, S.; Cheng, Z.; Rustam, S. Weakly-supervised DCNN for RGB-D object recognition in real-world applications which lack large-scale annotated training data. *arXiv* **2017**, arXiv:1703.06370.
42. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
43. Garboughe, H.; Rasti, P.; Rousseau, D. Deep learning-based detection of seedling development from indoor to outdoor. In Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP), Bratislava, Slovakia, 2–4 June 2021; Volume 1, pp. 1–11.
44. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
45. Minervini, M.; Giuffrida, M.V.; Perata, P.; Tsaftaris, S.A. Phenotiki: An open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *Plant J.* **2017**, *90*, 204–216. [[CrossRef](#)]
46. Intel RealSense Documentation—Intel RealSense Depth Tracking Cameras. Available online: <https://dev.intelrealsense.com/docs/docs-get-started> (accessed on 7 December 2019).
47. Granados, M.; In-Kim, K.; Tompkin, J.; Kautz, J.; Theobalt, C. Background Inpainting for Videos with Dynamic Objects and a Free-moving Camera. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012.
48. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
49. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
50. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of CNN and RNN for natural language processing. *arXiv* **2017**, arXiv:1702.01923.
51. Zhou, K.; Wang, W.; Hu, T.; Deng, K. Time Series Forecasting and Classification Models Based on Recurrent with Attention Mechanism and Generative Adversarial Networks. *Sensors* **2020**, *20*, 7211. [[CrossRef](#)]
52. Yuan, Y.; Lin, L. Self-Supervised Pre-Training of Transformers for Satellite Image Time Series Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 474–487. [[CrossRef](#)]
53. Garnot, V.S.F.; Landrieu, L.; Giordano, S.; Chehata, N. Satellite image time series classification with pixel-set encoders and temporal self-attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12325–12334.
54. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
55. Sampaio, G.S.; Silva, L.A.d.; Marengoni, M. 3D Reconstruction of Non-Rigid Plants and Sensor Data Fusion for Agriculture Phenotyping. *Sensors* **2021**, *21*, 4115. [[CrossRef](#)] [[PubMed](#)]
56. Jin, J.; Dundar, A.; Bates, J.; Farabet, C.; Culurciello, E. Tracking with deep neural networks. In Proceedings of the 2013 47th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 20–22 March 2013; pp. 1–5.
57. Srivastava, D.; Shamim, M.; Kumar, M.; Mishra, A.; Maurya, R.; Sharma, D.; Pandey, P.; Singh, K. Role of circadian rhythm in plant system: An update from development to stress response. *Environ. Exp. Bot.* **2019**, *162*, 256–271. [[CrossRef](#)]

RESEARCH

Open Access



ROSE-X: an annotated data set for evaluation of 3D plant organ segmentation methods

Helin Dutagaci¹, Pejman Rasti^{1,2,3}, Gilles Galopin² and David Rousseau^{1,2*}

Abstract

Background: The production and availability of annotated data sets are indispensable for training and evaluation of automatic phenotyping methods. The need for complete 3D models of real plants with organ-level labeling is even more pronounced due to the advances in 3D vision-based phenotyping techniques and the difficulty of full annotation of the intricate 3D plant structure.

Results: We introduce the ROSE-X data set of 11 annotated 3D models of real rosebush plants acquired through X-ray tomography and presented both in volumetric form and as point clouds. The annotation is performed manually to provide ground truth data in the form of organ labels for the voxels corresponding to the plant shoot. This data set is constructed to serve both as training data for supervised learning methods performing organ-level segmentation and as a benchmark to evaluate their performance. The rosebush models in the data set are of high quality and complex architecture with organs frequently touching each other posing a challenge for the current plant organ segmentation methods. We report leaf/stem segmentation results obtained using four baseline methods. The best performance is achieved by the volumetric approach where local features are trained with a random forest classifier, giving Intersection of Union (IoU) values of 97.93% and 86.23% for leaf and stem classes, respectively.

Conclusion: We provided an annotated 3D data set of 11 rosebush plants for training and evaluation of organ segmentation methods. We also reported leaf/stem segmentation results of baseline methods, which are open to improvement. The data set, together with the baseline results, has the potential of becoming a significant resource for future studies on automatic plant phenotyping.

Keywords: X-ray, Rosebush, Segmentation, Database, Machine learning

Background

Recent agricultural and genetic technologies require high throughput phenotyping systems which can benefit significantly from the automation of inspection and measurement. Automatic plant phenotyping through 3D data has been a recent research topic in computer vision; however, the scarcity of labeled and complete models of real plants is a roadblock for applying recent machine learning techniques that rely on a vast amount of annotated data. Also, benchmarking data sets are indispensable

for proper comparison of current and future phenotyping methods that operate on 3D data such as volumetric models or point clouds.

The production of annotated data sets has become even more important since the recent bloom of deep learning [1], performance of which was shown to be notably boosted by the availability of large annotated data sets [2]. The success of deep learning methods has triggered the interest in data collection and labeling in specific applications of computer vision such as plant imaging [3]. Most of the freely available annotated plant shoot data sets so far have been in the form of collections of 2D images acquired in the visible spectrum from top or side view. Among the available 2D data

*Correspondence: david.rousseau@univ-angers.fr

² INRA, UMR1345 Institut de Recherche en Horticulture et Semences, 42

Georges Morel CS 60057, 49071 Beaucoze, France

Full list of author information is available at the end of the article



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

sets reported in [3] some are provided with annotated ground truth [4, 5], which is very valuable for phenotyping through computer vision and machine learning. In this article, we are interested in providing 3D annotated models of plants.

Among the most related data sets, some provide multiple images of plants that would allow 3D reconstruction; however, they do not include complete 3D plant models [6–9]. Uchiyama et al. [7] provided a data set containing multiple RGB and depth images of Komatsuna plant together with the manually annotated leaf labels. The data set contains calibration images to be used for estimating 3D geometry from the plant images. Cruz et al. [8] constructed a database named “MSU-PID” containing fluorescence, IR, RGB, and top view depth images of Arabidopsis and bean plants. 3D reconstructions of plants are not available in the database. Bernotas et al. [9] provided an annotated Arabidopsis data set with 3D information acquired using the photometric stereo technique. The data set includes 221 manually annotated Arabidopsis rosettes, which are partially reconstructed using only top-down views of the plants, providing 2.5D information rather than full 3D models. Wen et al. [10] introduced a database of the 3D models of plants and organs from different species, cultivars, and multiple growth periods, however, at present, the majority of the models in the data set correspond to isolated organs, such as models of single leaves or fruits, rather than full plants.

Due to the improvement of the sensitivity of the sensors and the democratization of the technology, X-ray Computer Tomography (CT) is now widely used for plant imaging [11]. While X-ray imaging is the most adopted tool to monitor roots in real soil conditions [12], it is also being employed for the characterization of the aerial parts of plants [13–19]. The use of X-ray imaging has focused on the acquisition of very thin parts enhanced with dye [13, 17, 18] or the internal 3D analysis of the aerial part [14–16, 19].

Rosebushes have been studied with computer vision techniques applied on LiDAR and RGB image data [20, 21] to produce global characterization of the shoot and from there estimate its ornamental value. In contrast to these optics-based methods, X-ray CT imaging, although more expensive, provides complete and occlusion-free volumetric information of the 3D geometric structure of the shoot. Such accurate imaging that is able to capture internal structures provides a means to construct full 3D models of real plants. These models can later be used to guide computer vision and pattern recognition techniques that can operate on data acquired with low-cost imaging devices to inspect a large number of plants used in typical phenotyping experiments.

We provide the ROSE-X data set of 11 complete 3D models of real potted rosebush plants with complex architecture acquired through X-ray computed tomography. The rosebushes we captured through X-ray CT imaging have complex architecture and show significantly high amounts of self-occlusion from all viewpoints, i.e., they possess major challenges for optics-based 3D plant reconstruction methods. These models are suitable to be transformed to other data structures, e.g., full or partial point clouds corresponding to the visible surface of the shoot, similar to what would be obtained with optical systems used for 3D reconstruction of plant shoot such as LiDAR or Time-of-Flight (ToF) cameras [22]. This conversion will make it possible to train and evaluate algorithms that operate on point clouds originating from the visible surface. In addition, with the data available for the occluded parts, these models will make it possible to design algorithms that predict complex plant architectural structure from incomplete input.

The 3D voxel space of each rosebush in the data set is fully annotated through labeling each voxel with its corresponding botanical organ class; “organ” referring to the plant units such as leaves, branches, and flowers. Such ground truth data facilitate the detailed description of the architecture and morphology of the plant, and can be used to train automatic phenotyping algorithms aiming to extract both architectural and organ-level traits. Architectural and organ-level trait analysis of 3D data requires an initial stage of classification of points into their respective categories. Current practice is to segment the acquired data of the plant shoot into branches and leaves. In this paper, we focus on leaf-stem segmentation algorithms as one of the phenotyping applications where our data set can serve both as training data and as a benchmark. We chose four representative methods for stem-leaf segmentation: (1) unsupervised classification using local features from point clouds, (2) support vector machine (SVM) classification using local features from point clouds, (3) random forest-based classification of local features from volumetric data, and (4) 3D U-Net applied on volumetric data. The later two were not previously applied to 3D plant organ segmentation problem. We trained and evaluated the methods on the new ROSE-X data set, and provided baseline performance results.

Methods

The ROSE-X data set

We introduce an open repository of complete 3D models of real rosebush plants with ground truth annotations at organ-level. The acquisition was performed through a 3D Siemens X-ray imaging system with a voltage range of 10–450 kV, using a tungsten transmission target and a 280-mA current. For this study, the system was operated

with an 80-kV voltage. The number of projections was 900, and each radiograph was an average of three exposures of 333 ms each to reduce the noise. The acquisition time per plant was 20 min. A total number of 11 rosebush plants with varying architectural complexity were imaged. The output data obtained from each acquisition session is a stack of X-ray images with a pixel spacing of 0.9766 mm and slice spacing of 0.5 mm. The data is represented in a 3D voxel space, where the intensity of each voxel reflected the material properties of the plant shoot at that voxel.

From the raw volume data, the 3D voxels belonging to the rosebushes and their pots were extracted through masking and thresholding. The masks were manually constructed to separate unrelated material coming from the imaging platform, and thresholding was performed to separate the plant voxels from the air. Table 1 gives the number of thresholded voxels, the number of voxels corresponding to the plant shoot, and the number of voxels on the surface of the plant shoot. The pot contains a significant portion of the voxels of the models; the large difference in the number of the voxels between models is due to different sizes of the pots. The plant shoot corresponds to the plant parts above the soil. Most of the voxels of the plant shoot are on the surface since leaves and petals and sepals of the flowers are very thin structures.

After the X-ray intensity values of the voxels corresponding to air and background material are set to zero, the remaining voxels are assigned to one of the following classes: (1) stem, (2) leaf, (3) flower, (4) pot, (5) tag. The background voxels corresponding to air were assigned “zero” values. The stem class includes both the main branches and the petioles since they have similar geometrical structures and are spatially connected. The plant shoot is composed of the stem, leaf, and flower classes.

Table 1 Number of voxels in the models (also the number of points in the corresponding point cloud)

Model ID	# Thresholded voxels	# Plant shoot voxels	# Plant shoot surface voxels
S268650	794,618	312,212	275,954
S268660	588,101	157,029	127,158
S270230	657,195	205,686	175,800
S270240	642,192	169,276	142,474
S270250	818,568	347,013	301,786
S271780	2,091,739	305,534	264,634
S271790	2,072,313	200,346	171,963
S271800	2,011,882	164,108	138,065
S273080	1,153,337	176,155	145,284
S273090	1,909,986	192,755	166,246
S273110	1,254,316	294,528	257,992

Figure 1 displays the thresholded X-ray volume (a), the organ-level labels obtained through annotation (b), the labels corresponding to the plant shoot (c), and the stem and petiole structure (d) of a sample rosebush model from the data set. Table 2 gives the percentages of voxels of organ classes on the plant shoot and the surface of the plant shoot.

The manual annotation was carried out with the help of ilastik (Interactive Learning and Segmentation Toolkit) [23]: Using pixel classification tool of ilastik, on a rosebush model, we manually marked several voxels in regions belonging to each class to train the classifier. Then, we obtained full-volume predictions on all models generated by the trained classifier of ilastik. Through detailed inspection, we manually corrected the labels of all voxels incorrectly labeled by ilastik.

The data set is available online at [24]. We provide the 3D data in the following forms: (1) the raw X-ray image stack, (2) the binary volume mask indicating the voxels of only the shoot of the plant, the tag, and the pot, and the corresponding organ-level labels, (3) the binary volume mask indicating the voxels only on the surface of the plant shoot, and the corresponding organ-level labels, (4) the point cloud composed of the points of the shoot of the plant, the tag, and the pot with colors indicating organ-level labels, (5) the point cloud composed of the points on the surface of the plant shoot with colors indicating organ-level labels. The details of the file formats and label information are explained in the Additional file 1. Through these forms, it is possible to convert the 3D volumetric models to a labeled polygon mesh model and obtain 3D point clouds as viewed from any position around the plant through ray casting.

Baseline methods for leaf-stem segmentation

Vision-based plant phenotyping has been traditionally performed through analysis of 2D color images from which 3D characteristics of the plants (stem length, volume, leaf area, etc.) have been estimated. With the advance of 3D imaging technologies, phenotyping through the 3D capture and reconstruction of plants have gained considerable attention. In Table 3, characteristics of some of the 3D vision-based phenotyping methods that involve a segmentation stage to separate leaves from branches are summarized. 3D data was captured from various species of plants by structured light depth sensors [25, 26], laser scanners [27–31], ToF cameras [32], or from a set of color images through structure from motion [33, 34].

One of the disadvantages of these optics-based acquisition techniques is that they suffer from a high degree of self-occlusion of plants. As the architecture becomes more complex, more parts of the plants become heavily

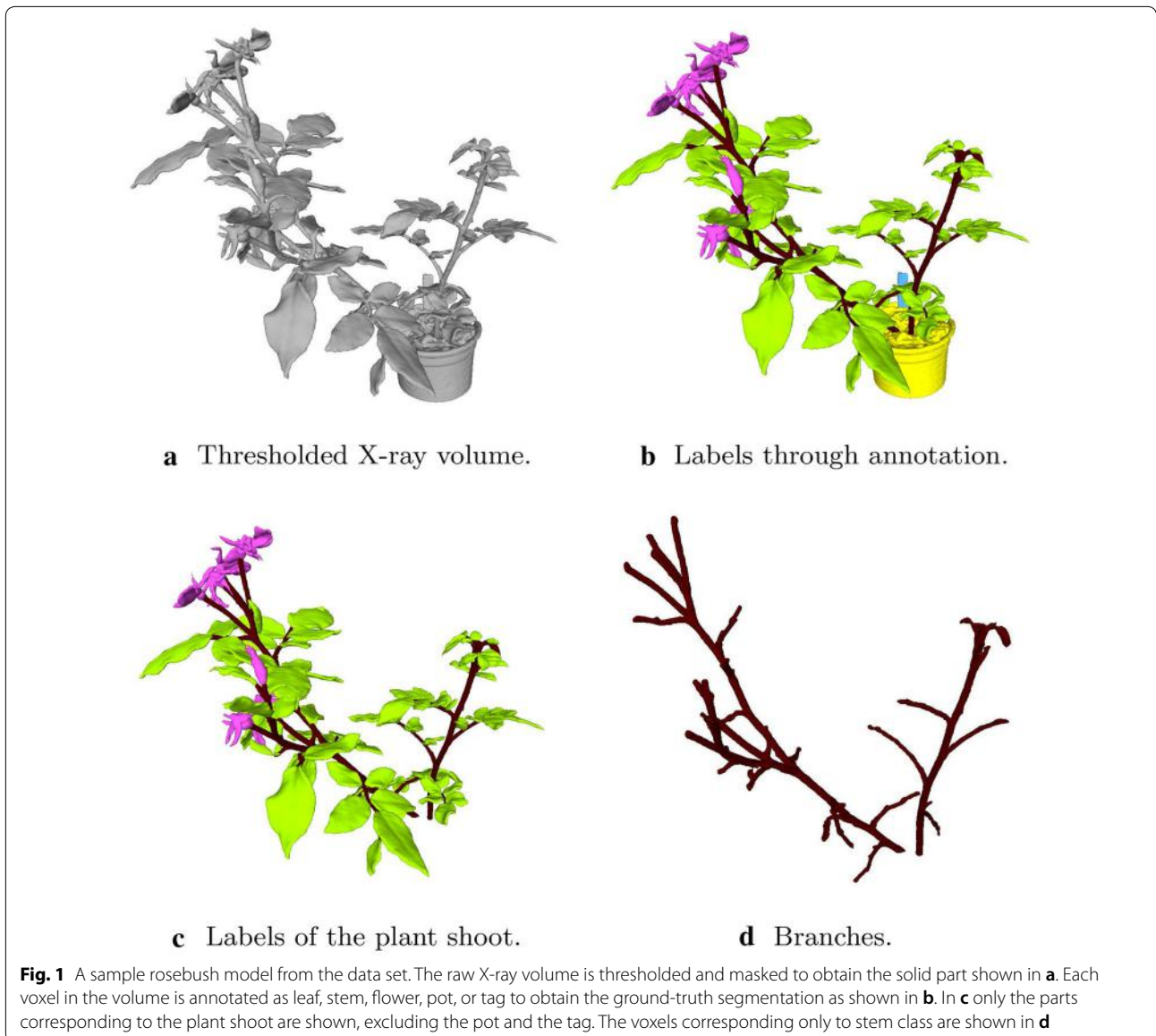


Table 2 Percentages of voxels (points) for organ classes in the plant shoot

Model ID	Leaf	Stem	Flower	Leaf on surface	Stem on surface	Flower on surface
S268650	79.06	13.08	7.86	83.99	9.43	6.58
S268660	70.53	17.06	12.41	77.37	12.66	9.97
S270230	77.07	14.36	8.57	83.44	10.40	6.17
S270240	71.30	16.60	12.10	79.92	11.70	8.38
S270250	75.22	12.33	12.45	80.64	8.93	10.43
S271780	80.97	13.46	5.57	86.35	9.79	3.86
S271790	75.76	13.96	10.28	81.26	10.12	8.62
S271800	73.84	17.09	9.07	81.70	12.57	5.73
S273080	69.20	21.72	9.08	77.50	15.99	6.51
S273090	75.08	19.20	5.72	82.64	13.97	3.39
S273110	79.91	17.07	6.02	83.78	12.27	3.95

Table 3 3D vision based phenotyping methods

	Imaging	Plant type	Application/traits	Segmentation approach
<i>Local surface features on point clouds</i>				
Dey et al. [33]	Structure from motion	Grapevine	Classification of 3D points into leaves, branches, and fruit (red)	Eigenvalues of local covariance matrix, SVM, CRF
Li et al. [25]	Structured light scanner	Anthurium, Dishia, Dancing bean	Leaf/stem segmentation for tracking events in time like budding and bifurcation	Local point features, MRF
Paulus et al. [28]	3D laser scanner	Grapevine, Wheat	Leaf/stem segmentation for grapevine	Local point features (FPFH), SVM, Region growing
Paulus et al. [35]	3D laser scanner	Barley	Leaf/stem segmentation for leaf area and stem height estimation	Local point features (FPFH), SVM, Region growing
Wahabzada et al. [30]	3D laser scanner	Grapevine, Wheat, Barley	Segmentation of leaf, stem, ear, and fruit parts	Local point features (FPFH), clustering, Region growing
Sodhi et al. [26]	Multi-view stereo & Kinect	Sorghum	Leaf/stem segmentation	Local point features (FPFH), SVM, CRF
Elnashef et al. [36]	Multi-view stereo	Corn, Cotton, Wheat	Leaf/stem segmentation	Eigenvalues of the second tensor
<i>Local features on volumetric models</i>				
Klodt et al. [37]	Multi-view stereo	Barley	Leaf/stem segmentation for the estimation of volume and surface area of the plant and the number of leaves	Eigenvalues of the second-moments tensor
Goldbach et al. [38]	Shape-from-silhouette	Tomato seedling	Leaf/stem segmentation for leaf length, width and area estimation	Breath-first flood-fill algorithm with a 26-connected neighbourhood
<i>Spectral clustering</i>				
Hétroy-Wheeler et al. [39]	Laser scanner	Tree seedlings	Segmentation of stems, leaves, and petioles for leaf surface area estimation	Graph construction, spectral embedding and clustering
Santos et al. [40]	Structure from motion	Sunflower, soybean	Leaf/stem segmentation for leaf surface area estimation	Graph construction, spectral embedding and clustering
<i>Geometric primitives</i>				
Binney and Sukhatme [31]	2D laser scanner	Tree branch	Segmentation of leaves and branches for estimation of branch locations, angles, radii, and lengths, and connectivity information between branches	Generative models for branches and branchpoints
Paproki et al. [41]	Multi-view stereo	Cotton	Stem, petiole and leaf segmentation for estimation of geometric properties such as stem height, leaf height and inclination angle	Region growing, tubular shape-fitting, clustering
Chaivatrakul et al. [32]	Time of Flight	Corn	Leaf/stem segmentation for stem diameter, leaf length, area, and angle estimation	Stem extraction by ellipse fitting and linking, and elliptical cylinder extrusion
Gélard et al. [34]	Structure from motion	Sunflower	Stem, petiole and leaf segmentation for leaf area estimation	Ring climbing for extraction of stems and petioles, clustering for segmenting leaves

occluded, making it impossible to capture some regions from any viewpoint. That disadvantage forced most automatic part segmentation and phenotyping research to be conducted on plants with relatively simple architectural and geometrical structure, such as plants with a single stem and well-separated leaves. With X-ray imaging, 3D information of the entire plant material can be captured. However, many phenotyping activities, such as growth monitoring, require the plants not to be moved, which makes X-ray imaging impractical. The bulk of the automatic phenotyping activities is bound to rely on optics-based acquisition devices. Although X-ray imaging will remain as an appropriate tool for applications such as root growth analysis, we envision that the ROSE-X data set will be mainly a resource for algorithms that operate on point clouds acquired with optics-based methods. The availability of complete models of real plants with high architectural complexity and full annotation will serve as a guiding resource for processing occluded point clouds of highly complicated plants acquired by RGB or depth cameras, or laser scanners.

Whether the data is in 3D volumetric form or is in the form of a 3D point cloud, semantic segmentation is required for particular phenotyping objectives, such as organ-level phenotyping, extraction of the architecture and event detection such as leaf growth and decay. Leaf-stem segmentation is the most commonly addressed problem in organ-level phenotyping. We can categorize leaf/stem segmentation methods for 3D phenotyping into the following groups: (1) segmentation using local surface features on point clouds [25, 26, 28, 30, 33, 35], (2) segmentation using local features on volumetric data [37, 38], (3) segmentation through spectral clustering [39, 40, 42], (4) segmentation by fitting geometric primitives [31, 32, 34, 41, 43]. Table 3 is organized using this categorization. In this work, instead of an exhaustive evaluation of all the available methods on our labeled data set, we selected four representative approaches as baseline methods for segmenting the shoot of the rosebush data into its branches and leaves. Two of these methods are based on local features extracted from the point cloud. The other two methods assume volumetric data as input, and have not been previously applied to the plant organ segmentation problem. For all methods, it is assumed that the plant shoot is already separated from the pot. In the following subsections, the baseline methods are described in detail.

Segmentation using local surface features on point clouds

One of the most common approaches to segment point clouds of plants is to use local features. Point neighborhoods on leaves and branches exhibit distinguishing distributions, which can be attributed to their sheet-like

or line-like structures, respectively. One of the simplest approaches is to represent such characteristics by the eigenvalues of the covariance matrix of the neighborhood. Researchers have devised the use of more sophisticated point features such as Fast Point Feature Histograms (FPFH) ([28, 35]) that provide a rich description of the local structure around a point. In this work, we opted to use the simplest point neighborhood descriptors for the baseline methods. For more information on 3D local features, we refer to the book [44] of Laga et al.

For a point x in the point cloud, the neighborhood can be defined as the set $\mathcal{N}_x = \{x_i : \|x - x_i\| < d\}$, where d is the radius of the neighborhood. The covariance matrix of the neighborhood is calculated as $C = \frac{1}{|\mathcal{N}_x|-1} \sum_{x_i \in \mathcal{N}_x} (x_i - \bar{x})(x_i - \bar{x})^T$, where $\bar{x} = \frac{1}{|\mathcal{N}_x|} \sum_{x_i \in \mathcal{N}_x} x_i$ is the mean of the points.

The relative magnitudes of the eigenvalues $\{\lambda_1, \lambda_2, \lambda_3\}$ of the covariance matrix with $\lambda_1 \leq \lambda_2 \leq \lambda_3$ can serve as local descriptors to discriminate leaf and stem points. For a thin flat structure, we expect λ_1 to be much smaller than both λ_2 and λ_3 . We also expect λ_2 and λ_3 to be close to each other. For a line-like structure we have a predominantly large value of λ_3 , with λ_1 and λ_2 being much smaller.

We used the eigenvalues of the local covariance matrix in two baseline stem/leaf segmentation methods. The first is an unsupervised method based on the Markov Random Fields (MRF) formulation given in [25]. The second is a supervised method where a classifier is trained with local features derived from the eigenvalues. This second approach aligns with the methods proposed in [26, 33].

Local features on point clouds—unsupervised (LFPC-u) : For this baseline method, we followed a simplified version of the stem/leaf classification method given in [25]. The eigenvalues are used to define local surface features on the point clouds and to search for a mapping f_B from a point x to one of the two labels for leaf (L) and stem (S) categories. The point cloud can be organized in a graph where the points $x \in \mathcal{X}$ correspond to the nodes and pairs of locally connected points $(x_i, x_j) \in \mathcal{E}$ constitute the edges. In our implementation, a pair (x_i, x_j) was considered to be an edge if the Euclidean distance between them is less than 1.4mm. The energy associated with a particular label mapping is defined as

$$E(f_B) = w_D \sum_{x \in \mathcal{X}} D_x(f_B(x)) + w_V \sum_{(x_i, x_j) \in \mathcal{E}} V(f_B(x_i), f_B(x_j)). \quad (1)$$

The weight factors w_D and w_V determine a compromise between the class likelihoods of individual nodes and the coherence across the edges. $D_x(f_B(x))$ corresponds to the data term (the unary potential) which gives the cost of

classifying a point x into a leaf or stem category. The term $V(f_B(x_i), f_B(x_j))$ gives the smoothness term (the pairwise potential) and is used to encourage labeling coherence between neighboring points. The energy function is minimized through min-cut algorithm [45] to obtain the optimum labels for the point cloud.

To determine the data and smoothness terms, an estimate of the curvature at point x is computed using the eigenvalues of the covariance matrix as $C(x) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}$. The range of the curvature values is $[0, 1/3]$, and leaf points are expected to have lower curvature values than stem points. A flatness feature is defined as $R(x) = \log(\max(C(x), c_\epsilon))$, where c_ϵ is set to 0.015. $R(x)$ is in the range $[R_L, R_S]$ with $R_L = \log(c_\epsilon)$ and $R_S = \log(1/3)$. Then, the data term is calculated as

$$D_x(f_B(x)) = \begin{cases} R(x) - R_L, & \text{if } f_B(x) = L. \\ R_S - R(x), & \text{if } f_B(x) = S. \end{cases} \quad (2)$$

The smoothness term also depends on the curvature $C(x)$, which is used as a measure of the discontinuity of the surface. The pairwise potential is set to be inversely proportional to the curvature since a high curvature value indicates a discontinuity which can be considered as the boundary of a plant part. The smoothness term is defined as

$$V(f_B(x_i), f_B(x_j)) = \begin{cases} \max\left(\frac{1}{C(x_i)}, \frac{1}{C(x_j)}\right), & \text{if } f_B(x_i) \neq f_B(x_j). \\ 0, & \text{if } f_B(x_i) = f_B(x_j). \end{cases} \quad (3)$$

Notice that this method is an unsupervised method in the sense that it does not require labeled training data to transform or organize features to boost their discriminating power. However, the weight factors w_D and w_V in Eq. (1) need to be fixed. Through experimentation on one rosebush reserved to train the methods, we found that $w_D = 0.9$ and $w_N = 0.1$ yielded the best results.

Local features on point clouds—supervised (LFPC-s): For the second baseline method, we selected to derive local features from the eigenvalues of the local covariance matrix, and used SVM as the classifier as in the work of Dey et al. ([33]). We defined the local features as follows:

$$F_1 = \frac{\lambda_1}{\sqrt{\lambda_2 \lambda_3}} \quad F_2 = \frac{\lambda_2}{\lambda_3} \quad F_3 = \frac{\lambda_1}{\sqrt{\lambda_1 \lambda_2 \lambda_3}} \quad F_4 = \frac{\lambda_1}{\lambda_2} \quad (4)$$

The size of the neighborhood from which the eigenvalues are computed determines the scale at which the local structures will be analyzed. The stem and the petioles of the plant shoot have varying widths, likewise the leaves exhibit a large size variability. Instead of fixing the radius, we extracted the features $\{F_1, F_2, F_3, F_4\}$ at various scales and concatenate them into a single feature vector. In our

tests, we used six scales, corresponding to neighborhoods of radii 2, 3, 4, 5, 6, and 7 mm. Using one of the rosebush models with ground truth labels, we trained a two-class linear SVM classifier.

Segmentation using local features on volumetric data (LFVD)

The point cloud data acquired from optic-based sensors such as RGB cameras or laser scanners can be converted to binary volumetric data using a 3D occupancy grid. The regular structure of 3D volume allows to apply standard filtering and feature extraction tools such as smoothing and estimation of first and second order derivatives. The software ilastik [23] can extract various types of features from 3D volume data: the color features correspond to the raw intensity values smoothed by a Gaussian filter. The edge features are the eigenvalues of the structure tensor, eigenvalues of the Hessian matrix, the gradient magnitude of the difference of Gaussians and Laplacian of Gaussian. The texture features correspond to eigenvalues of the structure tensor, eigenvalues of the Hessian matrix, and orientation features are the raw structure tensor and Hessian matrix entries. In our tests, the mentioned features are extracted from data smoothed by Gaussian filters with scales 0.7, 1.0, 1.6, 3.5, 5.0, and 10.0 mm.

The voxels of the original X-ray data possess intensity values which are determined by the intensity of the X-rays passing through the voxels and the material properties. X-ray intensity values in our models depend on the material properties of plant parts; e.g., leaves have very low intensity values compared to branches. In order to have comparable results between the volume-based and surface-based baseline methods, we used the binary volume mask, indicating the voxels of only the shoot of the plant. We further set the values of the voxels which are not on the surface of the plant-shoot, i.e., interior voxels, to zero, so that only the voxels on the surface of the plant-shoot will remain.

We employed ilastik [23] to extract intensity, edge, and texture features from one binary plant model and to train a random forest classifier [46] using the ground-truth labels. Once the classifier is trained on one model; it is tested on all the other models on the data set.

CNN on volume data (3D U-Net)

As a representative of deep learning methods, we selected 3D U-Net [47], which is originally proposed to provide dense volumetric segmentation maps for biomedical images. It is an extension of the 2D U-net architecture developed by Ronneberger et al. [48]; all the 2D operations in the 2D u-net are replaced with their 3D counterparts. The input volume is first passed through an analysis path with four resolution layers, each of which is composed of two $3 \times 3 \times 3$ convolutions with

Rectified Linear Units (ReLU) and one $3 \times 3 \times 3$ max pooling operation. Max pooling corresponds to down-sampling by using the maximum value from each of a cluster of neurons at the prior layer. Then a synthesis path is applied with four resolution layers each consisting of one $2 \times 2 \times 2$ upconvolution operator followed by two $2 \times 2 \times 2$ convolutions with ReLU. The high-resolution features obtained at the analysis path are also provided to the synthesis path through shortcut connections between layers of equal resolution. The size of the input voxel grid to the network is $144 \times 144 \times 144$, and the output is a volumetric data of the same size giving the label of each voxel. The architecture graph can be found in [47]. For more information on deep learning and the definitions of the classical layers that constitute the basis of deep neural networks, we refer to the book [49] of Goodfellow et al.

As we did with the baseline method based on local volumetric features, we only used the thresholded voxels on the surface of the shoot, so the input is binary devoid of the intensity information. We used one rosebush model to train the network. We extracted 25 subvolumes of size $144 \times 144 \times 144$ from various locations of the full volume of the model such that each subvolume contained leaf and stem instances. 20 of the subvolumes were used for training and 5 of them were used for validation. For a test model, we regularly partitioned the volume to non-overlapping subvolumes and provided the subvolumes to the network as inputs to get the corresponding segmentation.

Results

In this paper, we concentrated on the problem of partitioning the plant models into its leaf and stem (branch) parts; so the training and evaluation of the baseline methods are performed using the ground truth labels corresponding to the leaves and stems only. In our evaluation, we ignored the predictions generated on the flower parts.

There are many metrics for segmentation evaluation, such as Matthews Correlation Coefficient [50], Cohen's κ coefficient [51], Dice Similarity Coefficient [52], all with their advantages and all applicable in the framework of our benchmark. In this paper, we used precision (also known as Positive Predictive Value), recall (also known as sensitivity), and Intersection over Union

(IoU) to evaluate the baseline methods. Recall for the leaf class (R_{leaf}) is the ratio of the number of correctly labeled leaves (true positives) to the total number of leaf points in the ground truth (true positives + false negatives). Precision for the leaf class (P_{leaf}) is the ratio of the number of correctly labeled leaves (true positives) to the total number of points classified as leaf points by the algorithm (true positives + false positives). Recall (R_{stem}) and precision (P_{stem}) for the stem class are defined in the same way. Intersection over Union (IoU) metric for each class (IoU_{leaf} and IoU_{stem}) is defined as the ratio of all the true positives to the sum of true positives, false negatives and false positives.

For a single fold of the experimental evaluation, we selected one rosebush model for training and tested the algorithms on the remaining 10 models. For the unsupervised method based on local features on point clouds, we used the training model to optimize the weights of the data and smoothness terms. The results were averaged over the test models and over 5-fold experiments. A different rosebush model is reserved as training data for each fold. Table 4 gives the performances of the baseline leaf/stem segmentation methods. The visual results for a sample test rosebush are given in Fig. 2. The predicted labels of the rosebush model are displayed as a volume or as a point cloud depending on the type of the data the corresponding method processes. Figure 3 gives the stem points predicted by each baseline method. Correct predictions of the stem points with their connectivity maintained are especially important for establishing the architectural structure of the plant.

We can observe from Table 4 and Fig. 2 that the voxel classification method through local features (LFVD) gives the best overall performance for leaf/stem classification. It is a supervised method combining multi-scale volumetric local features with the random forest classifier. For this particular data set, it can model well the scale variations of leaf and stem points as well as their geometrical variations due to their locations on the organ (in the middle or at the border). The recall rate for the stem class is around 90%, meaning that 10% of the points on the branches are missed. Most missed stem points are on the petioles, which are in between close leaflets and possess an almost planar structure (Fig. 4c). The

Table 4 Performances of the baseline leaf/stem segmentation methods (%)

Method	R_{leaf}	R_{stem}	P_{leaf}	P_{stem}	IoU_{leaf}	IoU_{stem}
LFPC-u	95.74 ± 1.74	88.03 ± 1.82	98.23 ± 0.33	75.01 ± 9.76	94.10 ± 1.54	67.96 ± 8.18
LFPC-s	97.79 ± 0.46	80.50 ± 1.29	97.19 ± 0.48	83.67 ± 4.88	95.10 ± 0.46	69.57 ± 3.87
LFVD	99.38 ± 0.13	90.01 ± 1.17	98.53 ± 0.43	95.38 ± 1.02	97.93 ± 0.47	86.23 ± 0.31
3D U-Net	81.06 ± 1.68	97.41 ± 1.43	99.63 ± 0.29	54.0 ± 5.77	81.72 ± 1.71	53.58 ± 5.54

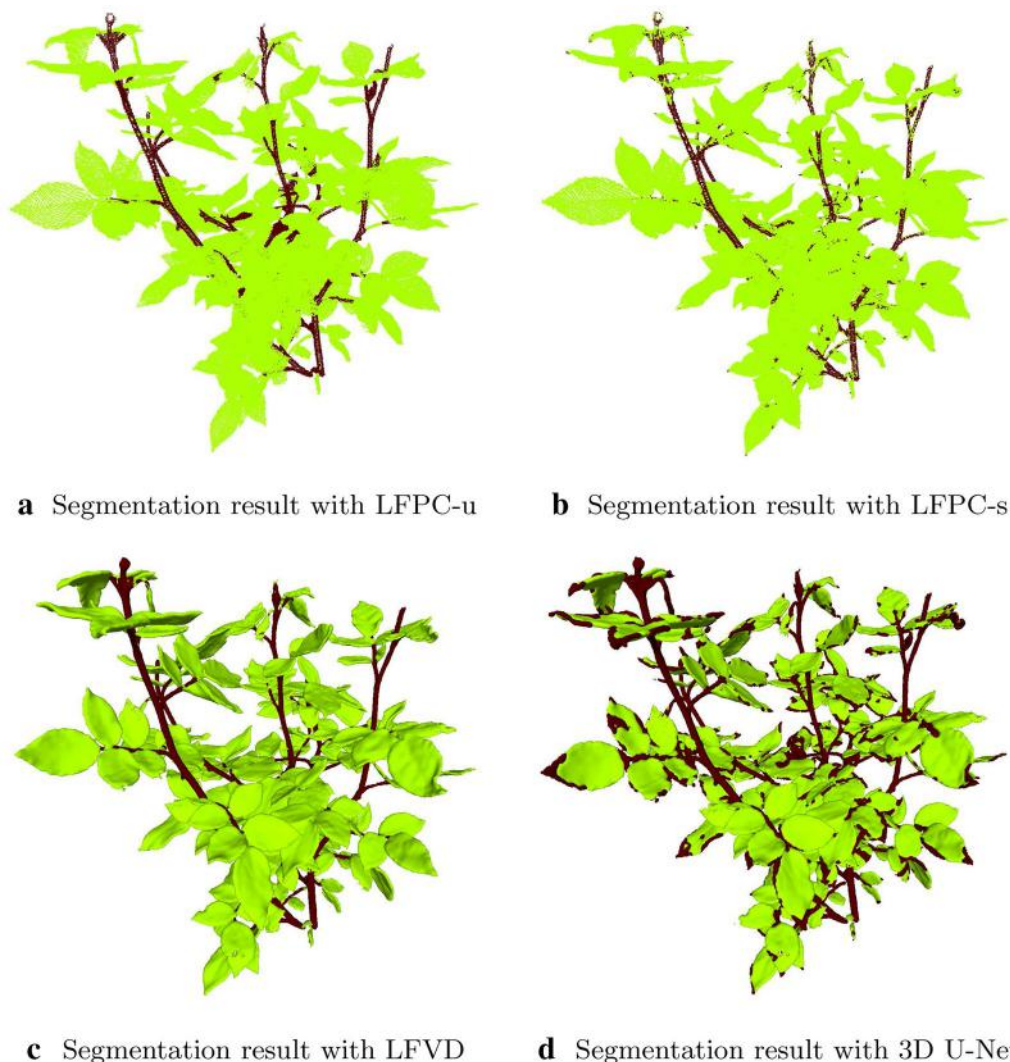


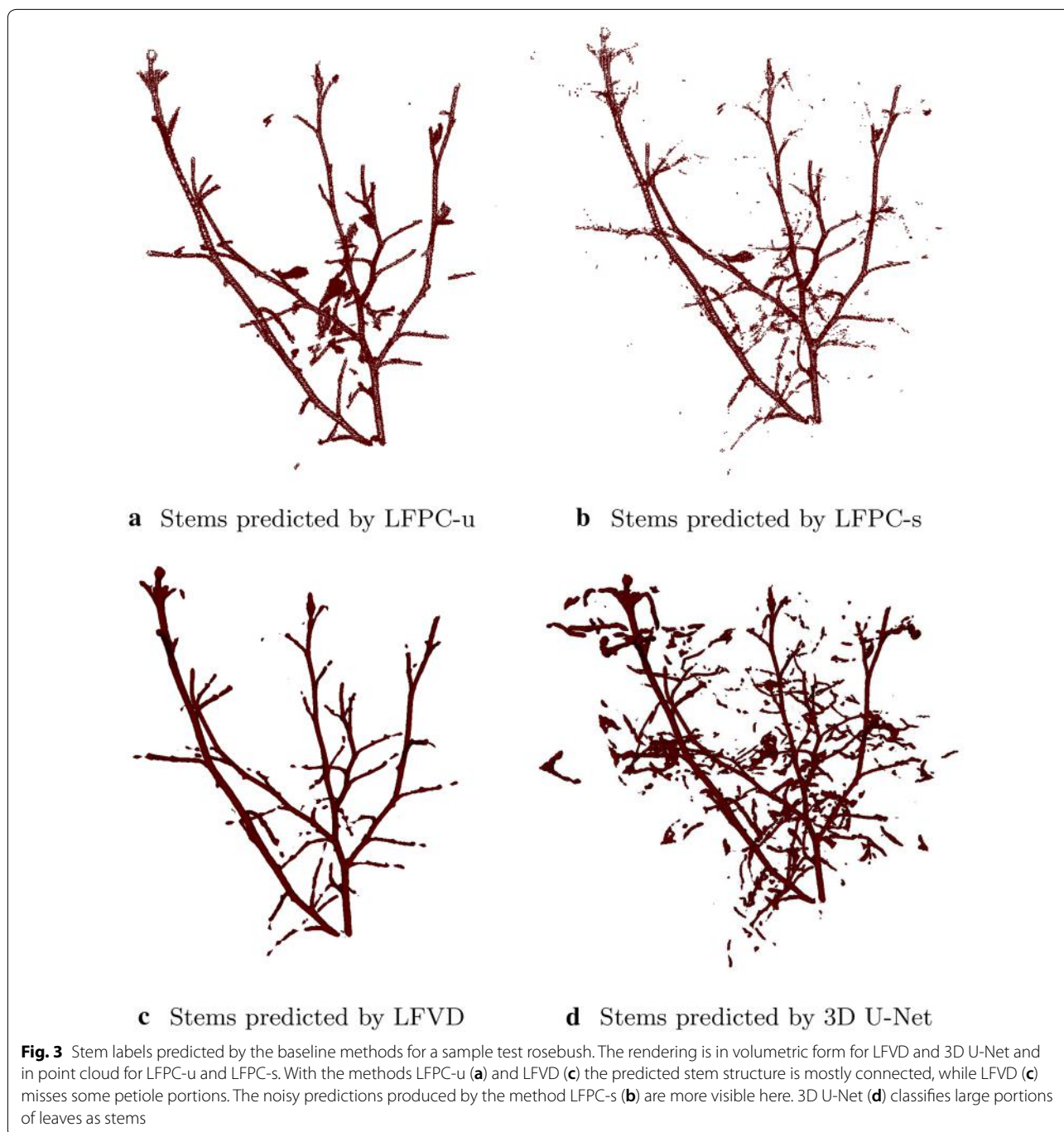
Fig. 2 Leaf and stem labels predicted by the baseline methods for a sample test rosebush. The rendering is in volumetric form for LFVD and 3D U-Net and in point cloud for LFPC-u and LFPC-s. The methods LFPC-u (a) and LFVD (c) produced smooth results, while the labels predicted by LFPC-s (b) are slightly noisy. 3D U-Net (d) wrongly classifies leaf borders as stems

discontinuities in the stem-branch structure predicted by LFVD (Fig. 3c) generally correspond to the petiole portions just in between opposite leaflets.

The classification results obtained by LFPC-u are smooth (Fig. 2a) and the stem structure is mostly connected (Fig. 3a) due to the regularization imposed by the MRF formulation. However, smoothing labels of adjacent points in regions of low curvature leads to an entire leaf or a portion of it to be classified as stem if there is a smooth transition of normals at the boundary as seen in Fig. 4a. This propagation of labels through boundaries with low curvature causes a relatively low stem precision rate (Table 4). Likewise, smooth petiole and leaf boundaries lead to the classification of petiole points as

leaves affecting the stem recall rate negatively. Although this method is unsupervised in the sense that it does not involve a classifier that learns feature transformations through labeled training data, the weights of the data and smoothness terms in Eq. 1 should be optimized for different plant species.

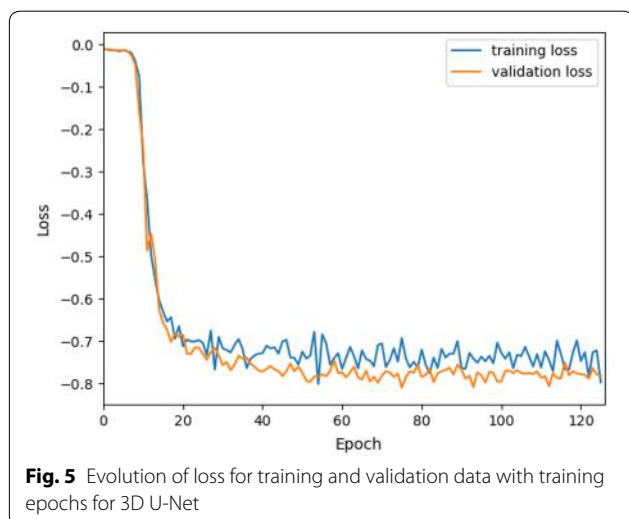
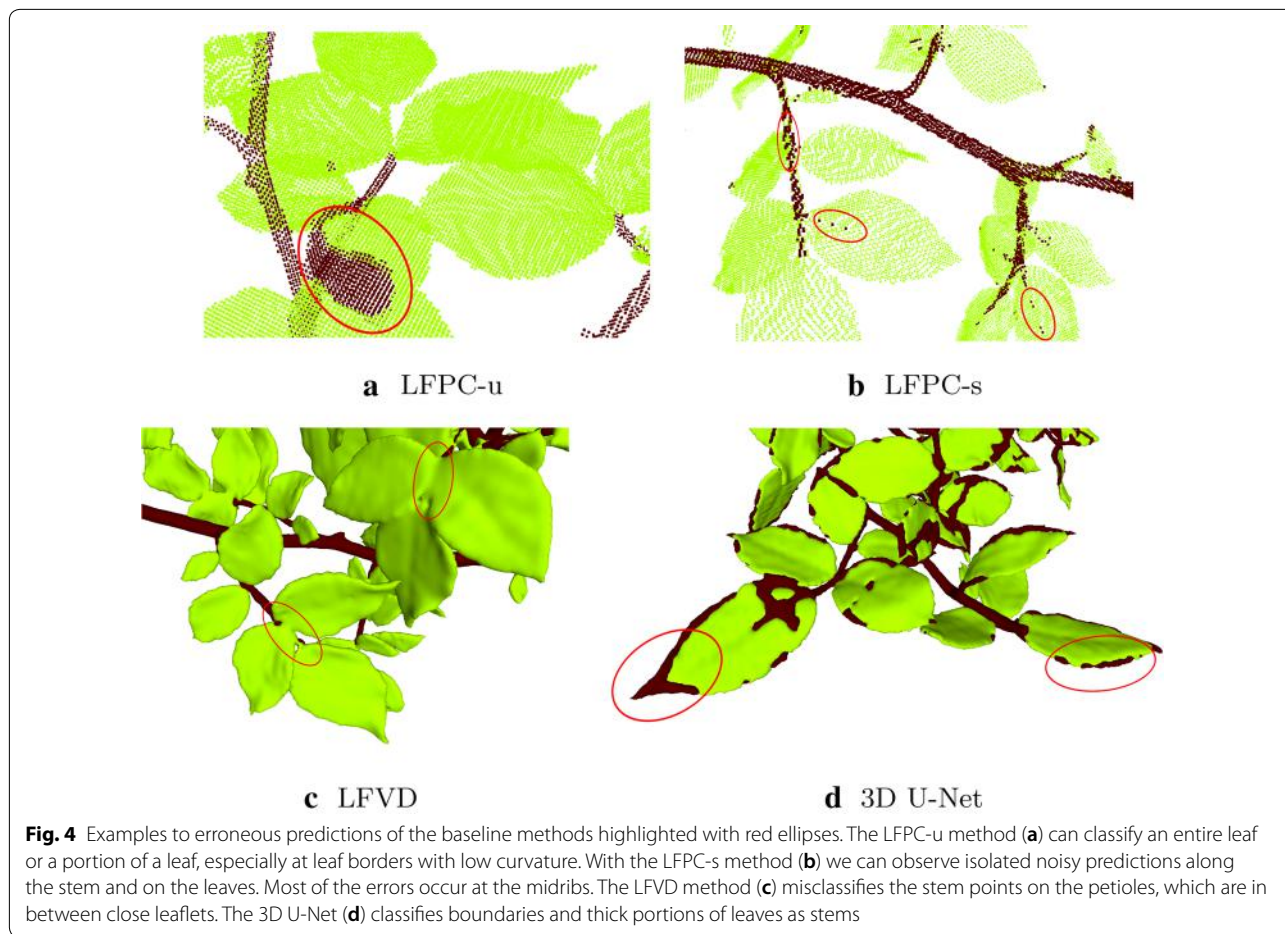
The performance of LFPC-s is slightly higher than that of LFPC-u in terms of the IoU metric (Table 4). Notice that we did not incorporate the MRF formulation for the baseline method LFPC-s, although it is completely applicable through setting the data term using SVM scores. Since no smoothness constraint is imposed on the labels, we can observe isolated noisy predictions along the stem and on the leaves (Fig. 2b). The predicted stem structure



has unconnected small regions due to some leaf points classified as stems (Fig. 4b). Most of these errors occur at the midribs which are usually the thickest parts of the leaves.

3D U-net gives the lowest performance as compared to the other methods. Boundaries and thick portions of leaves are classified as stems as can be observed from Fig. 4d. We give in Fig. 5 the evolution of the training and

validation loss. The dice coefficient function is used as the loss function in 3D U-Net algorithm, which shows a value in a range of 0 to 1. In this case, a negative is multiplied to values for optimization purposes. The curves in this figure show that the model can converge fast after about 50 epochs with the minimum overfitting between training and validation. However, the CNN network did not model the variations of leaves since we used sub-volumes



from a single rosebush model for training to have a fair comparison with other baseline supervised methods. The 3D U-net has far more parameters to learn than the other

methods; therefore, more training data is required for it to be properly trained. Besides, we directly applied the original 3D U-net architecture [47], which was designed for bio-medical data, without modification. In order to improve the results with deep learning, one can either increase the training data by using more than one rosebush model, employ data augmentation strategies, alter the 3D U-Net architecture or propose a new architecture suitable for 3D segmentation of plants. However, detailed analysis of the modifications on these lines is beyond the main objective of this work. We leave the design of 3D CNN architectures specific to plant organ segmentation as an open research problem, to the solution of which our entire labeled data set can contribute.

The methods LFPC-u, LFPC-s, and LFVD were run on a computer with an Intel processor of 3.5 GHz and 128 GB RAM. LFPC-u and LFPC-s were coded with MATLAB, while LFVD was implemented with Python. The average processing time for segmentation of a single model with LFPC-u is 4.2 min. The training time of the SVM classifier for LFPC-s is 5.1 min on average. The segmentation time for a test model with LFPC-s is 1.6 min.

The training time of the Random Forest classifier for LFVD is 13.4 min, and the testing time is 3.3 min. The 3D U-Net was trained using Python on a computer with an Intel processor of 2.2GHz and 8 GPUs of 64 GB. The training time is 3 h, while segmentation time for a new test model is 4.3 min on average.

Discussion

The ROSE-X data set includes high resolution 3D models of real rosebush plants, each of which was annotated at the voxel level with the corresponding botanical organ class. In this article, we focused on the step of segmentation of leaves and stems of automatic phenotyping pipelines. We provided a benchmark for proper comparison of current and future approaches for leaf/stem segmentation.

In this article, the focus has been on leaf segmentation from the stem. This is the first essential step in analyzing the shape and the architecture of the plant. Other questions can be addressed with the ROSE-X data set including issues raised by breeders, producers or consumers such as the study of interactions between genotype and environment on the one hand and phenotype and visual perception on the other. Such issues require the analysis of the growth and morphogenesis of the plant through effective phenotyping. With this objective in mind, it is possible to consider petiole segmentation, the distinction between leaflet and leaves, the detection of meristem along the stem, the analysis of the different part of the flower and the stage of development.

Also, the extraction and encoding of the architectural structure of the plant in the form of an organized collection of the main stem, second and higher order branches, and the branching locations is an important phenotyping task. Another task would be to extract geometrical characteristics of the individual architectural components and their spatial relationships, such as the length and width of the branch segments, petioles and their branching angles, leaf length, width, and area, together with the leaf inclination angles. These advanced botanical traits would be accessible with the spatial resolution of the 3D images of the proposed data set ROSE-X.

In order to evaluate the accuracy of phenotyping methods that aim to extract such more advanced botanical traits, we will release a forthcoming extension of the data set, with extended ground truth data in the form of geometrical properties of individual organs such as leaves, leaflets, petioles, stem segments, branching locations, and the spatial relationship between them.

We present the rosebush models in volumetric form, however, our main concern is to provide labeled data of plants with complex architecture for phenotyping methods that use the visible surface points of the plants as

input. The conversion of the volumetric form to a point cloud via sampling or via ray casting from an arbitrary viewpoint is straightforward. As part of the future work, we will generate partial point clouds from the models as seen from around the plant, and apply phenotyping methods that rely on partial data.

Another important issue is the applicability of leaf/stem classification methods trained with the rosebush data set to other plant species. Future work will involve the expansion of the data set with 3D models of different species, and the adaptation of the classifiers learned from one species to others.

Conclusion

This paper introduces a data set composed of 11 complete 3D models acquired through X-ray scanning of real rosebush plants. The models are stored in a voxel grid structure. We also provide the ground truth data, where each voxel stores the corresponding organ class label. The plant models are free from self-occlusion, however they possess complex architectural structure. As a sample application where the data set can be of use, we chose leaf-stem segmentation and compared the classification performances of four baseline methods. We observed that the volumetric approach (LFVD), where a random forest classifier is trained with local features, yielded the best performance. However, other baseline methods tested in this work are also open to further improvement, and there are yet the state-of-the-art techniques (Table 3) to be evaluated on our dataset. The data set is suitable to be annotated with more advanced traits and can be used as a benchmark for evaluation of automatic phenotyping methods that go beyond classifying plant points as leaves and stems.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13007-020-00573-w>.

Additional file 1. Description of the provided annotated dataset.

Acknowledgements

Auhtors thank Paul Salat for contribution in fine curation of the data set.

Authors' contributions

HD and DR conceived and designed this study. GG and DR carried out the acquisition of X-ray images. HD performed the implementation and analyzed the result of the image segmentation methods. PR contributed in the process of the data via the 3D U-Net. HD and DR wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

Helin Dutagaci and Pejman Rasti gratefully acknowledge Région des Pays de la Loire and Angers Loire Métropole for the funding of their Post-Doc positions.

Availability of data and materials

Entire data set will be released after acceptance on a website of University of Angers. Only a single sample is made available in the submitted manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable for that section.

Consent for publication

Not applicable for that section.

Author details

¹ LARIS, UMR INRA IRHS, Université d'Angers, 62 Avenue Notre Dame du Lac, 49000 Angers, France. ² INRA, UMR1345 Institut de Recherche en Horticulture et Semences, 42 Georges Morel CS 60057, 49071 Beaucoze, France. ³ ESAIP, école d'ingénieur informatique et environnement, Saint Barthélemy d'Anjou, France.

Received: 2 August 2019 Accepted: 21 February 2020

Published online: 04 March 2020

References

- Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
- Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009, p. 248–55. <https://doi.org/10.1109/CVPR.2009.5206848>.
- Lobet G, Draye X, Périlleux C. An online database for plant image analysis software tools. *Plant Methods*. 2013;9(1):38. <https://doi.org/10.1186/1746-4811-9-38>.
- Chitwood DH, Otoni WC. Morphometric analysis of Passiflora leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. *GigaScience*. 2017;. <https://doi.org/10.1093/gigascience/giw008>.
- Minervini M, Fischbach A, Scharf H, Tsafaris SA. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognit Lett*. 2016;81:80–9. <https://doi.org/10.1016/j.patrec.2015.10.013>.
- Veley KM, Berry JC, Fentress SJ, Schachtman DP, Baxter I, Bart R. High-throughput profiling and analysis of plant responses over time to abiotic stress. *bioRxiv*. 2017;. <https://doi.org/10.1101/132787>.
- Uchiyama H, Sakurai S, Mishima M, Arita D, Okayasu T, Shimada A, Taniguchi R. An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset. In: 2017 IEEE international conference on computer vision workshops (ICCVW); 2017, p. 2038–45. <https://doi.org/10.1109/ICCVW.2017.239>.
- Cruz JA, Yin X, Liu X, Imran SM, Morris DD, Kramer DM, Chen J. Multi-modality imagery database for plant phenotyping. *Mach Vis Appl*. 2016;27(5):735–49. <https://doi.org/10.1007/s00138-015-0734-6>.
- Bernotas G, Scorza LCT, Hansen MF, Hales IJ, Halliday KJ, Smith LN, Smith ML, McCormick AJ. A photometric stereo-based 3D imaging system using computer vision and deep learning for tracking plant growth. *GigaScience* 2019;8(5). <https://doi.org/10.1093/gigascience/giz056.giz056>. <http://oup.prod.sis.lan/gigascience/article-pdf/8/5/giz056/28704193/giz056.pdf>.
- Wen W, Guo X, Wang Y, Zhao C, Liao W. Constructing a three-dimensional resource database of plants using measured in situ morphological data. *Appl Eng Agric*. 2017;33(6):747–56. <https://doi.org/10.13031/aea.12135>.
- Perez-Sanz F, Navarro PJ, Egea-Cortines M. Plant phenomics: an overview of image acquisition technologies and image data analysis algorithms. *GigaScience*. 2017;. <https://doi.org/10.1093/gigascience/gix092>.
- Atkinson JA, Pound MP, Bennett MJ, Wells DM. Uncovering the hidden half of plants using new advances in root phenotyping. *Curr Opin Biotechnol*. 2019;55:1–8. <https://doi.org/10.1016/j.copbio.2018.06.002>.
- Staedler YM, Masson D, Schönenberger J. Plant tissues in 3D via x-ray tomography: simple contrasting methods allow high resolution imaging. *PLoS ONE*. 2013;8(9):1–10. <https://doi.org/10.1371/journal.pone.0075295>.
- Hughes N, Askew K, Scotson CP, Williams K, Sauze C, Corke F, Doonan JH, Nibau C. Non-destructive, high-content analysis of wheat grain traits using X-ray micro computed tomography. *Plant Methods*. 2017;13(1):76. <https://doi.org/10.1186/s13007-017-0229-8>.
- Gomez FE, Carvalho G, Shi F, Muliana AH, Rooney WL. High throughput phenotyping of morpho-anatomical stem properties using X-ray computed tomography in sorghum. *Plant Methods*. 2018;14(1):59. <https://doi.org/10.1186/s13007-018-0326-3>.
- Du J, Zhang Y, Guo X, Ma L, Shao M, Pan X, Zhao C. Micron-scale phenotyping quantification and three-dimensional microstructure reconstruction of vascular bundles within maize stalks based on micro-CT scanning. *Funct Plant Biol*. 2017;44(1):10–22. <https://doi.org/10.1071/FP16117>.
- Schneider JV, Rabenstein R, Wesenberg J, Wesche K, Zizka G, Habersetter J. Improved non-destructive 2D and 3D X-ray imaging of leaf venation. *Plant Methods*. 2018;14(1):7. <https://doi.org/10.1186/s13007-018-0274-y>.
- Wang Z, Verboven P, Nicolai B. Contrast-enhanced 3D micro-CT of plant tissues using different impregnation techniques. *Plant Methods*. 2017;13(1):105. <https://doi.org/10.1186/s13007-017-0256-5>.
- Mathers AW, Hepworth C, Baillie AL, Sloan J, Jones H, Lundgren M, Fleming AJ, Mooney SJ, Sturrock CJ. Investigating the microstructure of plant leaves in 3D with lab-based X-ray computed tomography. *Plant Methods*. 2018;14(1):99. <https://doi.org/10.1186/s13007-018-0367-7>.
- Garbez M, Chéné Y, Belin É, Sigogne M, Labatte J-M, Hunault G, Symoneaux R, Rousseau D, Galopin G. Predicting sensorial attribute scores of ornamental plants assessed in 3D through rotation on video by image analysis: a study on the morphology of virtual rose bushes. *Comput Electron Agric*. 2016;121:331–46. <https://doi.org/10.1016/j.compag.2016.01.001>.
- Chéné Y, Rousseau D, Belin É, Garbez M, Galopin G, Chapeau-Blondeau F. Shape descriptors to characterize the shoot of entire plant from multiple side views of a motorized depth sensor. *Mach Vis Appl*. 2016;27(4):447–61. <https://doi.org/10.1007/s00138-016-0762-x>.
- Vázquez-Arellano M, Griepentrog H, Reiser D, Paraforos D. 3-D imaging systems for agricultural applications—a review. *Sensors*. 2016;. <https://doi.org/10.3390/s16050618>.
- Sommer C, Strähle C, Köthe U, Hamprecht FA. Ilastik: Interactive learning and segmentation toolkit. In: Eighth IEEE international symposium on biomedical imaging (ISBI 2011). Proceedings; 2011, p. 230–3. <https://doi.org/10.1109/ISBI.2011.5872394>.
- The ROSE-X Dataset. <https://uabox.univ-angers.fr/index.php/s/rnPm5EHFK6Xym9t>.
- Li Y, Fan X, Mitra NJ, Chamovitz D, Cohen-Or D, Chen B. Analyzing growing plants from 4D point cloud data. *ACM Trans Graph*. 2013;32(6):157. <https://doi.org/10.1145/2508363.2508368>.
- Sodhi P, Vijayarangan S, Wettergreen D. In-field segmentation and identification of plant structures using 3D imaging. In: 2017 IEEE/RISJ international conference on intelligent robots and systems (IROS); 2017, p. 5180–7. <https://doi.org/10.1109/IROS.2017.8206407>.
- Paulus S, Behmann J, Mahlein A-K, Plümer L, Kuhlmann H. Low-cost 3D systems: suitable tools for plant phenotyping. *Sensors*. 2014;14(2):3001–18. <https://doi.org/10.3390/s140203001>.
- Paulus S, Dupuis J, Mahlein A-K, Kuhlmann H. Surface feature based classification of plant organs from 3D laserscanned point clouds for plant phenotyping. *BMC Bioinformatics*. 2013;14(1):238. <https://doi.org/10.1186/1471-2105-14-238>.
- Chaudhury A, Brophy M, Barron JL. Junction-based correspondence estimation of plant point cloud data using subgraph matching. *IEEE Geosci Remote Sens Lett*. 2016;13(8):1119–23. <https://doi.org/10.1109/LGRS.2016.2571121>.
- Wahabzada M, Paulus S, Kersting K, Mahlein A-K. Automated interpretation of 3D laserscanned point clouds for plant organ segmentation. *BMC Bioinformatics*. 2015;16(1):248. <https://doi.org/10.1186/s12859-015-0665-2>.
- Binney J, Sukhatme GS. 3D tree reconstruction from laser range data. In: 2009 IEEE international conference on robotics and automation; 2009, p. 1321–6. <https://doi.org/10.1109/ROBOT.2009.5152684>.
- Chaivivatrakul S, Tang L, Dailey MN, Nakarmi AD. Automatic morphological trait characterization for corn plants via 3D holographic reconstruction. *Comput Electron Agric*. 2014;109:109–23. <https://doi.org/10.1016/j.compag.2014.09.005>.

33. Dey D, Mummert L, Sukthankar R. Classification of plant structures from uncalibrated image sequences. In: 2012 IEEE workshop on the applications of computer vision (WACV); 2012, p. 329–36. <https://doi.org/10.1109/WACV.2012.6163017>.
34. Gélard W, Devy M, Herbulot A, Burger, P. Model-based segmentation of 3D point clouds for phenotyping sunflower plants. In: Proceedings of the 12th international joint conference on computer vision, imaging and computer graphics theory and applications—volume 4: VISAPP, (VISIGRAPP 2017); 2017, p. 459–67. <https://doi.org/10.5220/0006126404590467>.
35. Paulus S, Dupuis J, Riedel S, Kuhlmann H. Automated analysis of barley organs using 3D laser scanning: an approach for high throughput phenotyping. *Sensors*. 2014;14(7):12670–86. <https://doi.org/10.3390/s140712670>.
36. Elnashef B, Filin S, Lati RN. Tensor-based classification and segmentation of three-dimensional point clouds for organ-level plant phenotyping and growth analysis. *Comput Electron Agric*. 2019;156:51–61. <https://doi.org/10.1016/j.compag.2018.10.036>.
37. Klodt M, Cremers D. High-resolution plant shape measurements from multi-view stereo reconstruction. In: Agapito L, Bronstein MM, Rother C, editors. *Computer vision—ECCV 2014 workshops*. Springer, Cham; 2015, p. 174–84.
38. Golbach F, Kootstra G, Damjanovic S, Otten G, van de Zedde R. Validation of plant part measurements using a 3d reconstruction method suitable for high-throughput seedling phenotyping. *Mach Vis Appl*. 2016;27(5):663–80. <https://doi.org/10.1007/s00138-015-0727-5>.
39. Hétroy-Wheeler F, Casella E, Boltcheva D. Segmentation of tree seedling point clouds into elementary units. *Int J Remote Sens*. 2016;37(13):2881–907. <https://doi.org/10.1080/01431161.2016.1190988>.
40. Santos TT, Koenigkan LV, Barbedo JGA, Rodrigues GC. 3D plant modeling: localization, mapping and segmentation for plant phenotyping using a single hand-held camera. In: Agapito L, Bronstein MM, Rother C, editors. *Computer vision—ECCV 2014 workshops*. Springer, Cham; 2015, p. 247–63.
41. Paproki A, Sirault X, Berry S, Furbank R, Fripp J. A novel mesh processing based technique for 3D plant analysis. *BMC Plant Biol*. 2012;12(1):63. <https://doi.org/10.1186/1471-2229-12-63>.
42. Boltcheva D, Casella E, Cumont R, Hétroy F. A spectral clustering approach of vegetation components for describing plant topology and geometry from terrestrial waveform LiDAR data. In: Lintunen A, editor. *7th international conference on functional-structural plant models*, Saariselkä, Finland; 2013. <https://doi.org/10.13140/2.1.1114.1928>.
43. Nguyen CV, Fripp J, Lovell DR, Furbank R, Kuffner P, Daily H, Sirault X. 3D scanning system for automatic high-resolution plant phenotyping. In: 2016 international conference on digital image computing: techniques and applications (DICTA); 2016, p. 1–8. <https://doi.org/10.1109/DICTA.2016.7796984>.
44. Laga H, Guo Y, Tabia H, Fisher R, Bennamoun M. *3D Shape analysis: fundamentals, theory, and applications*. Hoboken: Wiley-Blackwell; 2019.
45. Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans Pattern Anal Mach Intell*. 2004;26(9):1124–37. <https://doi.org/10.1109/TPAMI.2004.60>.
46. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. <https://doi.org/10.1023/A:1010933404324>.
47. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. *Medical image computing and computer-assisted intervention—MICCAI 2016*. Cham: Springer; 2016. p. 424–32.
48. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical image computing and computer-assisted intervention—MICCAI 2015*. Cham: Springer; 2015. p. 234–41.
49. Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge: The MIT Press; 2016.
50. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. 2011;2(1):37–63.
51. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213–20. <https://doi.org/10.1037/h0026256>.
52. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26(3):297–302. <https://doi.org/10.2307/1932409>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Article

Supervised Image Classification by Scattering Transform with Application to Weed Detection in Culture Crops of High Density

Pejman Rasti , Ali Ahmad, Salma Samiei , Etienne Belin  and David Rousseau *

LARIS, UMR INRA IRHS, Université d'Angers, 62 avenue Notre Dame du Lac, 49000 Angers, France; pejman.rasti@univ-angers.fr (P.R.); ali.ahmad@insa-lyon.fr (A.A.); salma.samiei@univ-angers.fr (S.S.); etienne.belin@univ-angers.fr (E.B.)

* Correspondence: david.rousseau@univ-angers.fr

Received: 23 December 2018; Accepted: 23 January 2019; Published: 26 January 2019



Abstract: In this article, we assess the interest of the recently introduced multiscale scattering transform for texture classification applied for the first time in plant science. Scattering transform is shown to outperform monoscale approaches (gray-level co-occurrence matrix, local binary patterns) but also multiscale approaches (wavelet decomposition) which do not include combinatory steps. The regime in which scatter transform also outperforms a standard CNN architecture in terms of data-set size is evaluated (10^4 instances). An approach on how to optimally design the scatter transform based on energy contrast is provided. This is illustrated on the hard and open problem of weed detection in culture crops of high density from the top view in intensity images. An annotated synthetic data-set available under the form of a data challenge and a simulator are proposed for reproducible science. Scatter transform only trained on synthetic data shows an accuracy of 85% when tested on real data.

Keywords: weed detection; scatter transform; deep learning; machine-learning classification; annotation; synthetic data; local binary pattern

1. Introduction

Deep learning is currently tested world-wide in almost all application domains of computer vision as an alternative to purely handcrafted image analysis [1]. When inspecting the convolutional coefficients in the first layers of deep neural networks, these are very similar to Gabor wavelets. While promoting a universal framework, deep neural networks seem to systematically converge toward tools that humans have been studying for decades. This empirical fact is used by computer scientists in the so-called transfer learning where the first layers of an already trained network are re-used [2]. This has also triggered interest by mathematicians to revisit the use of wavelets to produce universal machine-learning architectures. This interdisciplinary cross-talk resulted in the proposal of the so-called scatter transform [3], which is roughly a cascade of wavelet decomposition followed by non-linear and pooling operators. If this deep architecture bares some similarity with the standard deep learning, it does not include the time-consuming feed-forward propagation algorithm. However, it proved its comparable efficiency to deep learning while offering a very rational way of choosing the parameters of the network compared to the rather empirical current art of tuning neural networks.

Despite its intrinsic interest to address multiple scales problems compared to deep learning, scatter transform since its introduction in 2013 has been applied only on a relatively small variety of pattern recognition computer vision problems notably including iris recognition [4], rainfall classification in radar images [5], cell-scale characterization [6,7], or face recognition, [8]. Also, in these applications

scatter transform has shown its efficiency, but it was not systematically compared with other techniques in a comprehensible way. We propose to extend the scope of investigation of the applicability of scatter transform algorithm to plant science with a problem of weed detection in a background of culture crops of high density. This plant science problem is important for field robotics where the mechanical extraction of weed is a current challenge to be addressed to avoid the use of phytochemical products. From a methodological point of view, this classification problem here will also serve as a use case to assess the potential of the scatter transform when compared with other single scale and multiple scales techniques.

A large variety of platforms, sensors, and data process already exist to monitor weeds at various temporal and spatial scales. From remote sensing supported by satellites to cameras located on unmanned aerial vehicles (UAVs) or on ground-based platforms, many systems have been described and compared for the weed monitoring in arable culture crops [9–11]. Related to the observation scale of our use case, by focusing on the imaging scales of UAVs and ground-based platforms, some studies exploiting RGB data have addressed crop weed classification with a large variety of machine-learning approaches. The problem of segmentation of crop fields from typical weeds, performing vegetation detection, plant-tailored feature extraction, and classification to estimate the distribution of crops and weeds has recently been solved with convolutional neural networks in the field [12,13] and in real-time [14]. Earlier, Aitkenhead, M. et al. [15] evaluated weed detection in fields of crop seedlings using simple morphological shape characteristic extraction and self-organizing neural network. Bayesian classifier was used in [16] for plant and weed discrimination. Shape, texture features [12,17–19] or wavelet transform [20,21] coupled with various classifiers including support vector machine (SVM), relevance vector machine (RVM), fuzzy classifier, or random forests were also shown to provide successful pipelines to discriminate between plant and weeds.

The above list of reference is of course not exhaustive and new pipelines will continue to appear because of the large variety of crops shape and imaging platform. In this context, scatter transform constitutes a candidate of possible interest worth to be assessed on a plant–weed classification problem. Also, by comparison with the existing work on weed detection, the computer vision community has focused on the relatively low density of crops and weed where the soil constitutes a background to be classified in addition to crop and weed. In this paper, we consider the case of culture crops of high density, i.e., where the soil is not visible from the top view. In this case, the culture is the background and the object to be detected are weeds of wild type. The contrast in color between the background and the weed, in this case, is obviously here very low by comparison with lower density culture.

2. Material and Methods

We start by introducing the computer vision problem considered, the data-set, the expected scales included in these images and the algorithms tested for comparison with the multiscale scatter transform algorithm.

2.1. Images and Challenges

We consider the situation of a culture crops of a high density of plants (mache salad) with the undesired presence of some weeds. Images were acquired with the imaging system fixed on a robot as displayed in Figure 1. Acquisition trials, as visible in Figure 1, were done under plastic tunnels without additional light. Some sample images are given in Figure 2. Examples of weed detected in such images are shown in Figure 3 to illustrate the variability of shapes among these wild types of weeds. The computer vision task considered in this article consists in detecting the weeds from the top view as shown in the ten real-world images of Figure 2. This is challenging indeed since the intensity or color contrast between weed and crop is very weak. Also, due to the lighting conditions during acquisition, the global intensity may vary from one image to another. The contrast between weeds and plants rather stands in terms of texture since the shape of the plant considered is rather round while

the weeds included in the data-set Figure 3 are much more indented. Therefore, this computer vision problem is well adapted to test scatter transform which is a texture-based technique.



Figure 1. Global view of the imaging system fixed on a robot moving above mache salads of high density. RGB images are captured by a JAI manufactured camera of 20 M pixels with a spatial resolution of 5120×3840 pixels, mounted with a 35 mm objective. The typical distance of plants to camera is of 1 m.

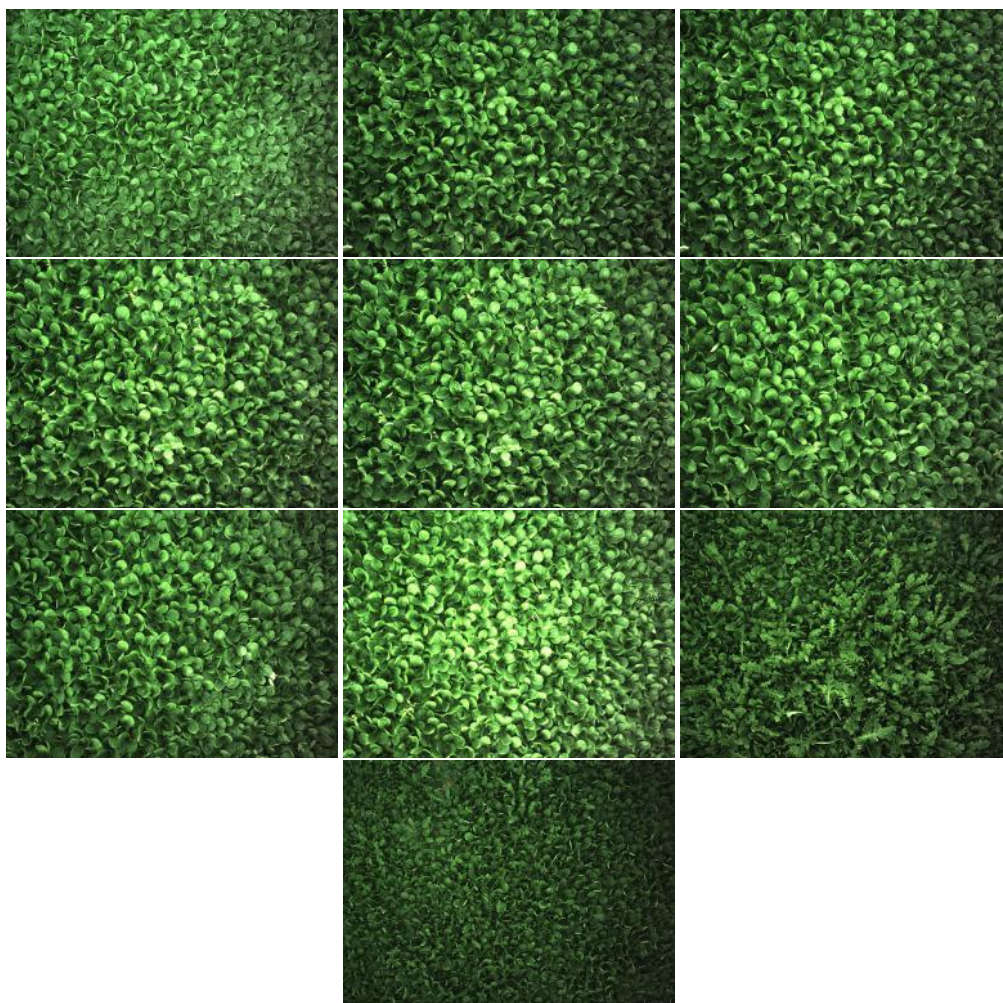


Figure 2. Set of 10 RGB images from top view for the detection of weed out of plant used as testing data-set in this study.

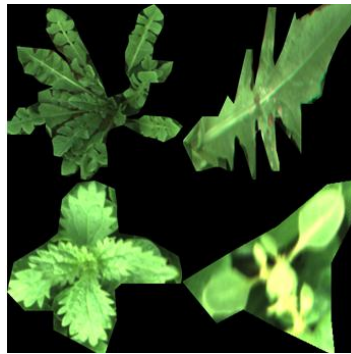


Figure 3. Illustration of different types of weeds used for the experiment.

A ground truth of the position of the weed in the ten images of Figure 2 was produced under the form of finely segmented weed and bounding box patches including these weeds. The total number of weeds being relatively low (21), we decided to generate a larger data-set with synthetic images. To simulate images similar to the real images acquired, we created a simulator which places weeds (among the 21 found in real images) from the annotated weed data-set in images of plants originally free from any weed along the pipeline shown in Figure 4.

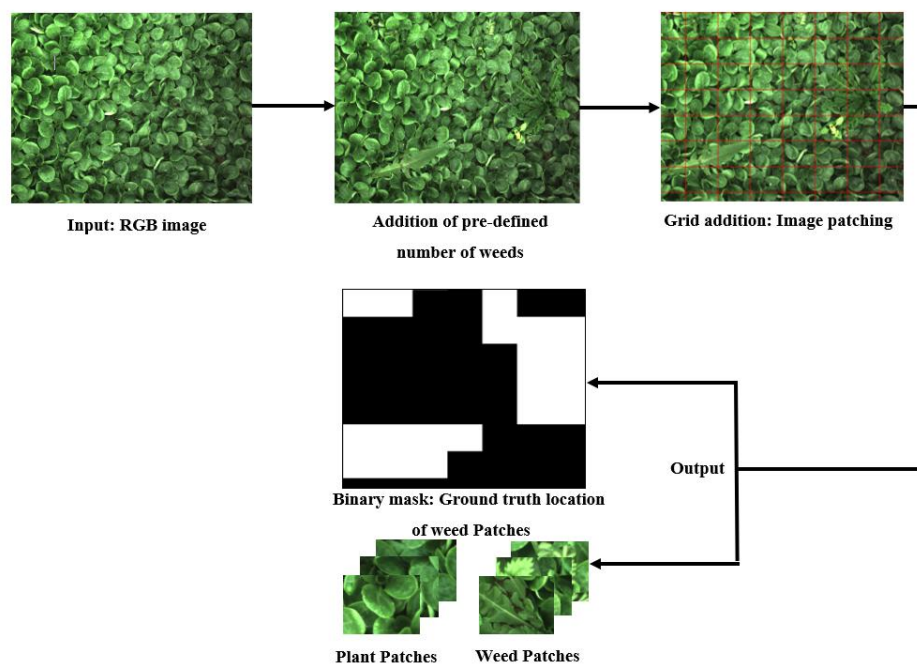


Figure 4. Simulation pipeline for the creation of images of plant with weed of Figure 3 similar to the one presented in Figure 2.

2.2. Scales

With a spatial resolution of 5120 by 3840 pixels included in the images of our data-set, and as illustrated in Figure 5, multiple anatomical structures of the dense weed/plant culture are accessible in our images. From tiny to coarse sizes, i.e., scales, this includes texture in the limb, the veins, and the leaf. There are possibly discriminant features between the two classes (weed/plant) to be found in these three scales either taken individually or combined with each other. To offer the possibility of a multiple scale analysis, together with a reasonably small computation time, classification is done at the scale of patches chosen as double size of the typical size of leaves, $2 \times \max\{S_w, S_p\}$, with rectangles

of 250 by 325 pixels where $S_w = 163$ pixels and $S_p = 157$ on average. With this constraint, we also keep for the patch the same ratio between height and width as in the original image for a periodic patch grid.

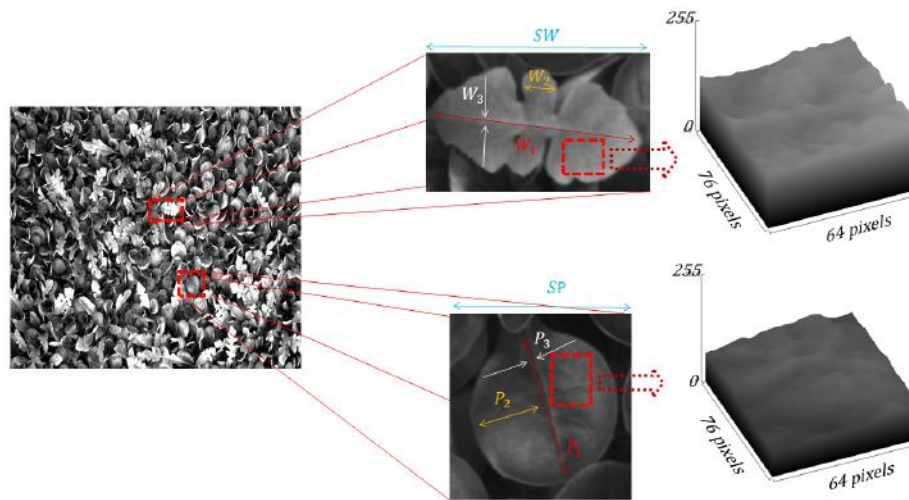


Figure 5. Anatomical scales where (W_i, P_i) presents the scales of weeds and plants respectively; (W_1, P_1) points toward the texture of the limb, (W_2, P_2) indicates the typical size of leaflet and (W_3, P_3) stands for the width of the veins. S_w and S_p show the size of a leaf of weed and plant, respectively. The classification of weed and plant is done at the scale of a patch taken as $2 \times \max(S_p, S_w)$ in agreement with a Shannon-like criteria.

2.3. Data-Set

With the simulator of Figure 4, we produced a total amount of 3292 patches containing weed and 3292 patches only with plants. The binary classification (weed/plant) is realized on these patches. This balanced data-set serves both for the training and the testing stages to assess the performance of different machine-learning tools. The data sets together with the simulator are proposed as supplementary material under the form of a free executable and a set of images (<https://uabox.univ-angers.fr/index.php/s/iuj0knyzOUgsUV9>).

2.4. Classifiers

In this section, we describe how we apply the scatter transform [3] on the weed detection problem introduced in the previous section. For comparison, we then propose a set of alternative techniques. This paper uses independent k-fold cross-validation to measure the performance of the scatter transform coupled to the classifier depicted in Figure 6 and compare other feature extractors coupled to the same classifier. The performances of these classifiers are measured by the metric of the accuracy of correct classification by

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

where TP indicates that the prediction is positive, and the actual value is positive. FP indicates that the prediction value is positive, but the actual value is negative. TN indicates that the prediction value is negative, and the actual value is negative. FN indicates that the prediction value is negative, but the actual value is positive.

2.4.1. Scatter Transform

A scattering transform defines a signal representation which is invariant to translations and potentially to other groups of transformations such as rotations or scaling. It is also stable to deformations and is thus well adapted to image and audio signal classification. A scattering transform is implemented with a convolutional network architecture, iterating over wavelet decompositions and complex modulus. Figure 6 shows a schematic view of a scatter transform network working as a feature extractor and coupled to a classifier after dimension reduction.

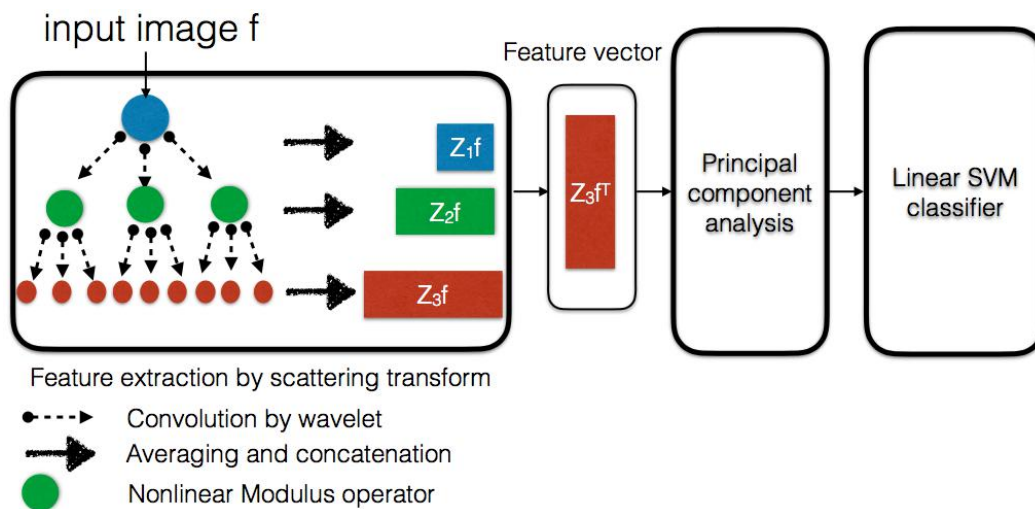


Figure 6. Schematic layout of the weed/plant classifier based on the scattering transform with three layers. The feature vector transmitted to the principal component analysis (PCA) step consists in the scatter vector $Z_m f$ of the last layer of Equation (2) after transposition.

The scatter vectors Z_m at the output of the first three layers $m = 1, 2, 3$ for an input image f are defined by

$$\begin{aligned}
 Z_1 f &= \{|f| \star \phi\} \\
 Z_2 f &= \{\dots, |f \star \psi_{j,\theta}| \star \phi, \dots\} \\
 Z_3 f &= \{\dots, |f \star \psi_{j,\theta}| \psi_{k,\phi} \star \phi, \dots\},
 \end{aligned}
 \tag{2}$$

where the symbol \star denotes the spatial convolution, $|\cdot|$ stands for the L_1 norm, ϕ is an averaging operator, $\psi_{j,\theta}$ is a wavelet dilated by 2^j and rotated by θ . The range of scales $j = \{0, 1, \dots, J\}$ and the number of orientations $\theta = \{0, \pi/L, \dots, \pi(L-1)/L\}$ are fixed by integers J and L . The number of layers is between $m = 1$ to $m = M$. In our case, we considered as mother wavelet the Gabor filter with implementation provided under MATLAB in (<https://www.di.ens.fr/data/scattering/>) for scatter transform.

Scatter transform differs from a pure wavelet decomposition because of the non-linear modulus operator. With this nonlinearity, decomposition of the image is not done on a pure orthogonal basis (whether wavelet basis is orthogonal or not) and this opens the way of a possible benefit in the concatenation of several layers with a combination of wavelet decompositions at different scales. Interestingly, these specific properties of the scatter transform match the intrinsic multiscale textural nature of our weed detection problem which therefore constitutes an appropriate use case to assess the potential of the scatter transform in practice. A visualization of output images for various filter scale j at $m = 2$ for a given orientation is shown in Figure 7. It clearly appears in Figure 7 that the various scales (texture of the limb and veins at $j = 3$, border shape at $j = 4$ and global leaf shape at $j = 8$) presented in Section 2.2 can be captured with the different scaling factor applied on the wavelet. In our study, we empirically picked $L = 8$ orientations and investigated up to $J = 8$ scales since there are

no other anatomical items larger than the leaf itself. The number of layers tested was up to $M = 4$ as proposed in [3] since the energy after some layers although none zero is logically vanishing.

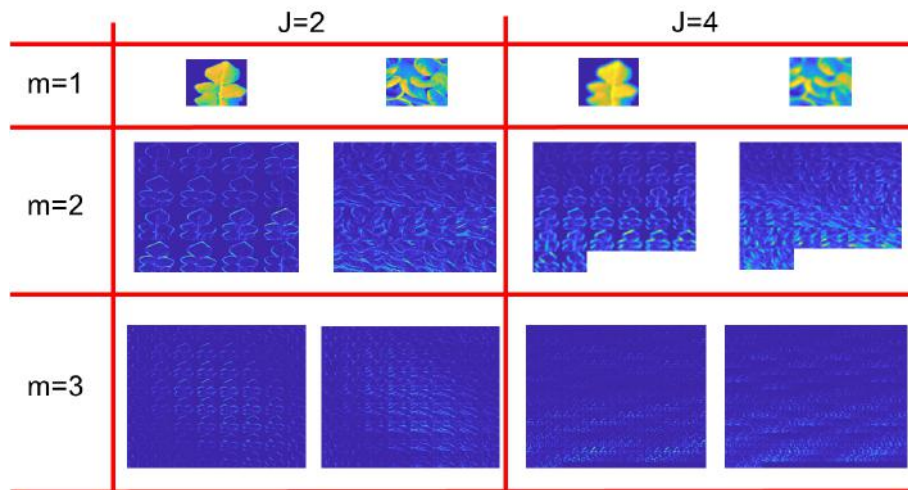


Figure 7. Output images for each class (weed on left and plant on right) and for each layer m of the scatter transform.

In the application of scatter transform to classification found in the literature so far, the optimization of the architecture was done a posteriori after supervised learning. This is rather time-consuming. We investigated the possibility to select a priori the best architecture by analyzing the distribution of relative energy E_m at the output of each layer as given by

$$E_m = \frac{\|Z_m f\|^2}{\|f\|^2} . \tag{3}$$

We computed these energies for the whole data-set as given in Table 1. As noticed in [3], the relative energy is progressively vanishing when the number of layers increases. This observation advocates for the use of a limited number of layers. However, these energies are computed on the whole population of patches including both plants and weeds and therefore it tells nothing about where to find the discriminant energy between each class throughout the feature space produced by the scatter transform. Tables 2 and 3 show the average relative energy for the weeds’ patches data-set, \bar{E}_{w_m} , and plants’ patches data-set, \bar{E}_{p_m} , for different layers m and various maximum scale J .

To show this discriminant energy between each class, various criterion could be proposed. We tested the percentage of energy similarity, Q_m , between the two classes defined by

$$Q_m = \frac{\operatorname{argmin}(\bar{E}_{w_m}, \bar{E}_{p_m})}{\operatorname{argmax}(\bar{E}_{w_m}, \bar{E}_{p_m})} \times 100. \tag{4}$$

According to this criterion, the best architecture of the scatter transform can be chosen at the point of η where the minimum Q_m between each class is found as a function of J by $\eta = \operatorname{argmin}_J(Q_m(J))$. The energy similarity $Q_m(J)$ are represented in Figure 8 and this clearly demonstrates that the contrast between classes is more pronounced on coefficient with small relative energy. This observation, not stressed in the original work of [3], indicates that it should be possible to draw benefit from the contribution of these small discriminative coefficients and thus this demonstrates the interest of the combinatory step of the scatter transform.

Table 1. Average percentage of energy of scattering coefficients E_m on frequency-decreasing paths of length m (scatter layers), with $L = 8$ orientations and various filter scale range, J , for the whole database of plants and weeds patches.

	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$J = 1$	96.18	2.35	-	-	-
$J = 2$	91.81	4.61	0.28	-	-
$J = 3$	85.81	8.46	0.89	0.03	-
$J = 4$	85.81	13.15	1.97	0.17	0.006
$J = 5$	81.46	15.36	3	0.36	0.024
$J = 6$	79.04	16.81	3.44	0.53	0.048
$J = 7$	80.74	17.05	3.49	0.63	0.071

Table 2. Average percentage of energy of scattering coefficients E_m on frequency-decreasing paths of length m (scatter layers), depending upon the maximum scale J and $L = 8$ filter orientations for the weed class patches.

	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$J = 1$	99.90	0.0985	-	-	-
$J = 2$	99.71	0.2798	0.0098	-	-
$J = 3$	99.07	0.8832	0.0443	0.0016	-
$J = 4$	97.55	2.2669	0.1663	0.0080	0.0003
$J = 5$	95.10	4.3892	0.4667	0.0343	0.0020
$J = 6$	92.07	6.8696	0.9522	0.0983	0.0076
$J = 7$	89.26	9.0102	1.5049	0.1979	0.0196

Table 3. Average percentage of energy of scattering coefficients on frequency-decreasing paths of length m (scatter layers), depending upon the maximum scale J and $L = 8$ filter orientations for the plant class patches.

	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$J = 1$	99.92	0.0711	-	-	-
$J = 2$	99.76	0.2339	0.0040	-	-
$J = 3$	99.17	0.7984	0.0281	0.0003	-
$J = 4$	97.75	2.0899	0.1380	0.0041	0.00003
$J = 5$	95.41	4.1411	0.4215	0.0254	0.0006
$J = 6$	92.34	6.6553	0.9078	0.0892	0.005
$J = 7$	89.37	8.9341	1.4817	0.1944	0.0171

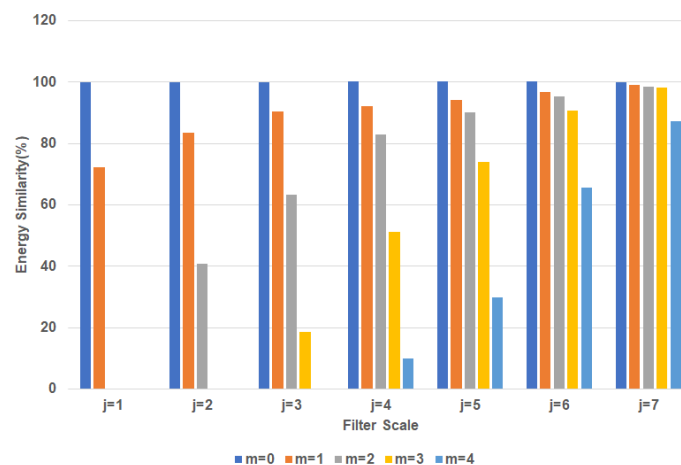


Figure 8. Energy similarity, $Q_m(J)$, between energy of weeds and plants data sets based on Tables 2 and 3.

Also, from the observation of Figure 8, our approach indicates that a priori the best discriminant energy between each class is to be expected with a scatter architecture corresponding to $M = 4$ and $J = 4$ which provides the minimum energy similarity, η , between the energy of images of the weeds' class and the plants' class.

2.4.2. Other Methods

To assess the possible interest of the scatter transform in our weed detection problem, we consider several alternative feature extractor algorithms. First, since the scatter transform by construction works on a feature space which includes multiple scales, it is expected to perform better than any state of the art monoscale method, i.e., working on a feature space tuned on a single size, when applied on a multiple scales problem (such as the one we have here with veins, limb, leaf). Second, since the scatter transform works on a combination of wavelet decomposition between scales it should perform slightly better than a pure wavelet decomposition chosen on the same wavelet basis but without the use of the non-linear operator nor the scales combination. Finally, because scatter transform shares some similarities with convolutional neural networks it should also be compared with the performance obtained with a deep learning algorithm. Based on this rationale, we propose the following alternative feature extractor for comparison with the feature extractor of the scatter transform where the same PCA followed by a linear SVM is used for the classification.

Local binary pattern: Under the original form of [22] and as used in this article, for a pixel positioned at (x, y) , local binary pattern (LBP) indicates a sequential set of the binary comparison of its value with the eight neighbors. In other words, the LBP value assigned to each neighbor is either 0 or 1, if its value is smaller or greater than the pixel placed at the center of the mask, respectively. The decimal form of the resulting 8-bit word representing the LBP code can be expressed as follows

$$LBP(x, y) = \sum_{n=0}^7 2^n s(i_n - i_{x,y}) \quad (5)$$

where $i_{x,y}$ corresponds to the gray value of the center pixel, and i_n denotes that of the n th neighboring one. Besides, the function $\zeta(x)$ is defined as follows

$$\zeta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases} \quad (6)$$

The LBP operator remains unaffected by any monotonic gray scale transformation which preserves the pixel intensity order in a local neighborhood. It is worth noticing that all the bits of the LBP code hold the same significance level, where two successive bits value may have different implications. The process of Equation (5) is produced at the scale of the patch defined in the previous section. The $LBP(x, y)$ of each pixel inside this patch are concatenated to create a fingerprint of the local texture around the pixel at the center of the patch. Equations (5) and (6) are applied on all patches of an image.

Gray-Level Co-Occurrence Matrix: A statistical approach that can well describe second-order statistics of a texture image is provided by the so-called gray-level co-occurrence matrix (GLCM). GLCM was firstly introduced by Haralick et al. [23]. A GLCM is essentially a two-dimensional histogram in which the (i, j) th element is the frequency of event i co-occurring with event j . A co-occurrence matrix is specified by the relative frequencies $C(i, j, d, \theta)$ in which two pixels, separated by a distance d , occurs in a direction specified by the angle θ , one with gray-level i and the other with gray-level j . A co-occurrence matrix is therefore a function of distance d , angle θ and grayscales i and j .

In our study, as perceptible in images of Figure 2, the weed-plant structures are isotropic meaning that they show no specific predominant orientations. As a logical consequence, and as already stated in similar weed classification problem using GLCM [24–26], choosing multiple orientations θ would not improve the classification performance. We therefore arbitrarily chose a fixed $\theta = 0$ which enables to probe on average leaves positioned in all directions. For distance, d , it is taken at $d = 2$ pixels

which correspond to a displacement capable of probing the presence of edges, veins, and structures in the limb.

Gabor filter: Same Gabor filters as in the scatter transform were applied to the images to produce a feature space. By contrast with the scatter transform, no non-linearities are included in this process and only one layer of filters is applied. For a fair comparison in this experiment, scale range J and number of orientations L of the Gabor filter bank are chosen at the same value as in the scatter transform.

Deep learning: Representation learning, or deep learning, aims at jointly learning feature representations with the required prediction models. We chose the predominant approach in computer vision, namely deep convolutional neural networks [27]. The baseline approach resorts to standard supervised training of the prediction model (the neural network) on the target training data. No additional data sources are used. In particular, given a training set comprised of K pairs of images f_i and labels \hat{y}_i , we train the parameters θ of the network r using stochastic gradient descent to minimize empirical risk:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^K \mathcal{L}(\hat{y}_i, r(f_i, \theta)) \quad (7)$$

\mathcal{L} denotes the loss function, which is cross-entropy in our case. The minimization is carried out using the ADAM optimizer [28] with a learning rate of 0.001.

The architecture of network $r(\cdot, \cdot)$, shown in Figure 9, has been optimized on a hold-out set and is given as follows: five convolutional layers with filters of size 3×3 and respective numbers of filters 64, 64, 128, 128, 256 each followed by ReLU activations and 2×2 max pooling; a fully connected layer with 1024 units, ReLU activation and dropout (0.5) and a fully connected output layer for 2 classes (weeds, plants) and SoftMax activation. Given the current huge interest on deep learning many other architectures could be tested and possibly provide better results. As a disclaimer, we stress that the architecture proposed in Figure 9 is of course not expected to provide the best performance achievable with any neural network architecture. Here the tested CNN serves as a simple reference with a level of complexity of the architecture adapted to the size of the input image and training data sets.

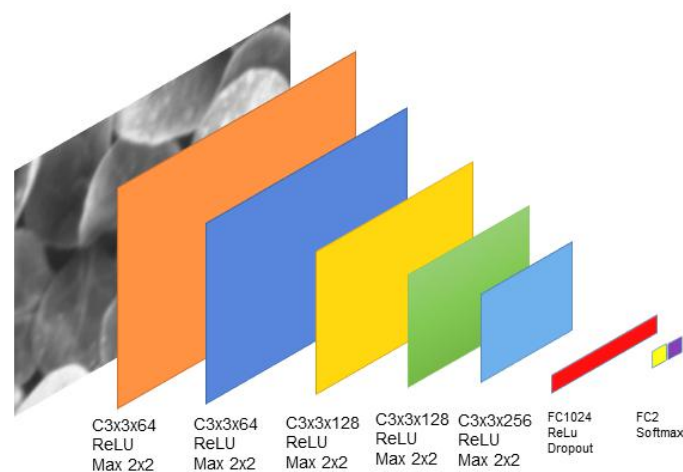


Figure 9. Architecture of the deep network optimized for the task on classification.

3. Result

In this section, we provide experimental results using the experimental protocol for the assessment of scatter transform (Section 2.4) as well as the different alternative feature extraction techniques chosen for comparison in Section 2.4.2.

The scatter transform produces a data vector containing the $Z_m f$ of Equation (2) whose dimension is reduced by a standard PCA and then applied to a linear kernel SVM. To compare the performance of different structures of scatter transform on the database, we used a different combination of filter scales, j , and the number of layers, m , to realize which structure is the best fit for our data. Table 4 shows the classification accuracy of these structures where 10-fold cross-validation approach is used for classification. The best weed/plant classification results with scatter transform are obtained for $J = 4$ and $m = 4$. This a posteriori exactly corresponds to the prediction done a priori from the energy-based approach presented in the method section.

Table 4. Percentage of correct classification for 10-fold cross-validation classification on simulation data with scatter transform for various values of m and J .

	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$	$J = 6$	$J = 7$	$J = 8$
$m = 1$	70.37%	77.89%	82.74%	86.17%	88.96%	91.94%	94.14%	95.05%
$m = 2$	—	91.95%	95.26%	95.54%	95.86%	95.82%	95.73%	95.55%
$m = 3$	—	—	95.41%	95.44%	95.21%	95.07%	95.03%	96.00%
$m = 4$	—	—	—	96.31%	96.02%	96.05%	96.16%	96.11%

We considered this optimal scatter transform structure with $J = 4$ and $m = 4$ and compared it with all alternative methods described in Section 2.4. Table 5 shows the recognition rates of weed detection on the data where a k-fold cross-validation approach of SVM classification with the different number of folds is used. Scatter transform appears to outperform all compared handcrafted methods. This demonstrates the interest of the multiscale and combinatory feature space produced by scatter transform. It is important to notice that to have a fair comparison of these alternative methods, we adapted the feature spaces of all algorithms to the same size. The minimum size of the whole feature space is selected, and feature space of other algorithms are reduced to that specific size. In our techniques, the minimum feature space belongs to the GLCM method which has a size of $N \times 19$ where N represents the number of samples. The PCA algorithm is adapted to our models to reduce the dimensions of the feature space generated by other techniques to the size of $N \times 19$.

As shown in Table 5 and Figure 10, when compared with CNN, like most handcrafted methods, scatter transform performs better for small data sets. The limit where CNN and scatter transform are found to perform equally is found to be 10^4 on the weed detection problem as given in Figure 10. This demonstrates the interest of the scatter transform in case of rather small data sets. It is, however, to be noticed that an intrinsic limitation of scatter transform is that it works only with patches to perform a classification while some architectures of convolutional neural network would also be capable of performing segmentation directly in the whole image (see for instance U-Net) [29].

Table 5. Percentage of correct classification by using k-fold Cross-validation on simulated data.

	5 Folds	6 Folds	7 Folds	8 Folds	9 Folds	10 Folds	Average std
Scatter Transform (0.6584×10^4 samples)	94.9%	95.2%	95.3%	95.7%	95.8%	95.8%	± 1.1
LBP (0.6584×10^4 samples)	85.5%	86.1%	86.3%	85.8%	86.9%	86.7%	± 0.4
GLCM (0.6584×10^4 samples)	87.4%	91.6%	90.9%	92.1%	92.4%	92.3%	± 0.7
Gabor Filter (0.6584×10^4 samples)	88.0%	88.2%	88.7%	88.6%	89.4%	89.3%	± 1.3
Deep Learning (0.6584×10^4 samples)	89.4%	89.9%	91.1%	91.5%	91.9%	92.1%	± 1.4
Deep Learning (2.8×10^4 samples)	97.6 %	97.9 %	97.9 %	98.2%	98.1%	98.3%	± 0.9

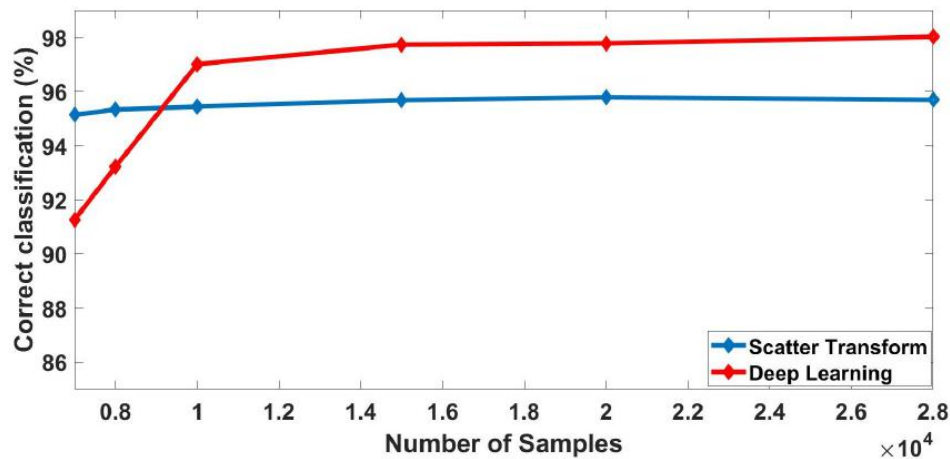


Figure 10. Comparison of the recognition accuracy between scatter transform and deep learning when the number of samples increases.

4. Discussion

So far, we focused in this article on detection of weeds in fields by the scatter transform algorithm with a comparison of other machine-learning techniques which have been trained and tested on synthetic images produced by the simulator of Figure 4. Our experimental results show that a good recognition rate of weeds detection (approximately 95%) can be achievable by the scatter transform algorithm. On the other hand, other alternative methods also work well for this problem with a minimum recognition rate around 85%. These experiments prove that texture-based algorithms can be useful for weed detection in culture crops of high density.

One may wonder how these classification results compare toward the literature on weed detection in less dense culture cited in the introduction section [12–21]. The performance in this literature varies from 75% to 99% of good detection of weed. It is, however, difficult to provide a fair comparison since in addition to the main difference with the absence of soil, the observation scales together with the acquisition conditions vary from one study to another.

One may wonder how these algorithms trained on synthetic data behave when they are applied to real images including plant background and weed not included in the synthetic data sets. We also tested our scatter transform classifier which was trained on synthetic data when applied on the real images of Figure 2. On average for all 10 real images, the accuracy found is 85.64%. Although this constitutes already interesting results, this indicates a bias between simulated data and real data. One direction could be to improve the realism of the simulator. In the version proposed here weeds were not necessarily acquired in the same lighting conditions as the plant. A simple upgrade could be to adapt the average intensity on the weed and the plant to compensate for this artifact or, since in plant and weed can indeed be of various intensity, to generate data augmentation with various contrast. However, simulators never exactly reproduce reality. Another approach to improve the performance of the training based on simulated data would be to add a step of domain adaptation after the scatter transform [30]. So far, the best and worst results obtained with scatter transform are given in Figure 11. A possible interpretation for the rather low performance in Figure 11b is the following. The density of weed in Figure 11b is very high compared to the other images in the training data-set. Consequently, the local texture in the patch may be very different from the one obtained when weeds appear as outliers. This demonstrates that the proposed algorithm, trained on synthetic data, is appropriate in the low density of weeds at an observation scale such as the one chosen for the patch where plant serves as a systematic background.

These performances could be improved in several ways. First, a large variety of weeds can be found in nature and it would be important to include more of this variability in the training data sets. Also, weeds are fast growing plants capable of winning the competition for light. Therefore,

high percentages of weed are expected to come with higher weeds than in very low percentage of the surface of weeds. This fact illustrated in Figure 11 is not included in the simulator where weeds of a fixed size are randomly picked. Such example of enrichment of the training data-set and simulator could be tested easily following the global methodology presented in this article to assess the scatter transform. Finally, we did not pay much effort on denoising the data. The proposed data have been acquired with a camera fixed on an unmanned vehicle. Compensation for variation of illumination in the data-set, or inside the images, themselves or compensation for the possible optical aberration of the camera used could also constitute directions of investigation to improve the weed/plant detection. All the methods presented in this paper (including scatter transform) have the capability to be robust to global variation of light intensity however the variation of light direction during the day may impact the captured textures. Increasing the data-set to acquire images at all hour of a working day or adding a lighting cabinet on the robot used would make the results even more robust [14,31–33].

The problem of weed detection in culture crops of high density is an open problem in agriculture which we believe deserves the organization of a challenge similar to the one organized on Arabidopsis in controlled conditions [34] for a biology community. Such challenges contribute to improving the state of the art as recently illustrated with the use of simulated Arabidopsis data to boost and speed up the training [35] in machine learning. This challenge is now open on the codalab platform (<https://competitions.codalab.org/competitions/20075>) together with the effort of proposing real data and the simulator (<https://uabox.univ-angers.fr/index.php/s/ij0knyzOUgsUV9>) developed for this article. These additional materials, therefore, contributes to the opening of the problem of weed detection in culture crops of high density to a wider computer vision community.

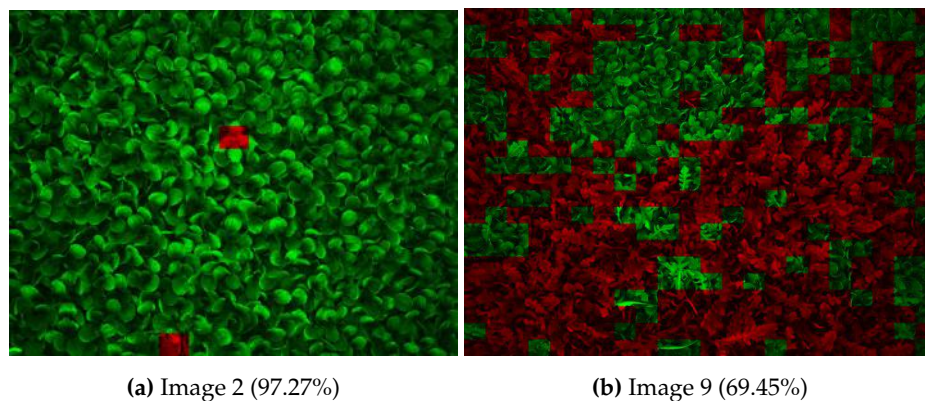


Figure 11. Visual comparison of the best and the worst recognition of weeds and plants by scatter transform.

5. Conclusions and Perspectives

In this article, we proposed the first application of the scatter transform algorithm to plant sciences with the problem of weed detection in a background of culture crops of high density. This open plant science problem is important for field robotics where the mechanical extraction of weed is a current challenge to be addressed to avoid the use of phytochemical products.

We assessed the potential of the scatter transform algorithm in comparison with single scale and multiscale techniques such as LBP, GLCM, Gabor filter, and convolutional neural network. Experimental results showed the superiority of the scatter transform algorithm with a weed detection accuracy of approximately 95% over the other single scale and multiscale techniques on this application. Though the comparison was not intended to be exhaustive among the huge literature on texture analysis, the variety of tested techniques contributes to confirm the effectiveness of using the scatter transform algorithm as a valuable multiscale technique for a problem of weed detection and opened an interesting approach for similar problems in plant sciences. Finally, an optimization method based on energy at the output of the scatter transform has been successfully proposed to select a priori the best scatter transform architecture for a classification problem.

Concerning the weed-plant detection, our optimal solution with scatter transform can serve as a first reference of performance and other machine-learning techniques could now be tested in the framework of the data challenge that we launched for this article (<https://competitions.codalab.org/competitions/20075>). As a possible perspective of the investigation, one could further optimize the scatter transform classifier proposed in this paper. For instance, the size of the grid could be fine-tuned or some hyperparameters could be added with non-linear kernels in the SVM step. Also, weed/plant detection was focused here on a binary classification since no distinction between the different weeds were included. In another direction, one could also envision to extend this work to a multiple types of weeds classification problem if more data were included.

Author Contributions: Conceptualization, P.R., A.A. and D.R.; Data curation, A.A. and S.S.; Formal analysis, E.B.; Funding acquisition, E.B. and D.R.; Methodology, D.R.; Resources, E.B.; Software, P.R., A.A. and S.S.; Supervision, D.R.; Validation, P.R., S.S. and D.R.; Visualization, P.R.; Writing—original draft, P.R. and D.R.

Funding: This research received no external funding.

Acknowledgments: Acquisitions of real-images were done in the framework of the project PUMAGri, supported from the Fonds Unique Interministeriel (FUI-BPI France). Authors thank Sixin Zhang from École Normale Supérieure, Paris France for useful discussions. Salma Samiei acknowledges Angers Loire Métropole for the funding of her PhD.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Janowczyk, A.; Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Informat.* **2016**, *7*. [[CrossRef](#)] [[PubMed](#)]
2. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [[CrossRef](#)]
3. Bruna, J.; Mallat, S. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1872–1886. [[CrossRef](#)] [[PubMed](#)]
4. Minaee, S.; Abdolrashidi, A.; Wang, Y. Iris recognition using scattering transform and textural features. In Proceedings of the Signal Processing and Signal Processing Education Workshop (SP/SPE), Salt Lake City, UT, USA, 9–12 August 2015; pp. 37–42.
5. Lagrange, M.; Andrieu, H.; Emmanuel, I.; Busquets, G.; Loubrié, S. Classification of rainfall radar images using the scattering transform. *J. Hydrol.* **2018**, *556*, 972–979. [[CrossRef](#)]
6. Li, B.H.; Zhang, J.; Zheng, W.S. HEp-2 cells staining patterns classification via wavelet scattering network and random forest. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 406–410.
7. Rakotomamonjy, A.; Petitjean, C.; Salaün, M.; Thiberville, L. Scattering features for lung cancer detection in fibered confocal fluorescence microscopy images. *Artif. Intell. Med.* **2014**, *61*, 105–118. [[CrossRef](#)] [[PubMed](#)]
8. Yang, X.; Huang, D.; Wang, Y.; Chen, L. Automatic 3d facial expression recognition using geometric scattering representation. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Voume 1, pp. 1–6.
9. Torres-Sánchez, J.; López-Granados, F.; De Castro, A.I.; Peña-Barragán, J.M. Configuration and specifications of an unmanned aerial vehicle (UAV) for early site specific weed management. *PLoS ONE* **2013**, *8*, e58210. [[CrossRef](#)] [[PubMed](#)]
10. Peña, J.M.; Torres-Sánchez, J.; Serrano-Pérez, A.; de Castro, A.I.; López-Granados, F. Quantifying efficacy and limits of unmanned aerial vehicle (UAV) technology for weed seedling detection as affected by sensor resolution. *Sensors* **2015**, *15*, 5609–5626. [[CrossRef](#)]
11. Fernández-Quintanilla, C.; Peña, J.; Andújar, D.; Dorado, J.; Ribeiro, A.; López-Granados, F. Is the current state of the art of weed monitoring suitable for site-specific weed management in arable crops? *Weed Res.* **2018**. [[CrossRef](#)]
12. Bakhshipour, A.; Jafari, A. Evaluation of support vector machine and artificial neural networks in weed detection using shape features. *Comput. Electron. Agric.* **2018**, *145*, 153–160. [[CrossRef](#)]

13. Lottes, P.; Khanna, R.; Pfeifer, J.; Siegwart, R.; Stachniss, C. UAV-based crop and weed classification for smart farming. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3024–3031.
14. Milioto, A.; Lottes, P.; Stachniss, C. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 2229–2235.
15. Aitkenhead, M.; Dalgetty, I.; Mullins, C.; McDonald, A.J.S.; Strachan, N.J.C. Weed and crop discrimination using image analysis and artificial intelligence methods. *Comput. Electron. Agric.* **2003**, *39*, 157–171. [[CrossRef](#)]
16. Marchant, J.; Onyango, C. Comparison of a Bayesian classifier with a multilayer feed-forward neural network using the example of plant/weed/soil discrimination. *Comput. Electron. Agric.* **2003**, *39*, 3–22. [[CrossRef](#)]
17. Prema, P.; Murugan, D. A novel angular texture pattern (ATP) extraction method for crop and weed discrimination using curvelet transformation. *ELCVIA Electron. Lett. Comput. Vis. Image Anal.* **2016**, *15*, 27–59. [[CrossRef](#)]
18. Ahmad, A.; Guyonneau, R.; Mercier, F.; Belin, É. An Image Processing Method Based on Features Selection for Crop Plants and Weeds Discrimination Using RGB Images. In *International Conference on Image and Signal Processing*; Springer: Berlin, Germany, 2018; pp. 3–10.
19. Haug, S.; Michaels, A.; Biber, P.; Ostermann, J. Plant classification system for crop/weed discrimination without segmentation. In Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), Steamboat Springs, CO, USA, 24–26 March 2014; pp. 1142–1149.
20. Bakhshipour, A.; Jafari, A.; Nassiri, S.M.; Zare, D. Weed segmentation using texture features extracted from wavelet sub-images. *Biosyst. Eng.* **2017**, *157*, 1–12. [[CrossRef](#)]
21. Bossu, J.; Gée, C.; Jones, G.; Truchetet, F. Wavelet transform to discriminate between crop and weed in perspective agronomic images. *Comput. Electron. Agric.* **2009**, *65*, 133–143. [[CrossRef](#)]
22. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [[CrossRef](#)]
23. Haralick, R.M.; Shanmugam, K.; Its'Haik, D. Textural features for image classification. *IEEE Trans. Syst. Man Cybernet.* **1973**, *6*, 610–621. [[CrossRef](#)]
24. Shearer, S.A.; Holmes, R. Plant identification using color co-occurrence matrices. *Trans. ASAE* **1990**, *33*, 1237–1244. [[CrossRef](#)]
25. Burks, T.; Shearer, S.; Payne, F. Classification of weed species using color texture features and discriminant analysis. *Trans. ASAE* **2000**, *43*, 441. [[CrossRef](#)]
26. Chang, Y.; Zaman, Q.; Schumann, A.; Percival, D.; Esau, T.; Ayalew, G. Development of color co-occurrence matrix based machine vision algorithms for wild blueberry fields. *Appl. Eng. Agric.* **2012**, *28*, 315–323. [[CrossRef](#)]
27. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1.
28. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980 .
29. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin, Germany, 2015; pp. 234–241.
30. Courty, N.; Flamary, R.; Tuia, D.; Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1853–1865. [[CrossRef](#)] [[PubMed](#)]
31. Slaughter, D.; Giles, D.; Downey, D. Autonomous robotic weed control systems: A review. *Comput. Electron. Agric.* **2008**, *61*, 63–78. [[CrossRef](#)]
32. Fadlallah, S.; Goher, K. A review of weed detection and control robots: A world without weeds. In *Advances in Cooperative Robotics*; World Scientific: Singapore, 2017; pp. 233–240.
33. Brown, R.B.; Noble, S.D. Site-specific weed management: sensing requirements—What do we need to see? *Weed Sci.* **2005**, *53*, 252–258. [[CrossRef](#)]

34. Scharr, H.; Minervini, M.; Fischbach, A.; Tsafaris, S.A. Annotated image datasets of rosette plants. In Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2014; pp. 6–12.
35. Ubbens, J.; Cieslak, M.; Prusinkiewicz, P.; Stavness, I. The use of plant models in deep learning: An application to leaf counting in rosette plants. *Plant Methods* **2018**, *14*, 6. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Local spatio-temporal encoding of raw perfusion MRI for the prediction of final lesion in stroke

Mathilde Giacalone^{a,1}, Pejman Rasti^{b,1}, Noelie Debs^a, Carole Frindel^a, Tae-Hee Cho^a, Emmanuel Grenier^c, David Rousseau^{b,*}

^aCREATIS, CNRS UMR-5220, INSERM U1206, Université Lyon 1, INSA Lyon Bât. Blaise Pascal, 7 avenue Jean Capelle, Villeurbanne 69621, France

^bLARIS, UMR IRHS INRA, Université d'Angers 62 avenue Notre Dame du Lac, Angers 49000, France

^cENS-Lyon, UMR CNRS 5669 'UMPA', and INRIA Alpes, project NUMED, Lyon F-69364, France

ARTICLE INFO

Article history:

Received 17 October 2017

Revised 28 July 2018

Accepted 31 August 2018

Available online 23 September 2018

Keywords:

Local binary pattern

Support vector machine

Dynamic susceptibility contrast perfusion

MRI

Stroke

ABSTRACT

We address the medical image analysis issue of predicting the final lesion in stroke from early perfusion magnetic resonance imaging. The classical processing approach for the dynamical perfusion images consists in a temporal deconvolution to improve the temporal signals associated with each voxel before performing prediction. We demonstrate here the value of exploiting directly the raw perfusion data by encoding the local environment of each voxel as a spatio-temporal texture, with an observation scale larger than the voxel. As a first illustration for this approach, the textures are characterized with local binary patterns and the classification is performed using a standard support vector machine (SVM). This simple machine learning classification scheme demonstrates results with 95% accuracy on average while working only raw perfusion data. We discuss the influence of the observation scale and evaluate the interest of using post-processed perfusion data with this approach.

© 2018 Published by Elsevier B.V.

1. Introduction

Cerebrovascular diseases represent a leading cause of disability and mortality worldwide (Towfighi and Saver, 2011; Feigin et al., 2014; Murray et al., 2015). Ischemic stroke (85% of all stroke cases) results from an acute occlusion of a cerebral artery. Early restoration of blood flow within the ischemic tissue (reperfusion), using intravenous thrombolysis and/or mechanical thrombectomy, is the most effective therapy to reduce infarct growth and promote clinical recovery (Goyal et al., 2016). The clinical benefit of reperfusion is highly dependent on the extent of the ischemic, but still viable, cerebral tissue (i.e. ischemic penumbra) (Nogueira et al., 2018; Albers et al., 2018). Late revascularization procedures, in patients with extensive, irreversible cerebral damage, may have little to no impact on their neurological recovery. Still, the rate of infarct growth within the ischemic penumbra is highly heterogeneous across patients, due to varying degree of cortical collateral flow, and may spatially vary over time due to regional hemodynamic fluctuations. This inter-individual heterogeneity complicates

both acute treatment decisions and the evaluation of novel therapeutic strategies.

Thus, numerous research groups around the world focus on the challenging topic of ischemic stroke prediction and classification. Their work has allowed for a better understanding of the underlying mechanisms involved during stroke, but the prediction of the evolution of the ischemic lesion is still an open question (Rekik et al., 2012; Wintermark et al., 2013). There is a great variability of stroke evolution among patients which current prediction models fail to describe, suggestive of a complex interaction between multiple co-factors. The ISLE challenge (Maier et al., 2017) for example is a testimony to the current interest of the research community for final stroke lesion prediction.

In this context, Magnetic Resonance Imaging (MRI) is often used for the prediction of stroke lesion, notably perfusion MRI, which is generally used to evaluate hemodynamic parameters maps indicative of the state of perfusion in the cerebral tissues (Davis et al., 2003). A perfusion MRI is a 4-dimensional spatio-temporal image. It is constituted of a time series of 3-dimensional MR images, the acquisitions of which are synchronized with the intravascular bolus injection of a contrast-agent. The temporal signature of the MR signal recorded in each voxel will depend on the state of perfusion of the tissue within the voxel and the MR temporal signals are therefore used in practice to extract semi-quantitative or quantitative hemodynamic parameters. Over the years, numerous

* Corresponding author.

E-mail address: david.rousseau@univ-angers.fr (D. Rousseau).

¹ These authors contributed equally to this work.

post-processing treatments have been proposed to approach perfusion MRI to a quantitative imaging modality (Willats and Calamante, 2013a). The objective of these post-treatments was motivated by the idea of detecting, via a simple thresholding method, the different pathological brain regions. However, there are still many difficulties to overcome in order to achieve an accurate and robust calculation of perfusion parameters and some researchers question the interest of such a quantitative approach in perfusion MRI (Perthen et al., 2002; Meijs et al., 2015). In the end, perfusion MRI has principally been used as an alternative modality to the PET imaging modality, a quantitative imaging modality which is rarely available in hospital centers. However, is this not missing the full potentiality of the perfusion MRI modality? Perfusion MRI is indeed a dynamic image modality but it is usually considered either as a temporal entity during post-treatment, by considering each voxel independently or as a spatial entity, when summarizing the information it contains by a 3-dimensional image. In practice, only a few recent approaches try to exploit the spatio-temporal nature of the data. For example, the spatio-temporal nature of the data was recently taken into account for the resolution of the ill-posed inverse problem of deconvolution, necessary for the extraction of quantitative hemodynamic parameters (He et al., 2010; Schmid, 2011; Frindel et al., 2014). Also, perfusion MRI has been used to quantify collateral flow (Kim et al., 2014), which plays a major role in stroke lesion evolution. Some research team also started to revisit the predictive power of the temporal signals of the raw MRI perfusion image, whether with simple approaches using local similarity descriptors, for example Frindel et al. (2012) or with more complex approaches using state of the art machine learning methods such as convolutional artificial neural networks (Ho et al., 2016).

In this paper, we investigate the potential of new descriptors directly extracted from raw perfusion MR images for the classification of tissue fate. Since the human eye is capable of distinguishing brain regions exhibiting pathological spatio-temporal behaviors on raw perfusion MRI, even without any specific post-processing of the data, we propose here to study the predictive potential of raw perfusion MRI when considering the local spatio-temporal behavior of each voxel as a texture.

Some recent works Huang et al. (2010), Scalzo et al. (2012) and Giacalone et al. (2017) have demonstrated the interest of a “regional” approach for the question of the prediction of tissue fate in stroke over a single voxel approach. This was done in Huang et al. (2010) and Scalzo et al. (2012) with a classifier or regression model between temporal parameter extracted from perfusion weighted images (PWI) and the gray level of the final FLAIR. In Giacalone et al. (2017) the regional approach of stroke was demonstrated with predictability metrics from information theory applied to binarized input-output data from PWI and FLAIR images. In these studies the considered images were post-processed.

We propose a new approach which consists in encoding into a patch directly the spatio-temporal information contained in the regional environment of each voxel without any post-processing. We then evaluate the potential of this patch for the voxel-wise prediction of tissue fate. To do so, each patch is described using texture descriptors which are used to classify the voxel associated with the patch depending on its chances of survival. Following this texture approach, we propose to use the local binary pattern (LBP) as texture descriptors and a support vector machine (SVM) classifier for the classification. We will address the importance of observation scale optimization and image denoising. We will also compare the performance of the proposed approach when post-processed perfusion data are used instead of raw perfusion data.

2. Material

MRI data were extracted from a cohort of patients acquired in Hospices Civils de Lyon (Hermitte et al., 2013). We worked on longitudinal data from four patients affected by an ischemic stroke of the anterior circulation. Those four patients did not receive any thrombolytic treatment and did not reperfuse on their own. The brain regions exhibiting a pathological hemodynamic behavior in the acute stage are therefore highly susceptible to end up dead and form the final ischemic lesion. All patients underwent the following MRI protocol on admission: diffusion-weighted-imaging (DWI; repetition time 6000 ms, field of view 24 cm, matrix 128×128 , slice thickness 5 mm), Fluid-attenuated-inversion-recovery (FLAIR; repetition time, 8690 ms; echo time, 109 ms; inversion time 2500 ms; flip angle, 150° ; field of view, 21 cm; matrix, 224×256 ; 24 sections; section thickness, 5 mm), T2-weighted gradient echo, MR-angiography and dynamic susceptibility-contrast perfusion imaging (DSC-PWI; echo time 40 ms, repetition time 1500 ms, field of view 24 cm, matrix 128×128 , 18 slices, slice thickness 5 mm; gadolinium contrast at 0.1 mmol/kg injected with a power injector). A follow-up MRI was performed at 1-month, including the same sequences minus the DSC-PWI. The MRI sequences used here were acquired with a 1.5 Tesla MRI scanner. A motion correction was applied on the raw perfusion MRI. We registered, slice by slice, all time points on the first time point with a maximum mutual information approach. This was done by registering each temporal point ($n+1$) on its previous temporal point (n) and by then applying recursively the transformation matrices obtained until all time point is aligned with the first time point. The segmentation mask of the final lesion was delineated for each patient on the one-month follow-up FLAIR-MRI by experts. The FLAIR MRI was co-registered on the image computed by averaging the temporal points acquired before the contrast-agent bolus arrival. The transformation matrix obtained was then used to register the segmentation mask of the final ischemic lesion visible on the one-month follow-up FLAIR-MRI. After registration, the final lesion were rebinarized by applying a 50% threshold to correct for the eventual partial volume effects introduced during registration.

The study proposed here is a voxel-by-voxel study, and therefore, only a sub-group of voxels were selected from each patient in order to obtain a good repartition between infarcted and non-infarcted voxels in our study dataset. In accordance with the recommendations of Jonsdottir et al. (2009), the sub-set of voxels was selected in such a way as to have 50% of infarcted voxels versus 50% of non-infarcted voxels in our dataset, with amongst the non-infarcted voxels, 60% situated in the ipsi-lateral hemisphere of the brain (where the final lesion is present) and 40% situated in the contra-lateral hemisphere (hemisphere not affected by the ischemic stroke). More precisely on the location of these training voxels, all infarcted voxels were included, the voxel located in the contra-lateral were chosen randomly while the voxels in the ipsi-lateral were chose in the close vicinity (measure with morphomathematic distance) of the infarcted voxels. In the end, from the four patients considered here, we extracted a total of 22,105 voxels for our pilot study dataset. A visual abstract of the annotated data set considered for the supervised classification (infarcted versus non infarcted) addressed from perfusion MRI in this study is given in Fig. 1.

3. Methods

3.1. Encoding of the local spatio-temporal signature of each voxel into a patch

The new encoding proposed for the perfusion MRI is motivated by the fact that the spatio-temporal signature of each voxel is dif-

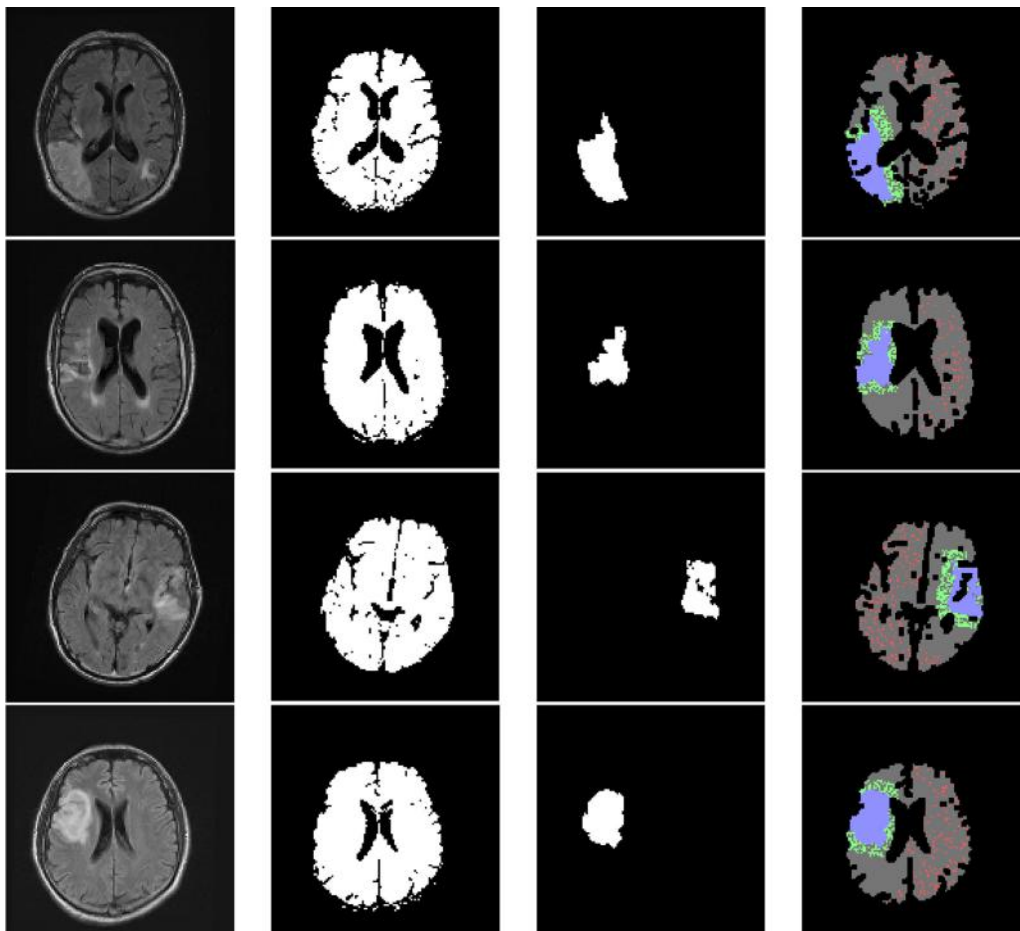


Fig. 1. Annotated data set considered for supervised classification of this study based on raw perfusion MRI. 1st line: patient 1 (slice $z = 12$), 2nd line: patient 2 (slice $z = 13$), 3rd line: patient 3 (slice $z = 10$), 4th line: patient 4 (slice $z = 14$). 1st column: FLAIR-MRI, 2nd column: brain masks, 3rd column: lesion masks obtained by the segmentation of the FLAIR-MRI, 4th column: voxel subsampling for classification - healthy pixels chosen shown in red (contra hemisphere) and green (ipsi hemisphere), and infarcted pixels shown in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

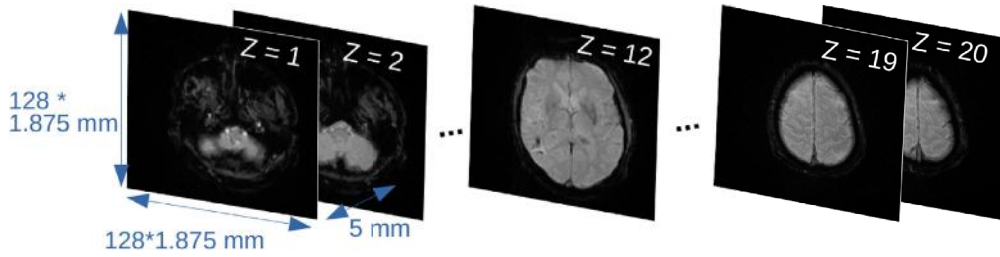
difficult to observe due to the 3D+ time nature of the data structure (See Fig. 2).

We propose to translate this spatio-temporal signature, into a discriminant texture easily identifiable by the human eye and useful to automatically characterize the state of the tissues in each voxel by simple texture analysis tools from computer vision. In order to do so, we propose to encode the information contained in the Moore neighborhood of order 1 of each voxel.

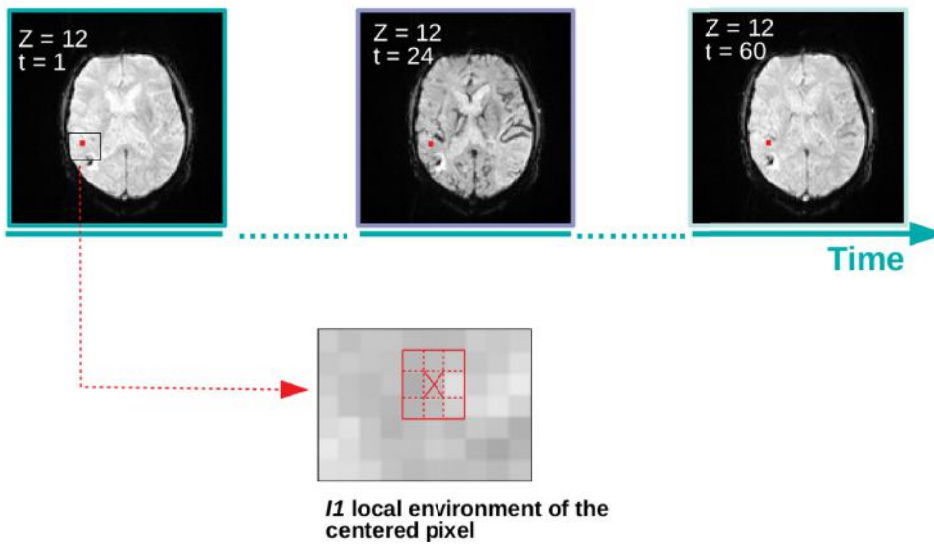
To do so, we unfold the temporal signals along a spatial dimension and then pile up, one on top of the other, the temporal signals of the 8 voxels in the Moore neighborhood of order 1 of each voxel of interest, creating thereby for each voxel a patch of size 9 by N_t , where N_t is the number of temporal acquisition points in the perfusion imaging sequence. This encoding method is illustrated in Fig. 3, with, in our case, $N_t = 60$. In order to take into account information from a larger neighborhood, it would be possible to consider larger Moore neighborhood orders. This would however decrease the amount of available voxels in the brain due to border effect. Instead, as introduced in Giacalone et al. (2017), we propose to add a preliminary image treatment, a time point by time point smoothing step with a mean filter of size $N_p \times N_p$. N_p can be seen as the observation scale for the neighborhood. A given voxel v_i and a given observation scale N_p are associated with a unique patch $I_{N_p}(v_i)$ generated from the perfusion image. This preliminary smoothing of the image allows to reduce noise in the patch, noise that might have created patterns which would not have been

relevant to our classification task and would have hindered the classification process. However, if the observation scale N_p is too large, we will lose the relevant local information and the classification precision might be negatively impacted. In this paper, we evaluate the impact of observation scale on the precision of tissue fate prediction and will compare the predictive potential of patches of various size N_S from $I_1, I_3, I_5, I_7, I_9, I_{11}$ to I_{13} .

Intuitively, we expected to obtain different patterns (or textures) on patches associated with voxels belonging to regions where the tissues end up as part of the final ischemic lesion and those associated with voxels in the rest of the brain. We will refer to these two groups of voxels as the pathological and healthy voxels. By looking at the patches obtained for these two groups of voxels, we notice a general tendency (see Fig. 4). The typical patch obtained for the healthy voxels exhibits a well-defined hypo-intensity segment of relatively narrow width and similar width for all the lines in the patch. As expected, this is indicative of a consistent behavior between neighboring voxels and a quick contrast-agent bolus passage. The typical patch obtained for the pathological voxels exhibits a hypo-intensity segment which is relatively spread out and not very contrasted with varying width for the different lines in the patch. As expected, this is indicative of a more erratic behavior in the pathological tissues, with a difficult passage of the contrast-agent bolus. However, are the textures obtained on these patches sufficiently discriminating to allow for tissue fate prediction? The question is to determine whether it is



((a)) Visualization of the spatial structure of PWI-MRI for patient 1. The acquisition was made on 20 spatial slices for 60 consecutive time steps of 1.5 second. 5 slices (taken at number $z=1$, $z=2$, $z=12$, $z=19$ and $z=20$) for one given time step ($t=1$) are shown.



((b)) Visualization of the temporal structure of PWI-MRI for patient 1. The acquisition was made on 20 spatial slices for 60 consecutive time steps. 3 time steps ($t=1$, $t=24$ and $t=60$) for one given slice ($z=12$) are shown. The local environment of one pixel with a neighborhood of 1 pixel $I1$ are shown in red.

Fig. 2. Illustration of the spatio-temporal signature of the raw MRI signals in perfusion MRI.

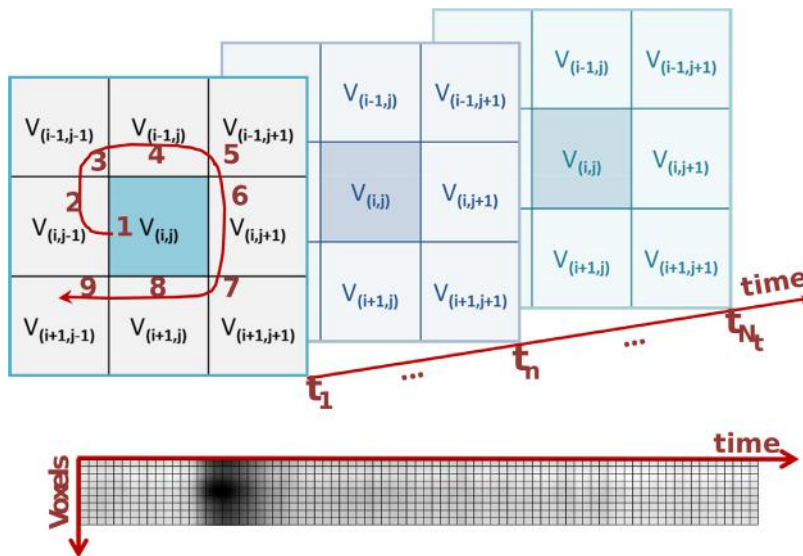


Fig. 3. Encoding of the spatio-temporal signature of perfusion MRI signals as a patch.

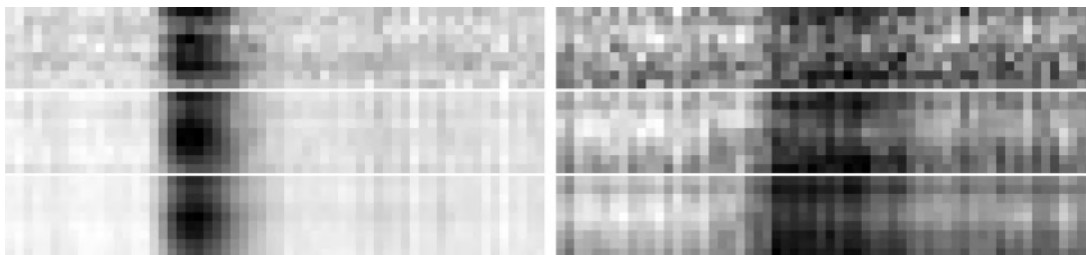


Fig. 4. Illustration of the typical patterns obtained for patches of type I_1 , I_3 and I_5 (from top to bottom) for healthy voxels (left) and pathological voxels (right).

possible to classify the voxels depending on their final state based on the pattern visible on the encoded patches. In order to do so, we propose here to use classical texture analysis and classification tools.

3.2. Local binary patterns as a texture descriptor

Since its introduction in Ojala et al. (2002), LBP has been extensively reported in the literature for image classification (Nanni et al., 2012). Notably, LBP have been shown to be valuable for medical image analysis (Nanni et al., 2010). For example, LBP have been used to identify malignant breast cells (Oliver et al., 2007) or to find relevant brain slices in magnetic resonance volumes (Unay and Ekin, 2008). They have also been used as texture features extracted from thyroid slices (Keramidas et al., 2008) and were investigated in automated cell phenotype image classification (Nanni and Lumini, 2008). More recently, algorithms using LBP texture features and support vector machine classifiers were proposed for the classification of medical images from saliency-based folded data (Camlica et al., 2015) and for the classification of skin cancer images to melanoma and non melanoma cases (Adjed et al., 2016) and for macula pathologies (Liu et al., 2011) or the breast cancer (Wan et al., 2017).

An important property of the LBP is notably its robustness to image mean intensity variations which can be caused, for example, by illumination variations in classical images. An other attractive property is their computational simplicity which renders possible their use in real-time image analysis applications. In our case, the LBP properties will present a great interest. First, the raw perfusion MRI signals are not quantitative and brightness invariance will therefore be necessary. Moreover, if the LBP are to be used eventually for patient management during clinical routine, where time is of the essence, a low computational cost will also be a highly desirable property. In this pilot study, we therefore propose to use LBP to describe the patterns observed on our patches.

LBP correspond to texture descriptors which describe the local environment of each voxel by a label computed by simple thresholding of the gray-level values of its neighboring voxels. The basic idea behind LBP is that an image is composed of micropatterns and that a histogram of these micropatterns contains information about the distribution of edges and other local features in the image. The conventional LBP operator with neighborhood (L, R) (Ojala et al., 2002) is computed at each voxel location by considering both the value of the voxel under consideration (q_c) and the values of the L voxels in the circular neighborhood of radius R around the voxel under consideration (q_l with $l \in \{1 \dots L\}$). Formally, the LBP operator with neighborhood (L, R) is defined as

$$LBP_{(L,R)} = \sum_{l=1}^L s(q_l - q_c) 2^{l-1}, \quad (1)$$

where $s(x) = 1$ if $x \geq 0$ and 0 otherwise. There are therefore 2^L distinct labels resulting from the different possible circular patterns around each voxel. Two types of patterns can be distinguished

in LBP: the uniform patterns, which have at most two transitions from $s(q_l - q_c) = 0$ to $s(q_{l+1} - q_c) = 1$ (or reversely), and the non-uniform patterns. Ojala et al. (2002) have observed that the uniform patterns constitute the majority of the patterns that can be observed in textured images. For example, they constitute slightly less than 90% of the patterns when using a $(8,1)$ neighborhood in textured images. This knowledge can therefore be used to reduce the number of possible labels by using a distinct label for all of the uniform patterns but by using only one label for all the non-uniform patterns. For example, if we consider a $(8,R)$ neighborhood, there is a total of $2^8 = 256$ possible patterns, only 58 of which are uniform patterns, and we can therefore reduce the number of possible labels from 256 to 59. Here, we propose to use LBP with a $(8,1)$ neighborhood (i.e. Moore neighborhood) and an encoding with uniform patterns, resulting in a total of 59 possible labels (see Fig. 5).

Once the LBP operator has been applied to a patch, the concatenated histograms of the sub-patches separating the labeled patch into contiguous segments can then be used as a feature vector to describe the texture in the initial patch. An important quality parameter of the LBP as a texture descriptor is the dimension of the sub-patches on which the histograms are being computed. Here, we propose to use as a feature vector for each voxel the concatenation of the histograms on the adjacent sub-patches of width N_h , separating each patch into N_t/N_h contiguous segments of equal width. Since we proposed to work with uniform patterns and a $(8,1)$ neighborhood, each histogram contains 59 values and we therefore obtain a feature vector of total length $N_t/N_h \times 59$. In this paper, we evaluate the impact of the width N_h on the precision of tissue fate prediction and will compare the predictive potential of feature vectors calculated by using segments of width 3, 4, 5, 6, 10, 12, 15, 20, 30 and 60 voxels respectively.

We now wish to use these feature vectors to classify each voxel depending on its final state status (class 1 for pathological voxels belonging to the final lesion or class 0 for healthy voxels). In order to do so, we propose to use a support vector machine classifier for the supervised learning of our tissue fate classification model.

3.3. Classification with a support vector machine classifier

The Support Vector Machine (SVM) method consists in finding the separating hyper-plane allowing to separate at best, within the input variable space, the data points in the training set depending on their final state status (class 0 or 1). A new point is then classified as belonging to class 0 or 1 depending on its position with respect to the separating hyper-plane. The farther the point is from the separating hyper-plane and the highest the confidence is concerning the class assigned to this point by the SVM classifier. The distance between the hyper-plane and the training points closest to it, points called support vectors, is defined as the margin. The separating hyper-plane selected with the SVM method corresponds to the hyper-plane resulting in the largest margin possible. The support vectors are therefore the only points used for the

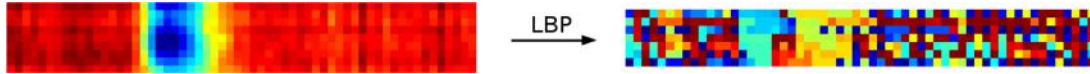


Fig. 5. Illustration of the LBP labels obtained (right) from a given patch (left).

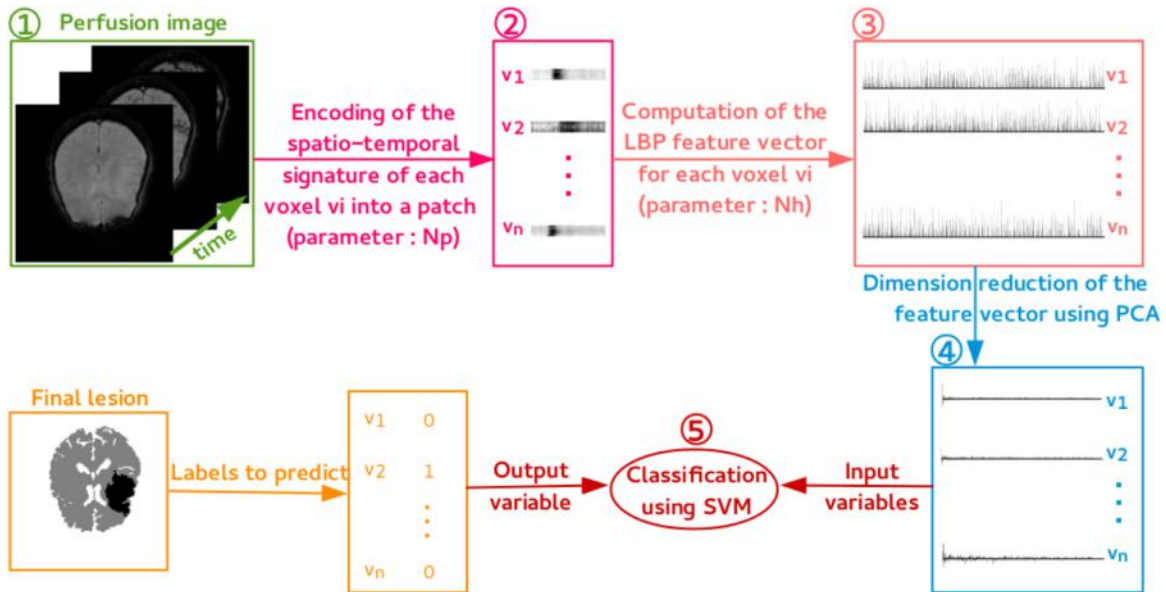


Fig. 6. Visual summary of the proposed classification approach.

optimization of the margin and the definition of the optimum separating hyper-plane, hence the name of this machine learning method. This method works in cases where the points in the two different classes are linearly separable. In numerous situations, the two classes are not linearly separable and, in this case, it is possible to transform the input variables into a new variable space, of higher dimension than the native space and in which the points are rendered linearly separable. In practice, the passage to another feature space is done via the use of a kernel function, and is therefore referred to as the kernel trick.

Here, after empirical testing, we decided to use an SVM classifier with a degree-3 polynomial kernel, taking as input variables the elements of the LBP feature vector of each voxel. The size of the feature vectors obtained is quite large, therefore, in order to reduce the dimension of the input vector, we proceed beforehand to a principal component analysis (PCA) and preserve only the first principal vectors which allow to explain 90% of the total variance contained in the training dataset. A diagram summarizing the classification approach proposed here is given in Fig. 6. To account for the variability in the quality of the classifications obtained, we used in this section a K-fold cross-validation technique which allows to assess how well the SVM classifier might generalize to an independent data set for tissue fate prediction. We divided our data set into 100 subsets of voxels and, for each possible combination, we used 99 of the subsets for the training of the SVM classifier (99% of the data, i.e. 21,884 training points) and 1 subset for testing the quality of the classification model obtained (1% of the data, i.e. 221 testing points).

4. Results

4.1. Segment width optimization

We show in Table 1 the influence of the segment width N_h used for the computation of the feature vectors on the performance of the classification. Performances are expressed in terms of the

minimum, mean, maximum and standard deviation σ of the accuracy, sensitivity and specificity obtained during cross-validation. Also, N_f , the number of components kept after the PCA to explain 90% of the total variance, is provided as a measure of the complexity of the model. As can be seen, the segment width N_h used for feature vector computation greatly influences the quality of the classification obtained. It appears from the analysis of Table 1, that decreasing the segment width increases the performances obtained, resulting in up to a 15% difference in accuracy between a segment width of $N_h = 60$ and $N_h = 3$. The contrast-agent bolus observed on our data in healthy tissues has a width of roughly 10 voxels. The optimum segment width found here, $N_h = 3$, is therefore approximately a third of the width of the contrast-agent bolus, which seems reasonable. The patterns to be encoded by LBP as shown in Fig. 3 can be seen as the concatenation of three sub-patterns: a darker one corresponding to the bolus in the middle of two brighter ones which mainly corresponds to noise. Too large a N_h would mix together local variations of these subpatterns and produce as shown in Table 1 a decrease of the prediction performances. The discriminant information lays in the relative sizes of these sub-patterns. Decreasing the size of the segment N_h comes with an increase of the features selected N_f in the prediction model and an increasing risk of overfitting.

4.1.1. Observation scale optimization

We now fix the segment width to the optimal value found in the previous section, $N_h = 3$, and investigate the influence of the observation scale N_p used for the patch computation. The results are provided in Table 2. It appears that the classification performance increases and that its variability decreases as the observation scale increases up to $N_p = 9$. For observation scales larger than 9, the performance seems to reach a plateau and its increase appears negligible compared to the variability of the results obtained. The optimization of the observation scale from $N_p = 1$ to $N_p = 9$ allows to improve the accuracy of another 13%. Interestingly, these results are in accordance with the

Table 1

Accuracy (%), Sensitivity (%), Specificity (%) obtained on the 100 test data sets during cross-validation when patches of type I_1 (i.e.: observation scale $N_p = 1$) are used for classification and different size N_h of segment width are evaluated for feature vector computation. N_f corresponds to the number of features selected after dimension reduction via PCA.

N_h	Accuracy (%)				Sensitivity (%)				Specificity (%)				N_f
	Min	Mean	Max	σ	Min	Mean	Max	σ	Min	Mean	Max	σ	
60	58.82	66.46	75.11	3.34	49.53	61.59	71.79	4.79	59.83	71.34	82.61	4.59	22
30	61.09	68.81	76.47	3.27	55.00	68.04	78.38	4.79	59.26	69.55	79.65	4.45	50
20	58.37	69.51	78.73	3.39	55.14	68.71	80.91	5.22	60.87	70.33	79.63	4.41	79
15	65.61	72.07	78.73	2.83	60.95	71.48	80.85	4.28	60.55	72.69	81.42	3.91	111
12	68.33	74.40	81.00	2.84	64.22	74.20	84.11	4.52	65.77	74.59	81.90	3.54	143
10	68.78	74.78	82.35	3.14	65.09	74.17	84.68	4.58	63.54	75.43	84.91	4.12	176
6	71.49	78.09	85.97	2.72	66.38	77.85	88.79	4.25	69.37	78.33	87.61	3.70	309
5	71.49	79.05	85.52	2.59	68.97	79.07	87.07	3.95	69.72	79.01	88.50	3.74	379
4	72.40	80.21	85.97	2.73	70.69	80.03	87.27	4.07	70.87	80.37	88.50	3.64	473
3	75.11	81.81	88.24	2.71	72.41	81.43	89.81	3.71	73.33	82.16	91.15	3.56	659

Table 2

Accuracy (%), Sensitivity (%), Specificity (%) obtained on the 100 test data sets during cross-validation when the segment width used for feature vector computation is fixed to $N_h = 3$ and different patch types I_{N_p} are evaluated. N_f corresponds to the number of features selected after dimension reduction via PCA.

Patch type	Accuracy (%)				Sensitivity (%)				Specificity (%)				N_f
	Min	Mean	Max	σ	Min	Mean	Max	σ	Min	Mean	Max	σ	
I_1	75.11	81.81	88.24	2.71	72.41	81.43	89.81	3.71	73.33	82.16	91.15	3.56	659
I_3	85.07	89.85	95.02	2.11	81.65	89.62	96.58	3.26	81.82	90.07	95.10	2.85	607
I_5	89.14	93.43	96.38	1.63	85.71	93.65	98.28	2.41	88.46	93.20	97.64	2.09	545
I_7	90.05	94.73	98.19	1.46	88.99	95.23	99.13	1.84	85.96	94.22	99.12	2.21	494
I_9	92.31	95.04	98.19	1.28	90.76	95.38	100.00	1.80	89.81	94.70	100.00	1.92	454
I_{11}	91.40	95.26	98.64	1.50	89.92	95.77	100.00	2.07	87.85	94.72	99.09	2.32	420
I_{13}	91.86	95.37	98.19	1.39	88.24	96.02	100.00	1.83	90.18	94.70	98.28	2.00	391

spatial scales found in the work of Scalzo et al. (2012) and Giacalone et al. (2017) pointing to the interest of a “regional” approach to predict tissue fate in stroke

4.2. Impact of perfusion MRI post-processing

Results obtained so far have been obtained on raw dynamic susceptibility images considered without any other post processing than the registration described in the Material section. This means without the usual deconvolution step by the arterial input function. In this last experiment, we compare the classification results obtained when using, as originally proposed in this article, raw perfusion images for the patch computation and those obtained when using, conventionally, images which underwent the usual post-processing treatments applied in quantitative perfusion MRI.² For this experiment a standard temporal Tikhonov deconvolution (Calamante et al., 2003) was used for comparison with the raw perfusion signals.

The results obtained using the optimal segment width N_h and observation scale N_p found in the previous experiments are given in Table 3. It appears that the performances in terms of accuracy, specificity and sensitivity are very close to each other. The potential gain in using post-processed images is possibly located in the size of the model, which is smaller (e.g. 350 features for the deconvolved data vs 454 for the raw perfusion data). This results are also illustrated in Fig. 7.

² The usual post-processing treatments applied in quantitative perfusion MRI consist in estimating the contrast-agent concentration signals from the raw perfusion signals, under the assumption of a linear relationship between the change in relaxation rate and the concentration, and then in proceeding to a temporal deconvolution of the concentration signals by the arterial input function (Calamante et al., 2003).

4.3. Sensitivity analysis

In the previous subsection we have detailed the performance of the classification scheme of Fig. 6 while optimizing the spatial and temporal scales of the spatio-temporal representation of the perfusion MRI. Each step of this classification scheme could also of course be optimized. In this section, we investigate some of these variants while considering as reference the setup and associated performance obtained in Tables 1 and 2 and keeping as common point the spatio-temporal encoding which is the core of the proposal of this article.

First, several variants of the LBP are tested. This includes the so-called local ternary patterns (LTP) (Tan and Triggs, 2010), and the Discriminative Features of LBP Variants (disCLBP) (Guo et al., 2012). Second, several variants of the dimension reduction technique have been tested with Locally Linear Embedding (LLE) (Roweis and Saul, 2000) and T-SNE (Maaten and Hinton, 2008) for comparison with PCA. In step 4 of Fig. 6, we used classical principal component analysis to reduce the dimension of the feature space. The dimension size of feature space is chosen fixed to have a fair comparison of all three classification approaches on our data. In order to estimate, an intrinsic dimension of the feature space, the Maximum Likelihood Estimates (MLE) algorithm of Levina and Bickel (2005) is used and estimated an intrinsic dimension of 69 for the feature space that we consider for N_f in the following experiment with all three dimensional reduction approaches tested. Third, several variants of the classifier have been tested. Since the data set considered here is relatively small for a machine learning approach, we selected classifier which are known to operate robustly with small data sets. This includes soft margin SVM, and Random Forest (RF) (Breiman, 2001), which are all tested with cross-validation approach with different number of folds. Concerning the RF classifier, the following parameters are considered to control depth of trees, maximum 100 decision splits, minimum 1

Table 3

Accuracy (%), Sensitivity (%), Specificity (%) obtained on the 100 test data sets during cross-validation when the segment width is fixed to $N_h = 3$ for feature vector computations and the observation scale is fixed to $N_p = 9$ for the patch computation. N_f corresponds to the number of features selected after dimension reduction via PCA. These results compare the quality of the prediction obtained if we use the raw perfusion signals (PWI) as input for our classification method or if we consider the contrast-agent concentration image (CAC) or the concentration image deconvolved with Tikhonov algorithm (Calamante et al., 2003) (IRF).

	Accuracy (%)				Sensitivity (%)				Specificity (%)				N_f
	Min	Mean	Max	σ	Min	Mean	Max	σ	Min	Mean	Max	σ	
PWI	92.31	95.04	98.19	1.28	90.76	95.38	100.00	1.80	89.81	94.70	100.00	1.92	454
CAC	90.50	94.59	99.10	1.49	89.29	95.18	99.10	1.99	88.89	94.03	99.17	2.02	365
IRF	90.05	93.94	97.74	1.56	89.62	94.05	98.21	1.87	87.85	93.87	99.15	2.38	350

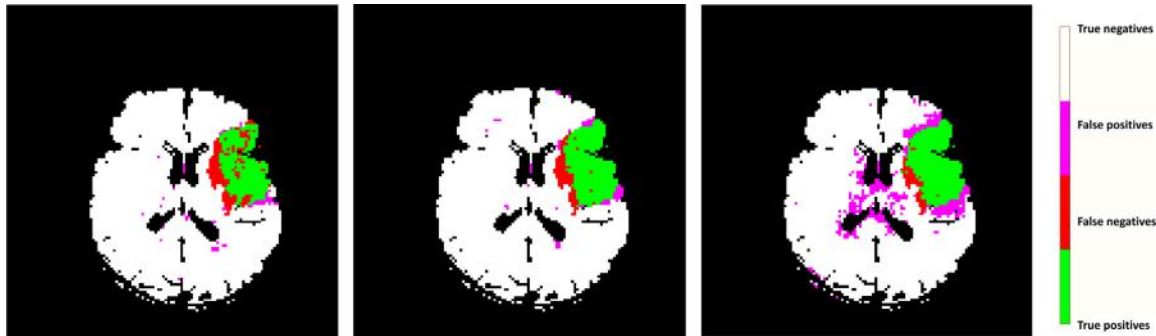


Fig. 7. Illustration of the classification performances obtained for one patient slice when using a segment width of $N_h = 3$ and an observation scale of $N_p = 9$ and (left) the raw perfusion signals, (middle) the concentration signals or (right) the concentration signals after deconvolution. The classification model used here was trained on all the voxels (from the four different patients) which did not belong to this specific slice.

Table 4

Accuracy (%) obtained by LTP where PCA, T-SNE and LLE are used as dimensional reduction functions.

Classifiers	DR	Number of folds in cross-validation									
		2	3	4	5	6	7	8	9	10	
RF	PCA	79.84	79.72	79.60	79.12	79.28	79.21	79.62	79.50	79.38	
	TSNE	83.99	84.74	84.43	84.60	84.55	84.75	84.60	84.53	84.58	
	LLE	83.78	84.30	84.43	84.67	84.59	84.75	84.76	84.59	84.64	
Soft margin SVM	PCA	92.52	94.34	94.78	95.17	95.41	95.39	95.81	95.69	95.77	
	TSNE	86.81	87.32	87.44	87.69	87.60	87.56	87.68	87.73	87.68	
	LLE	86.32	86.92	87.04	87.18	87.17	87.16	87.17	87.28	87.27	
Hard margin SVM	PCA	92.41	94.24	95.01	95.34	95.42	95.67	95.62	95.84	95.69	
	TSNE	90.09	89.96	90.31	90.31	90.17	90.28	90.43	90.60	90.38	
	LLE	88.55	88.83	89.19	89.49	89.32	89.36	89.29	89.48	89.53	

Table 5

Accuracy (%) obtained by disCLBP where PCA, T-SNE and LLE are used as dimensional reduction functions.

Classifiers	DR	Number of folds in cross-validation									
		2	3	4	5	6	7	8	9	10	
RF	PCA	79.68	80.42	80.33	80.10	80.65	80.22	80.36	80.64	80.31	
	TSNE	85.95	85.49	85.73	85.84	85.12	85.23	85.36	85.29	85.08	
	LLE	85.13	84.95	84.94	85.11	84.97	85.18	85.09	85.16	84.89	
Soft margin SVM	PCA	69.96	71.40	71.97	72.26	72.62	72.53	72.83	73.07	73.14	
	TSNE	88.93	88.80	89.04	89.01	89.01	89.44	89.09	89.13	89.10	
	LLE	85.33	86.03	86.17	86.28	86.41	86.43	86.51	86.48	86.54	
Hard margin SVM	PCA	69.92	71.11	72.03	72.41	72.71	72.73	72.90	73.01	73.21	
	TSNE	85.63	85.89	86.01	86.10	86.29	86.17	86.24	86.17	86.26	
	LLE	85.10	85.74	86.09	86.45	86.34	86.41	86.33	86.49	86.43	

leaf node observation, and Gini's diversity index as split criterion. The results are presented in Tables 4–6.

The obtained results in Tables 4–6 show that variants of the pipeline presented in the result section of this manuscript also obtain good performances. Possibilities of improvement from the performance obtained in the main result section of the manuscript are mainly found in Table 4 with the LTP method. A full optimization combining the best elements in the global pipeline of Fig. 6 is out of the scope of this article. The important point we want to stress here is that all these variants are based on the same spatio-

temporal encoding of the raw perfusion data. This further demonstrates the interest of this encoding.

5. Discussion

The result presented in Fig. 7 for the prediction on raw perfusion signals are intrinsically interesting since they demonstrate the possibility to perform prediction of similar quality without deconvolution. Also, errors found in Fig. 7 with our method are also intrinsically good errors since the badly classified voxels are not

Table 6
Accuracy (%) obtained by LBP where PCA, T-SNE and LLE are used as dimensional reduction functions.

Classifiers	DR	Number of folds in cross-validation								
		2	3	4	5	6	7	8	9	10
RF	PCA	80.64	81.40	81.47	81.69	81.29	81.05	81.40	81.14	81.58
	TSNE	85.85	85.28	85.25	85.26	85.65	85.33	85.05	85.36	85.40
	LLE	84.14	84.61	84.51	84.34	84.54	85.15	84.73	84.58	84.11
Soft margin SVM	PCA	71.10	73.32	74.54	75.11	75.49	75.37	75.58	75.66	75.85
	TSNE	85.94	86.50	86.63	86.73	86.61	86.66	86.65	86.71	86.69
	LLE	86.32	86.92	87.04	87.18	87.17	87.16	87.17	87.28	87.27
Hard margin SVM	PCA	72.03	73.67	73.97	75.22	75.43	75.68	75.68	75.54	75.68
	TSNE	88.79	89.54	89.60	89.29	89.79	89.65	89.27	89.05	89.07
	LLE	86.64	86.83	87.15	87.20	87.32	87.32	87.29	87.14	87.17

randomly positioned but precisely located on the frontier between infarcted and non infarcted regions, i.e. where the decision making is the most challenging.

Beside the intrinsic value of our predictor based on the proposed spatio-temporal encoding one may also wonder about the performances obtained with our prediction scheme in relation with the existing literature. In most recent studies using machine learning-based approaches in the framework of the ISLE challenge (Maier et al., 2017) or outside this challenge, for instance on the use of deep learning with neural network (Huang et al., 2010; Stier et al., 2015; Nielsen et al., 2018) best performances of between 85 and 95% accuracy are found.

It is important however to notice that such performances were obtained here while incorporating perfusion images and diffusion images. Diffusion in the acute stage of the stroke is known to be a highly predictive imaging for the final stroke lesion (Giacalone et al., 2017). In the small cohort considered here the lesions were quite compact lesions with a good similarity between the early perfusion lesion and the diffusion lesion. This may explain why we obtain prediction result similar to the one of the literature while working only on perfusion images. Larger dataset with patients having larger differences between perfusion and diffusion would be interesting to be incorporated to assess in such cases the predictive value of perfusion alone and the gain brought jointly using diffusion and perfusion along variant of the proposed encoding scheme.

Other points could be discussed. For instance local binary patterns have been used here to analyze the local spatio-temporal patches seen as textures. This encoding benefits from an interesting property of invariance to a gray level baseline variation. This is important for texture characterization in temporal MRI sequences where this baseline can indeed vary. Local binary patterns, under the form used in this article (Ojala et al., 2002), are not shift invariant. Such shift would happen in practice in multicentric studies where the trigger of the acquisition with the injection of the contrast agent is not normalized. However the typical duration of the bolus is an invariant parameter and it would be possible to normalize the local patch to be encoded with a raw detection of the peak of contrast. Other variants of the local binary pattern with additional properties of invariance could be tested (Nanni et al., 2010). Also, other texture approaches, possibly with native baseline and shift invariance properties, could be used (Mirmehdi, 2008). The patches temporal encoding proposed in this article could also be seen and analyzed as transient noisy patterns with wavelet-like approaches (Mallat, 2008). We are currently considering all these perspectives with larger datasets.

6. Conclusion

In this pilot study, we have proposed a new approach to encode the spatio-temporal signature of each voxel in raw perfusion data. We have then proposed a scheme based on local binary pat-

terns coupled with a support vector machine classifier to realize a supervised classification of pathological and healthy voxels in stroke. This approach provides promising results, with a precision of classification of 95% on average on the small data set evaluated here. This is promising indeed since a large part of the literature on stroke focuses on developing post-processing treatments (Willats and Calamante, 2013b) to denoise images while the classification here was realized from raw perfusion data only.

Acknowledgements

This work was performed within the framework of the Labex PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program “Investissements d’Avenir” (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

- Adjed, F., Faye, I., Ababsa, F., Gardezi, S.J., Dass, S.C., 2016. Classification of skin cancer images using local binary pattern and SVM classifier. In: AIP Conference Proceedings, 1787. AIP Publishing, p. 080006.
- Albers, G.W., Marks, M.P., Kemp, S., Christensen, S., Tsai, J.P., Ortega-Gutierrez, S., McTaggart, R.A., Torbey, M.T., Kim-Tenser, M., Leslie-Mazwi, T., et al., 2018. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N. Engl. J. Med.* 378 (8), 708–718.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32.
- Calamante, F., Gadian, D.G., Connelly, A., 2003. Quantification of bolus-tracking MRI: improved characterization of the tissue residue function using Tikhonov regularization. *Magn. Reson. Med.* 50 (6), 1237–1247.
- Camlica, Z., Tizhoosh, H.R., Khalvati, F., 2015. Medical image classification via SVM using LBP features from saliency-based folded data. In: Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, pp. 128–132.
- Davis, S., Fisher, M., Warach, S., 2003. *Magnetic Resonance Imaging in Stroke*. Cambridge University Press.
- Feigin, V.L., Forouzanfar, M.H., Krishnamurthi, R., Mensah, G.A., Connor, M., Bennett, D.A., Moran, A.E., Sacco, R.L., Anderson, L., Truelsen, T., et al., 2014. Global and regional burden of stroke during 1990–2010: findings from the global burden of disease study 2010. *Lancet* 383 (9913), 245–255.
- Frindel, C., Robini, M.C., Rousseau, D., 2014. A 3-D spatio-temporal deconvolution approach for MR perfusion in the brain. *Med. Image Anal.* 18 (1), 144–160.
- Frindel, C., Rousseau, D., Cho, T., Berthezène, Y., Wiart, M., Nighoghossian, N., 2012. Application d’une mesure de similarité locale pour la segmentation du système ventriculaire cérébral en irm de perfusion. 1er congrès de la Société Française de Résonance Magnétique en Biologie et Médecine. France
- Giacalone, M., Frindel, C., Grenier, E., Rousseau, D., 2017. Multicomponent and longitudinal imaging seen as a communication channel—an application to stroke. *Entropy* 19 (5), 187.
- Goyal, M., Menon, B.K., Van Zwam, W.H., Dippel, D.W., Mitchell, P.J., Demchuk, A.M., Dávalos, A., Majoie, C.B., van der Lugt, A., De Miquel, M.A., et al., 2016. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *The Lancet* 387 (10029), 1723–1731.
- Guo, Y., Zhao, G., Pietikäläinen, M., 2012. Discriminative features for texture description. *Pattern Recognit.* 45 (10), 3834–3843.
- He, L., Orten, B., Do, S., Karl, W.C., Kambadakone, A., Sahani, D.V., Pien, H., 2010. A spatio-temporal deconvolution method to improve perfusion CT quantification. *Med. Imaging, IEEE Trans.* 29 (5), 1182–1191.
- Hermitte, L., Cho, T., Ozanne, B., Nighoghossian, N., Mikkelsen, I.K., Ribe, L., Baron, J., Østergaard, L., Derex, L., Hjort, N., Fiehler, J., Pedraza, S., Hermier, M., Maucort-Boulch, D., Berthezène, Y., 2013. Very low cerebral blood volume predicts parenchymal hematoma in acute ischemic stroke. *Stroke* 44, 2318–2320.

- Ho, K.C., El-Saden, S., Scalzo, F., Bui, A.A., Arnold, C.W., 2016. Abstract WP41: predicting acute ischemic stroke tissue fate using deep learning on source perfusion MRI. *Stroke* 47 (Suppl 1). AWP41–AWP41.
- Huang, S., Shen, Q., Duong, T.Q., 2010. Artificial neural network prediction of ischemic tissue fate in acute stroke imaging. *J. Cerebral Blood Flow Metabolism* 30 (9), 1661–1670.
- Jonsdottir, K.Y., Østergaard, L., Mouridsen, K., 2009. Predicting tissue outcome from acute stroke magnetic resonance imaging. *Stroke* 40 (9), 3006–3011.
- Keramidas, E.G., Iakovidis, D.K., Maroulis, D., Dimitropoulos, N., 2008. Thyroid texture representation via noise resistant image features. In: *Computer-Based Medical Systems*, 2008. CBMS'08. 21st IEEE International Symposium on. IEEE, pp. 560–565.
- Kim, S.J., Son, J.P., Ryoo, S., Lee, M.-J., Cha, J., Kim, K.H., Kim, G.-M., Chung, C.-S., Lee, K.H., Jeon, P., et al., 2014. A novel magnetic resonance imaging approach to collateral flow imaging in ischemic stroke. *Ann. Neurol.* 76 (3), 356–369.
- Levina, E., Bickel, P.J., 2005. Maximum likelihood estimation of intrinsic dimension. In: *Advances in neural information processing systems*, pp. 777–784.
- Liu, Y.-Y., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J.S., Rehg, J.M., 2011. Automated macular pathology diagnosis in retinal oct images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. *Med. Image Anal.* 15 (5), 748–759.
- Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. *J. Mach. Learn. Res.* 9 (Nov), 2579–2605.
- Maier, O., Menze, B.H., von der Gabelntz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med. Image Anal.* 35, 250–269.
- Mallat, S., 2008. *A Wavelet Tour of Signal Processing: The Sparse way*. Academic press.
- Meijs, M., Christensen, S., Lansberg, M.G., Albers, G.W., Calamante, F., 2015. Analysis of perfusion MRI in stroke: to deconvolve, or not to deconvolve. *Magn. Reson. Med.*
- Mirmehdi, M., 2008. *Handbook of texture analysis*. Imperial College Press.
- Murray, C.J., Barber, R.M., Foreman, K.J., Ozgoren, A.A., Abd-Allah, F., Abera, S.F., Aboyans, V., Abraham, J.P., Abubakar, I., Abu-Raddad, L.J., et al., 2015. Global, regional, and national disability-adjusted life years (dalys) for 306 diseases and injuries and healthy life expectancy (hale) for 188 countries, 1990–2013: quantifying the epidemiological transition. *Lancet* 386 (10009), 2145–2191.
- Nanni, L., Lumini, A., 2008. A reliable method for cell phenotype image classification. *Artif. Intell. Med.* 43 (2), 87–97.
- Nanni, L., Lumini, A., Brahnham, S., 2010. Local binary patterns variants as texture descriptors for medical image analysis. *Artif. Intell. Med.* 49 (2), 117–125.
- Nanni, L., Lumini, A., Brahnham, S., 2012. Survey on LBP based texture descriptors for image classification. *Expert Syst. Appl.* 39 (3), 3634–3641.
- Nielsen, A., Hansen, M.B., Tietze, A., Mouridsen, K., 2018. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. *Stroke* STROKEAHA–117.
- Nogueira, R.G., Jadhav, A.P., Haussen, D.C., Bonafe, A., Budzik, R.F., Bhuvu, P., Yavagal, D.R., Ribo, M., Cognard, C., Hanel, R.A., et al., 2018. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N. Engl. J. Med.* 378 (1), 11–21.
- Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (7), 971–987.
- Oliver, A., Lladó, X., Freixenet, J., Martí, J., 2007. False positive reduction in mammographic mass detection using local binary patterns. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pp. 286–293.
- Perthen, J.E., Calamante, F., Gadian, D.G., Connelly, A., 2002. Is quantification of bolus tracking MRI reliable without deconvolution? *Magn. Reson. Med.* 47 (1), 61–67.
- Rekik, I., Allasounnière, S., Carpenter, T., Wardlaw, J., 2012. Medical image analysis methods in mr/ct-imaged acute-subacute ischemic stroke lesion: segmentation, prediction and insights into dynamic evolution simulation models. a critical appraisal. *NeuroImage* 1 (1), 164–178.
- Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290 (5500), 2323–2326.
- Scalzo, F., Hao, Q., Alger, J.R., Hu, X., Liebeskind, D.S., 2012. Regional prediction of tissue fate in acute ischemic stroke. *Ann. Biomed. Eng.* 40 (10), 2177–2187.
- Schmid, V.J., 2011. Voxel-based adaptive spatio-temporal modelling of perfusion cardiovascular MRI. *IEEE Trans. Med. Imaging* 30 (7), 1305–1313.
- Stier, N., Vincent, N., Liebeskind, D., Scalzo, F., 2015. Deep learning of tissue fate features in acute ischemic stroke. In: *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on. IEEE*, pp. 1316–1321.
- Tan, X., Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* 19 (6), 1635–1650.
- Towfighi, A., Saver, J.L., 2011. Stroke declines from third to fourth leading cause of death in the united states: historical perspective and challenges ahead. *Stroke* 42 (8), 2351–2355.
- Unay, D., Ekin, A., 2008. Intensity versus texture for medical image search and retrieval. In: *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on. IEEE*, pp. 241–244.
- Wan, S., Lee, H.-C., Huang, X., Xu, T., Xu, T., Zeng, X., Zhang, Z., Sheikine, Y., Connolly, J.L., Fujimoto, J.C., et al., 2017. Integrated local binary pattern texture features for classification of breast tissue imaged by optical coherence microscopy. *Med. Image Anal.* 38, 104–116.
- Willats, L., Calamante, F., 2013a. The 39 steps: evading error and deciphering the secrets for accurate dynamic susceptibility contrast mri. *NMR Biomed.* 26 (8), 913–931.
- Willats, L., Calamante, F., 2013b. The 39 steps: evading error and deciphering the secrets for accurate dynamic susceptibility contrast MRI. *NMR Biomed.* 26 (8), 913–931.
- Wintermark, M., Albers, G.W., Broderick, J.P., Demchuk, A.M., Fiebach, J.B., Fiehler, J., Grotta, J.C., Houser, G., Jovin, T.G., Lees, K.R., et al., 2013. Acute stroke imaging research roadmap ii. *Stroke*, STROKEAHA–113.

OPEN

Machine Learning-Based Classification of the Health State of Mice Colon in Cancer Study from Confocal Laser Endomicroscopy

Pejman Rasti¹, Christian Wolf², Hugo Dorez³, Raphael Sablong³, Driffa Moussata³, Salma Samiei¹ & David Rousseau^{1*}

In this article, we address the problem of the classification of the health state of the colon's wall of mice, possibly injured by cancer with machine learning approaches. This problem is essential for translational research on cancer and is a priori challenging since the amount of data is usually limited in all preclinical studies for practical and ethical reasons. Three states considered including cancer, health, and inflammatory on tissues. Fully automated machine learning-based methods are proposed, including deep learning, transfer learning, and shallow learning with SVM. These methods addressed different training strategies corresponding to clinical questions such as the automatic clinical state prediction on unseen data using a pre-trained model, or in an alternative setting, real-time estimation of the clinical state of individual tissue samples during the examination. Experimental results show the best performance of 99.93% correct recognition rate obtained for the second strategy as well as the performance of 98.49% which were achieved for the more difficult first case.

Classically the characterization of colon's pathology is realized from histology¹ but is now also investigated with *in vivo* imaging techniques which enable the oncological² early detection of abnormal physiological processes such as inflammation of dysplastic lesions. This includes chromoendoscopy³, confocal laser endomicroscopy^{4,5} or multiphoton microscopy⁶. These modern video-microscopies introduced in preclinical studies on mice with the promises of translational research⁷.

These imaging techniques are producing videos which for the inspection of one colon of one mouse corresponds to thousands of frames to be further multiplied by the number of mice inspected. Each frame of these videos can be different in the structure and texture as it is recorded over a colon's wall with movement of the probe, spurious presence of unexpected items between probes and colon, variation of contrast agent concentration. To draw benefit from such imaging protocols, the bottleneck is thus the automation of the image analysis. In this article, we consider one of these protocols and propose a fully automated solution for the classification of colon wall images into healthy, inflammation and dysplastic tissues.

We work with the confocal endomicroscopy imaging protocol of⁵ for the classification of the health state of the colon's wall of mice. Since its introduction, this protocol has seen widespread usage in multiple research groups⁸⁻¹⁰. So far, image analysis for the classification of colon's wall health state with this protocol has been relatively limited. The existing literature is based on handcrafted features^{5,8-10}.

In this article, we go beyond the sole characterization (feature handcrafting) and, for the first time on Mice colon in cancer study from confocal laser endomicroscopy, in the growing trend of machine learning applied to medical image analysis¹¹⁻¹³, propose a fully automated classification method based on supervised learning that we validate on thousands of images. This work is a priori challenging since the amount of data in preclinical studies, such as in our case, is rather limited compared to the usual amount of data available in medical applications of machine learning. Also, another a priori open question addressed in the preclinical study is the question of translational research, i.e. the reusability of the knowledge gained for animals on human or human on animals. We

¹Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRA IRHS, Université d'Angers, Angers, 49000, France. ²INSA-Lyon, INRIA, LIRIS, CITI, CNRS, Villeurbanne, France. ³Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, Lyon, 69621, France. *email: david.rousseau@univ-angers.fr

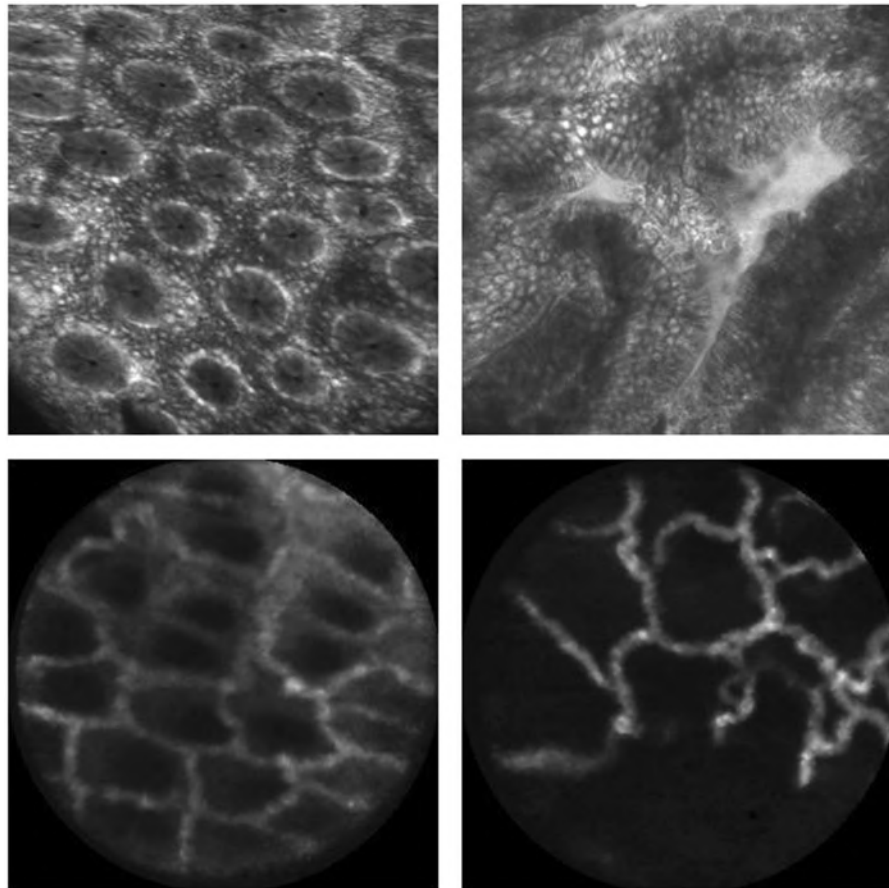


Figure 1. Top: Human samples of colon's wall images: healthy (left) and unhealthy (right) tissues observed from fluorescent confocal endomicroscopy. Bottom: Mouse samples of colon's wall images: healthy (left) and unhealthy (right) tissues observed from fluorescent confocal endomicroscopy.

address this question here, for the first time to our knowledge, in the perspective of machine learning. As the last innovation in our methodology to address a specific unsolved preclinical problem, we discuss different scientific use cases and corresponding strategies for training concerning some properties of confocal laser endomicroscopy. Images are acquired at the video frame rate while the expert holding the endoscopic probes moves it slowly to inspect the tissue when located close to the tissue of interest. Consequently, though the imaging system is producing vast amounts of images, a large number of images are very similar. We consider the possibility of taking benefit from this self-similarity in order to significantly reduce the size of the data set requested during the training stage. This training approach is vital for the expert in charge of the annotation of the training data sets since it is a highly time-consuming task. In a second configuration, we also discuss the performance obtained with different machine learning approaches when we learn on images corresponding to a given set of mice while applying the classification on a distinct cohort of mice. This cross-subject training is relevant for clinical purposes because it quantifies to which extent the disease observed is generic or patient-specific. The performances of these two training strategies compared to the best performance obtained with a brute force random sampling on a whole cohort for the training of the classification algorithm.

In the literature, several studies have focused on the classification of colon's health state from endomicroscopy. Up to our knowledge, this body of work based on the classical methodology of handcrafted feature design (taking into account domain knowledge), followed by supervised machine learning.

A method based on global descriptors proposed in⁵, whose introduced fractal box-counting metrics and illustrated them on two images. Vessel detection was proposed in⁸ after a Hessian-based filter in addition to length area and diameter measurements of vascular crypts of the colon's wall. Blood vessels of the colon's wall characterized in⁹ from Fourier analysis. Also, vascular networks of colon's wall were characterized in terms of graphs in¹⁰ after skeletonization on few hundreds of images.

Closest to our work is the method by Ștefănescu *et al.*, which is based on machine learning with neural networks of images of human tissues¹⁴ acquired with confocal laser endomicroscopy. However, the images are clearly different; in contrast, the field of view and resolution, as can be seen in Fig. 1. These differences motivate our proposition of designing a specific method for mice trained on mouse images. In contrast to¹⁴, we (i) propose a method based on representation learning¹⁵ as opposed to handcrafted features, and (ii) specifically discuss different experimental protocols and develop different training strategies adapted to these protocols.

	Healthy mice	Mice with cancer	Mice with inflammation
Training	5	7	7
Validation	1	2	2
Testing	3	4	7

Table 1. Number of mice in each dataset.

Left			Right			
Classifiers	Transfer learning	Accuracy		True Cancer	True Inflammation	True Healthy
Proposed CNN architecture	—	98.49% ± 0.6	Predicted Cancer	13107	0	0
DenseNet	X	94.54% ± 2.9	Predicted Inflammation	0	5012	46
VGG16 + linear SVM	X	90.60% ± 0.4	Predicted Healthy	0	75	2011
VGG16	X	89.62% ± 3.3				
ResNet50	X	75.93% ± 4.1				
VGG16	—	74.82% ± 3.2				
LBP features + linear SVM	—	83.01% ± 0.4				
Proposed method at ¹⁴	—	77.41% ± 1.3				

Table 2. Left: Results of cross-subject training with full data, where all images of 6 healthy mice, 9 mice with cancer, and 9 mice with inflammation used for training the system. Right: Confusion matrix of cross-subject performance where our proposed CNN architecture is used.

Results

In this section, we give experimental results using the experimental protocol and training strategies described in the method section as well as the different feature extraction and feature learning techniques.

Cross-subject training. For this protocol, the most challenging one of all considered cases, where generalization to unseen subjects (mice) is required, randomly chosen images of mice for three datasets of training, validation, and testing as shown in Table 1. While the training set is used to adjust the parameters of the model, the validation set is used to minimize overfitting and tune the parameters. The test set of unseen data is used to confirm the predictive power and that the model generalises. The final classification of trials is computed as the average performance of each fold. The number of healthy and unhealthy mice are not equal. We simulated cross-validation for this approach by changing mice between training, validation, and testing for each new experiment.

Table 2 gives results with the different feature representations and classifiers described in the method section. In addition, Table 3 shows classification accuracy of a transfer learning method with different freezing layers discussed in section. Our proposed architecture trained from scratch shows the best recognition rate compared to handcrafted features, and state of the art high-capacity architectures with pre-training. The experiments indicate that high-capacity networks overfit on this amount of target data even when they are pre-trained on large datasets of natural images. We conjecture that the shift in data distributions is too large in the case of this application. The last layer of the network, still trained from scratch even in the case of transfer learning, overfits on the small target data set. To sum up the essence of the contribution, we train a high-capacity model on a large scale data set, followed by fine-tuning of a low capacity SVM model on the small volume target data set.

Also, we studied the dependency of the classification results on the number of subjects in the training data, as illustrated in the Fig. 2. For this study, we chose the LBP based representation and the SVM classifier since it can work better when a small size of the database is available for training. As expected, the system performance increases significantly when additional mice are added to the training set, as each mouse potentially has its specific pattern for health, inflammation, and cancer tissues.

Figure 3 shows some cases of correctly and wrongly classified images with their coarse localization maps. As can be seen, these images are indeed difficult to assess as the miss classified images have a similar pattern with another class.

Cross-sample training with all samples. Let us recall that in another use case of cross-sample training, subjects (mice) are mixed between training and test sets. In our setup, the 7 fold cross-validation approach used where almost 75% of images are dedicated for training and 25% of images for testing purposes, which corresponds to the proportions chosen for a similar problem in¹⁴, albeit for human colon's walls. When needed, the validation set was chosen from the training set. Table 4 gives the prediction performance of the different classifiers on this data. We report means and standard deviations of ten runs.

In this more natural case, where correlations between subsequent frames in the input video can be exploited, our CNN architecture still outperforms other models and feature learning methods with a close to perfect performance of 99.33%. Even transfer learning of deep networks cannot compete in this section, where generalization to unseen subjects is not an issue. We conjecture that the reason is that pre-training on the large-scale data set

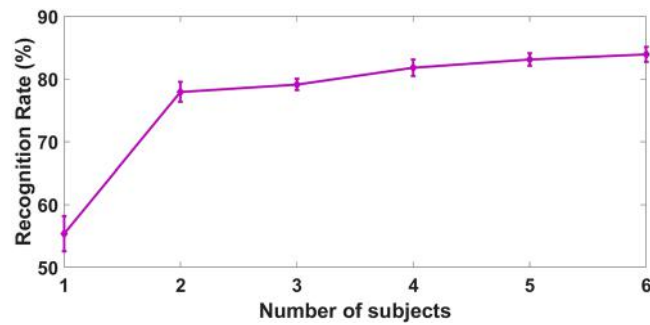


Figure 2. Dependency on the number of training subjects for cross-subject training (LBP features + SVM classifier).

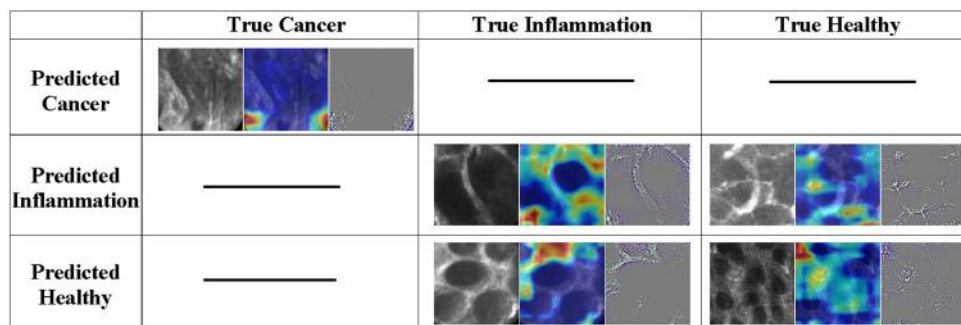


Figure 3. Example of correctly and miss classified images of the proposed CNN architecture for the cross-subject training strategy. Each cell consists from left to right of a grayscale image, a coarse localization map of the important regions in the image for the network⁴⁰, and a high-resolution class-discriminative visualization⁴⁰. Cells with dashed lines mean that there is no miss classified images for that class.

No. Freezing Conv. layers	1	2	3	4	5	6	7	8	9	10	11	12	13
Accuracy	40.8% ± 17.4	65.6 ± 29.9%	89.6 ± 3.3%	89.2% ± 3.9	42.8% ± 21.9	43.4% ± 23.25	70% ± 24.1	52.8% ± 22.2	75.4% ± 23.9	82.2% ± 9.4	65.8% ± 29.9	41.2% ± 18.3	33% ± 0

Table 3. Results of cross-subject training with different numbers of frozen layers when transferring the VGG16 network from ImageNet to the target dataset.

Left			Right			
Classifiers	Transfer learning	Accuracy		True Cancer	True Inflammation	True Healthy
Proposed CNN architecture	—	99.93% ± 0.13	Predicted Cancer	13994	0	0
LBP features + linear SVM	—	97.7% ± 0.39	Predicted Inflammation	0	4032	0
VGG16 + linear SVM	X	85.9% ± 0.4	Predicted Healthy	0	5	1849
VGG16	X	82.12% ± 4.1				
ResNet50	X	79.94% ± 4.6				
DenseNet	X	79.51% ± 3.8				
VGG16	—	78.49% ± 1.27				

Table 4. Left: Results of cross-sample training with full data. Right: Confusion Matrix of cross-sample performance where our proposed CNN architecture is used.

learns a representation tailored for high generalization, which requires encoding invariances to large deformation groups into the prediction model. These invariances help to recognize natural classes, like animals and objects from daily life, even though their viewpoints and shapes might be profoundly different. It is clearly not the

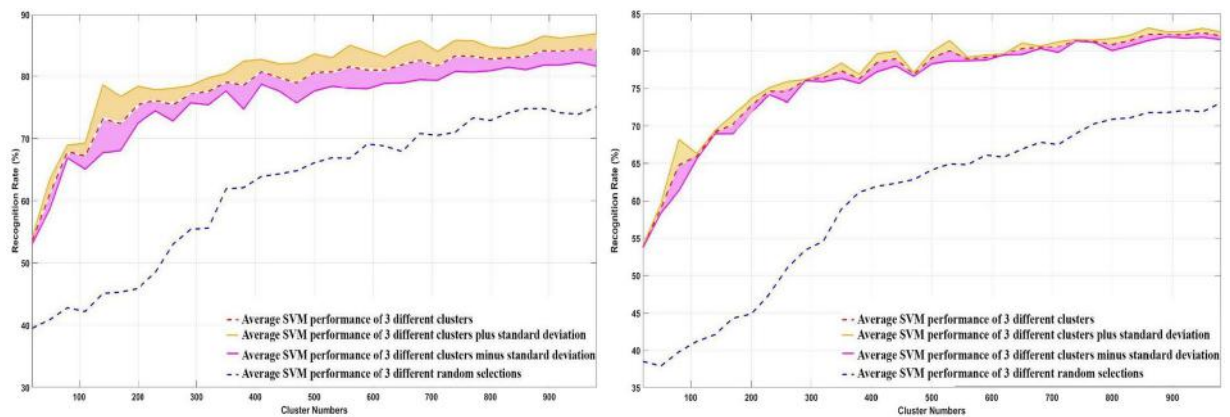


Figure 4. Average of recognition rate of cross-subject (left) and cross-sample (right) training respectively with sample selection in solid red line versus a random selection of data in dashed blue line as a function of the number of images in the training dataset. Yellow and purple lines show the average recognition rate plus and minus standard deviation respectively.

objective for our cross-sample use case, where generalization is less an issue than encoding extremely fine-grained similarities between samples which are very close in feature space.

Overall deep learning methods with a pre-training, the best results were obtained by the VGG16 model pre-trained on ILSVRC and fine-tuned on our target data set, where after fine-tuning a linear SVM classifier was trained on the last feature layer of the deep network. Interestingly, this performance is comparable to what was obtained in¹⁴ for a similar colon's wall classification but on humans.

Cross-sample and cross-subject training with sample selection. We tested the performance of the handcrafted pipeline when the number of input data is limited. For this approach, images of each state are divided into training and testing sets, and then the training set is split into an increasing number of clusters based on their similarities. We stop at around 1000 clusters when a plateau of performance is reached. Then, a random image of each cluster in each state is selected to train the model, and the model is tested on the test data. Figure 4 shows the average recognition rate of the system after three trials as a function of the number of clusters, i.e., the size of the data set for the training for both cross-subject and cross-sample approaches. As visible in Fig. 4, the performance of both cross-sample and cross-subject training with sample selection overpasses the random selection of images with a gain approximately constant of 13% of recognition rate in all the range. However, at its maximum level, the performance is lower than the best performance obtained in Table 4. This approach can also be used for real-time applications as there is no need to use clustering on test data.

Methods

Experimental protocols and associated training strategies. Our main objective is to automate the classification process of mouse tissues into three classes, healthy, inflammation, and cancer tissues. Below, we describe two different medical use cases, where these predictions are helpful. In other words, two different approaches of splitting data into training and testing for our experiments are introduced, which refers to two different clinical problems where prediction is required on subjects or samples.

Scientific use cases. *Cross-subject predictions.* This use case arises when a prediction must be made on unknown subjects (unknown mice) using a model which has been created (trained) during an off-line training phase. The underlying scientific question addressed by this use case is whether locally acquired samples of tissue can be correctly classified without any additional information from the same subject. Alternatively, in other words, we would like to study whether prediction models based on machine learning can generalize to unseen subjects; it quantifies to which extent the observed diseases are generic or patient-specific.

In a real-world scenario, the corresponding prediction model is static in a sense that different predictions on new subjects will be based on the same model acquired by the medical personnel at a single instant (software updates not with standing). It means a model is trained on a given set of subjects, and will then apply it to new subjects (previously unseen). Decoupling training and prediction is the main advantage of this use case, as the prediction model does not require re-training between predictions, and results can be obtained using the same model on any new subject.

Cross-sample predictions. The second use case focuses more on individual tissue samples. This situation arises when one or more subjects are studied in detail, and a large number of tissue samples need to be classified. The underlying scientific question is, whether tissue annotation can be done semi-automatically when a large number of tissues need to be annotated from a low number of subjects. Alternatively, in other words, we would like to study whether a prediction model based on machine learning can generalize to different regions from the same or different subjects.

In a real-world scenario, the corresponding prediction model is dynamic, as (on-line) re-training is necessary for regular intervals. The medical personnel uses an application, which allows them to view tissue samples and annotate them in real-time, available in the additional information section.

The two uses cases are inherently different. Cross-subject predictions are usually more difficult, as the shift between the training data distribution and testing data distribution is generally higher, putting higher requirements on the generalization performance of the predictors. In practice, both cases can be addressed using fully supervised machine learning.

Proposed training strategies. We propose three different training strategies to address the scientific use cases described above.

Cross-subject training. This training strategy is designed to cover the cross-subject use case. The data set is split cross-subject wise, i.e., that subjects (mice) whose samples are in the training set are not present in the test set. It should be considered that the colon's wall of a subject can sometimes consist of all three labels at the same time, which means that a part of the colon's wall show cancer tissues. Another part show some inflammation tissues, and the rest can be considered as healthy tissues. Thus, it is essential to design a classifier that tries to label every image independently. Later a subject could be labeled based on the majority of labels of its images.

Cross-sample training with all samples. This strategy corresponds to the cross-sample use case. The data set is split into training and test sets by randomly sampling images of each type to be classified (health, inflammation, and cancer). In particular, this approach selects images without information on whether they are consecutive in video frames, or whether they belong to a given subject. In this strategy, images from one subject (a mouse) can be in both training and testing sets, but it does not mean that the same images are used in training and testing. As the microprobe captured images through the colon's wall of subjects, each image is taken from one specific part (tissue) of the colon's wall.

Cross-sample training with sample selection. In an alternative training strategy for the cross-sample use case, we address the fact that images correspond to video frames which are acquired in the continuity of a local probe inspection process. Therefore, consecutive images are visually similar with a high probability. This temporal correlation between frames can lead to skewed (unbalanced) data distribution and, if not dealt with, to sub-optimal performance.

We propose an unsupervised sample selection processing based on clustering. Features are extracted from each image, which includes standard deviation, mean, variance, and the skewness of the raw pixel values. The features are clustered with k-means, and a single sample is picked from each cluster for training. The rest of the images of the database are used for testing.

Features, feature learning and classification. Independently of the training strategy, we proposed two different procedures, including both feature extraction and classification methods. The first is based on handcrafted features, whereas the second resort to automatic learning of the intermediate representation.

Handcrafted features. In this methodology, we handcraft feature representations instead of learning them. Handcrafted representations have been optimized by the computer vision community over decades of research, including theoretical analysis and experiments. In our setting, we resort to the local binary patterns (LBP)¹⁶, a state-of-the-art handcrafted descriptor which has been used in a variety of tasks in computer vision, among which are face recognition, emotion recognition, and others, see the survey in¹⁷. Notably, LBPs have been shown to be valuable for medical image texture analysis¹⁸.

Under the original form of¹⁶ and as used in this article, for a pixel positioned at the point (x, y) , LBP indicates a sequential set of the binary comparison of its value with the eight neighbors. In other words, the LBP value assigned to each neighbor is either 0 or 1, if its value is smaller or greater than the pixel placed at the center of the mask, respectively. The decimal form of the resulting 8-bit word representing the LBP code can be expressed as follows:

$$LBP(x, y) = \sum_{n=0}^7 2^n s(i_n - i_{x,y}) \quad (1)$$

where $i_{x,y}$ corresponds to the grey value of the center pixel, and i_n denotes that of the n^{th} neighboring one. Besides, the function $s(x)$ is defined as follows:

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases} \quad (2)$$

The LBP operator remains unaffected by any monotonic gray scale transformation, which preserves the pixel intensity order in a local neighborhood. It is worth noticing that all the bits of the LBP code hold the same significance level, where two successive bit values may have different implications. The process of Eq. (1) is realized at the scale of a patch size of $N \times N$ pixels. The $LBP(x, y)$ of each pixel inside this patch are concatenated to create a fingerprint of the local texture around the pixel at the center of the patch. Eqs. (1) and (2) are applied on all patches of an image. Finally, all histogram outputs of patches (after applying LBP on them) are concatenated and considered as the feature vector of an image. This patch size N , in this study, is chosen in the order of an average

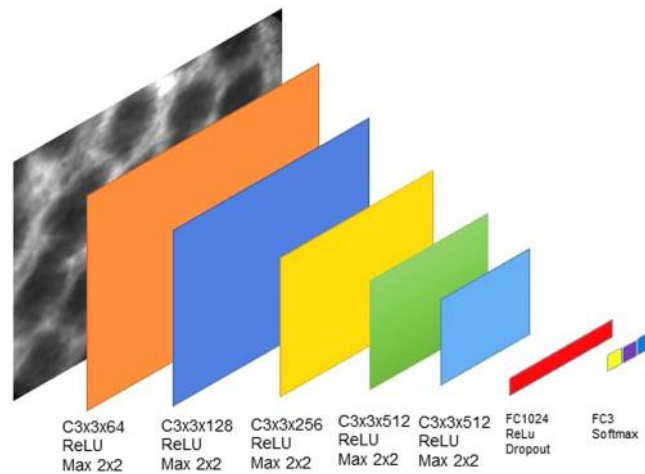


Figure 5. The proposed architecture of the deep network optimized for the task on the cross-validation set.

size of vesicular crypts on health images. In our database, a patch size of 8×8 can almost cover a healthy vesicular crypt. At the next step, a linear SVM is applied to classify the images based on their LBP features.

Representation learning. Representation learning, or deep learning, aims at jointly learning feature representations with the required prediction models. We chose the predominant approach in computer vision, namely deep convolutional neural networks¹⁹, which have proven to be well suited for standard tasks in the medical domain like cell segmentation²⁰, tumor detection, and classification²¹, brain tumor segmentation²², De-noising of Contrast-Enhanced MRI Sequences²³ and several other purposes¹⁵. We train two different models, one which was designed for the task and trained from scratch, and one which has been adapted from (and pre-trained on) image classification.

Training from scratch. The baseline approach resorts to a standard supervised training of the prediction model (the neural network) on the target training data corresponding to the respective training strategies described in section. No additional data sources are used. In particular, given a training set comprised of K pairs of images x_i and labels \hat{y}_i , we train the parameters θ of the network f using stochastic gradient descent to minimize empirical risk:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^K \mathcal{L}(\hat{y}_i, f(x_i, \theta)) \quad (3)$$

\mathcal{L} denotes the loss function, which is cross-entropy in our case. The minimization is carried out using the ADAM optimizer²⁴ with a learning rate of 0.001.

The architecture of our proposed architecture $f(\cdot, \cdot)$, shown in Fig. 5, has been optimized on a cross-validation set and is given as follows: five convolutional layers with filters of size 3×3 and respective numbers of filters 64, 128, 256, 512, 512 each followed by ReLU activations and 2×2 max pooling; a fully connected layer with 1024 units, ReLU activation and dropout ($p = 0.5$) and a fully connected output layer for 3 classes (health, inflammation and cancer) and softmax activation.

Transfer learning. Deep learning addresses complex prediction problems through neural networks with high capacity, i.e., highly non-linear functions with a large number of parameters, whose estimation typically requires a large amount of annotated training data. If this data is not available, the trained networks tend to overfit on the training data and thus generalize poorly to unseen data.

A standard solution to this problem is transfer learning or domain adaptation. The idea is to learn high capacity models on large alternative source data sets whose content is sufficiently correlated with the target application and then transfer the learned knowledge to the target data. Various techniques have been proposed, which differ, among other in the way this transfer is performed and whether labels are available for the target data set (supervised techniques, e.g.^{25,26}) or not (unsupervised techniques, e.g.²⁷).

We perform supervised transfer using classical weight freezing and fine-tuning²⁵, which transfers knowledge by first solving Eq. 3 on the target data set, and then using the obtained parameters θ^* as initialization (starting point) for the training of the network on the target data set. The assumption is somehow grounded by the existence of standard features in images from natural scenes, which transfer well to images from other domains.

We transfer knowledge from the well-known image classification task ILSVRC 2012 (aka *ImageNet*), a dataset of roughly one million images and 1000 classes²⁸. Our model architectures optimized for this task, and as described above, is very likely to underfit on this transfer learning setting. Its hyper-parameters, among which are its architecture and the number of parameters, has been optimized over a validation set, which is very much smaller than the ILSVRC data by roughly a factor of 500. Its design capacity will, therefore, tend to be much too

small for the knowledge encoded in the source data (ILVSR). For this reason, we take “classical” and well-known high-capacity models for the ILVSR task, namely VGG16²⁹, DenseNet³⁰, and ResNet50³¹. From the pre-trained model, we remove the task-specific output layer (designed for 1000 classes) and replace it with a new layer for three classes. Among all possible combinations of freezing layers which tested, the model with freezing at the first 3 layers and fine-tuning the other layers on the validation data set returned the best performance shown in the Table 3. The results of the transfer learning method with different freezing layers on our database show the transferability of features from ImageNet database in the spirit of²⁵.

We would like to point out that the two different strategies (training from scratch vs. pre-training and transfer) are compared using two different model architectures. Our goal is to compare strategies, and different strategies can possibly have different optimal architectures. Network architectures need to be adapted to various parameters of the problem, namely the complexity of the task and the number of training samples. As mentioned above, in our case, there is a big difference between the small size of our dataset and the large size of typical computer vision datasets like the ImageNet/ILSVRC dataset (1 M images). Therefore, this involves optimizing parameters (through SGD) as well as the hyper-parameters (through model-search). Only if both are optimized, the potentials of the two strategies are compared. In contrast, comparing two identical architectures would have been inconclusive, as one of two architectures would have been better suited to the task at hand.

Research involving animals. All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted.

Ethical standards. This study was approved by the institutional review board of the Université Claude Bernard Lyon 1 (reference number: DR2014-62-v1) and complied with ethics committee standards.

Annotating software. The annotating software tool has been specially developed for this study but is applicable to any video endoscopy annotation for cancer. It is freely available at <https://uabox.univ-angers.fr/index.php/s/AZ21Zl6LDYRcd8P> together with a demo video and some data sample.

Database

The experiments involving animals were led in accordance with the rules of the University Lyon 1 Ethics Committee on animal experimentation. Animals were acclimated for two weeks prior to the experiment in the following environment: a 12-hour day/night rhythm in 300 cm² plastic cages (for four animals) with straw bedding, pellet food, and tap water. The temperature of each cage was monitored and kept between 19 and 21 °C. To induce colitis, mice were chemically treated with a single injection of azoxymethane (AOM, intraperitoneal injection, 10 mg/kg body weight) at the beginning and then, during six months, with dextran sulfate sodium in drinking water (DSS, concentration of 2%). During the experiment, a pressure sensor placed on the mouse's chest in order to monitor the respiratory index of animals. Analyzed images used in this article chosen at the extrema of the respiratory cycle, where the movements are the slowest to minimize artifacts due to these movements. Mice anesthetized with 3% isoflurane and aspiration flow set at 0.4 L/min during the induction phase. A 25 µL solution of Fluorescein Isothiocyanate FITC-Dextran 5% (Sigma Aldrich), used as a contrast agent, is injected in retro-orbital of the mouse's eye before the CEM investigation.

The anesthesia maintained during imaging with 1.4 to 1.7% isoflurane vaporization and aspiration flow set up on 0.4 L/min. The endoscopic test was conducted using a mini multi-purpose rigid telescope dedicated to small animals (Karl Storz). Acquisition of images made by using a 488 nm confocal endomicroscope CEM (CellVizio c, Mauna Kea Technologies) combined with a 0.95 mm outer diameter Proflex MiniZ microprobe (PF-2173, Mauna Kea Technologies). The microprobe was inserted through the operating sheath of this endoscope and positioned on the mice's colon walls. During the acquisitions, the depth assessed was approximately 58 µm for a lateral resolution of 3.5 µm and a frame rate of 12 fps. The output image size is 329 × 326 µm² corresponding to a matrix of 292 × 290 pixels¹⁰.

In total, 38 mice were included in the study for a total of 66788 images which have been annotated as healthy tissue images (6474 images from 9 mice), cancer tissue images (46566 images from 13 mice) or inflammation tissue images (13748 images from 16 mice) by two experts together at the same time with a pre-knowledge of mice diseases. Images were also labeled according to the mice from which they were acquired. Annotation was realized with the help of an application (available in the additional information section) especially developed for this study freely available, as pointed in the supplementary material section. It enables the classification of images according to the three classes studied in this article but also other classes of interest in biomedical studies of the colon's wall. This application is made available as supplementary material to this study. As mentioned in⁵, some of the raw images do not carry any information for diagnosis. This can be due to misposition of the probe which does not receive enough signal, a decrease of the fluorescence, saturation of the imaging sensor due to too high amount of fluorescence, due to residues, due to contrast agent extravasation or presence of some light-absorbing objects within mucous film located between the probes and the tissue. To prevent the expert from spending time on annotating such non-relevant images and improve the learning process, we decided, as usually done in video endomicroscopy^{32,33} to withdraw them automatically and only keep the informative frame. A simple test based on the computation of the skewness of the gray level histogram of the images demonstrated to be very efficient for this task. Images with a skewness higher than -5 (as an empirical threshold) were kept. The skewness captures the dissymmetry of the histogram around its mean value. This is useful to detect saturated or underexposed images. We estimated, on some 6000 images, that this simple statistical test performs 98% of good detection for the detection of images carrying no useful diagnostic information with a false alarm of 1%. Additionally, in order to assess the influence of these artifactual images if they would not have been removed, an additional experiment has been

done on all raw data (without removing noisy data). This experiment showed a reduction of 2% (on average) on the recognition performance of each training strategy by using our proposed CNN model. This demonstrates the interest of the denoising step but also quantify the robustness of our model.

Based on the training strategies, the database was spilled into three datasets of training (for training of our model), validation (to optimize hyper-parameters), and testing (to report performance on). In the cross-subject training strategy, images of each subject (mouse) were transferred into one of the datasets of training, validation, and testing. The exact number of mice in each dataset shown in Table 1. In the cross-sample training strategy, 75% of the whole database transferred to the training dataset, and the rest of the data belonged to the testing dataset. In this case, the validation dataset was extracted from the training dataset for deep learning experiments. This splitting database approach made a guaranty that the test dataset was not seen during training and validation of the model.

Conclusion

In this paper, we have presented three classification approaches to classify three states of health, inflammation, and cancer on mice colon's wall. Fully automated machine learning-based methods are proposed, including deep learning, transfer learning, and classical texture-based classification. Different training strategies are compared in order to find the best approach for this specific problem. The images processed in this paper were acquired in the framework of a preclinical study on colon mice. In this type of study (preclinical), the size of the database is not comparable with other domains in machine learning as also underlined in³⁴ on the different types of images, we found that a custom deep learning model shows superiority over handcrafted features and well-known deep learning-based architectures. The best classification performance on this type of images are achieved with our proposed CNN model which are trained on colon's wall images.

In the cross-sample case, where generalization to unseen subjects is not an issue, Deep learning gave a performance of 99.93% of correct classification. Similar to the cross-sample, in the cross-subject approach where classification on un-seen objects is an issue, our proposed CNN method showed a performance of 98.49% of correct classification. These are usual order of magnitude of performance obtained with nowadays machine learning approaches when vast data sets are available, but this can be considered as excellent performance indeed here since we worked with the typical small data sets available in preclinical studies.

This work corresponds to the first fully automated classification algorithm for mice colon's wall images reported in the literature. Similar works were carried on the human colon's wall with the same imaging system. The comparison of the closest work¹⁴ with our algorithm shows a comfortable margin of a 14% of accuracy. This is an interesting result which demonstrates that in the perspective of machine learning, there is no guarantee of translational research between human and animal. Also, a novel unsupervised sampling strategy based on the specific similarities of images in the acquisition of images with endomicroscopy in the colon has been designed. The interest of this sampling strategy has been demonstrated in terms of amount of data required in the training data sets to reach a plateau of performance. However, the performance of this sampling strategy is lower than brute forces classical approaches. It would be possible to improve the metric of similarity used to select the images in the training data sets automatically. This was based on first-order statistics in this study, but other approaches could be used to include more dynamical information. However, due to the multi-scale sources of temporal noise (movement of the probes³⁵, passing of unexpected items between probe and tissues, biological movement, etc.) it would be an open question to determine a reasonable time scale for this smoothing.

Our clustering method is somewhat related to active learning, where the agent requests feedback on data from a user. The comparison is a little bit a stretch, as no new data is collected from decisions by an agent. In our current implementation, the dataset stays stable, and only a subset is actively chosen.

However, we plan to investigate active learning as future work, where a classifier is trained on a subject followed by continued examination of the subject on new samples. Here, an agent could quickly provide decisions on (i) which samples should be added to the training set, and (ii) into which direction the user should emphasize its search in order to optimize performance and discovery. This leads to an exploitation/exploration trade-off known from Reinforcement learning.

Direct perspectives of other sampling strategies are possible. It would now be possible to apply the classification scheme developed here to produce a score on individual mice quantifying the number of images with the disease. Such a quantification could then be compared with clinical scores realized on other types of imaging systems in a multimodal perspective such as the one recently shown with magnetic resonance imaging³⁶. Also, the machine learning approach presented with a discussion on the different training strategies could be transposed to other bioimaging problems. In confocal endomicroscopy, this includes, for instance, the characterization of other colon's diseases observed in confocal microscopy³⁷ or other parts of the digestive system³⁸ or also to other organs³⁹ which have received interest and could benefit from machine learning approaches to perform automated characterization of tissues.

Received: 8 May 2019; Accepted: 9 December 2019;

Published online: 27 December 2019

References

1. Sirinukunwattana, K. *et al.* Gland segmentation in colon histology images: The glas challenge contest. *Med. image analysis* **35**, 489–502 (2017).
2. Brady, M., Highnam, R., Irving, B. & Schnabel, J. A. Oncological image analysis. *Med. image analysis* **33**, 7–12 (2016).
3. Becker, C., Fantini, M. & Neurath, M. High resolution colonoscopy in live mice. *Nat. protocols* **1**, 2900–2904 (2006).
4. Wang, H.-W., Willis, J., Canto, M., Sivak, M. V. & Izatt, J. A. Quantitative laser scanning confocal autofluorescence microscopy of normal, premalignant, and malignant colonic tissues. *IEEE Transactions on biomedical engineering* **46**, 1246–1252 (1999).

5. Waldner, M. J., Wirtz, S., Neufert, C., Becker, C. & Neurath, M. F. Confocal laser endomicroscopy and narrow-band imaging-aided endoscopy for *in vivo* imaging of colitis and colon cancer in mice. *Nat. protocols* **6**, 1471–1481 (2011).
6. Cicchi, R. *et al.* Multiphoton morpho-functional imaging of healthy colon mucosa, adenomatous polyp and adenocarcinoma. *Biomed. optics express* **4**, 1204–1213 (2015).
7. Evans, J. P. *et al.* From mice to men: Murine models of colorectal cancer for use in translational research. *Critical reviews oncology/hematology* **98**, 94–105 (2016).
8. Mielke, L., Preaudet, A., Belz, G. & Putoczki, T. Confocal laser endomicroscopy to monitor the colonic mucosa of mice. *J. immunological methods* **421**, 81–88 (2015).
9. JA Konda, V. *et al.* *In vivo* assessment of tumor vascularity using confocal laser endomicroscopy in murine models of colon cancer. *Curr. Angiogenesis* **2**, 67–74 (2013).
10. Bujoreanu, D. *et al.* Robust graph representation of images with underlying structural networks. application to the classification of vascular networks of mice's colon. *Pattern Recognit. Lett.* **87**, 29–37 (2017).
11. Na, K.-S. Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. *Sci. reports* **9**, 3335 (2019).
12. Singh, S. P. *et al.* Machine learning based classification of cells into chronological stages using single-cell transcriptomics. *Sci. reports* **8**, 17156 (2018).
13. Min, X., Yu, B. & Wang, F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on copd. *Sci. reports* **9**, 2362 (2019).
14. Ștefănescu, D. *et al.* Computer aided diagnosis for confocal laser endomicroscopy in advanced colorectal adenocarcinoma. *PLoS one* **11**, e0154863 (2016).
15. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).
16. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis machine intelligence* **24**, 971–987 (2002).
17. Nanni, L., Lumini, A. & Brahnem, S. Survey on lbp based texture descriptors for image classification. *Expert. Syst. with Appl.* **39**, 3634–3641 (2012).
18. Nanni, L., Lumini, A. & Brahnem, S. Local binary patterns variants as texture descriptors for medical image analysis. *Artif. intelligence medicine* **49**, 117–125 (2010).
19. Ravi, D. *et al.* Deep learning for health informatics. *IEEE journal biomedical health informatics* **21**, 4–21 (2017).
20. Akram, S. U., Kannala, J., Eklund, L. & Heikkilä, J. Cell segmentation proposal network for microscopy image analysis. In *Deep Learning and Data Labeling for Medical Applications*, 21–29 (Springer, 2016).
21. Akselrod-Ballin, A. *et al.* A region based convolutional network for tumor detection and classification in breast mammography. In *Deep Learning and Data Labeling for Medical Applications*, 197–205 (Springer, 2016).
22. Zhao, X. *et al.* A deep learning model integrating fcnn and crfs for brain tumor segmentation. *Med. image analysis* **43**, 98–111 (2018).
23. Benou, A., Veksler, R., Friedman, A. & Raviv, T. R. De-noising of contrast-enhanced mri sequences by an ensemble of expert deep neural networks. In *Deep Learning and Data Labeling for Medical Applications*, 95–110 (Springer, 2016).
24. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. In *ICML* (2015).
25. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 27, 3320–3328 (Curran Associates, Inc., 2014).
26. Douarre, C., Schielein, R., Frindel, C., Gerth, S. & Rousseau, D. Transfer learning from synthetic data applied to soil–root segmentation in x-ray tomography images. *J. Imaging* **4**, 65 (2018).
27. Ganin, Y. & Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *ICML* (2015).
28. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *IJCV* **115**, 211–252 (2015).
29. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR* (2015).
30. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *CVPR* **1**, 3 (2017).
31. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
32. Oh, J. *et al.* Informative frame classification for endoscopy video. *Med. Image Analysis* **11**, 110–127 (2007).
33. Bashar, M. K., Kitasaka, T., Suenaga, Y., Mekada, Y. & Mori, K. Automatic detection of informative frames from wireless capsule endoscopy images. *Med. Image Analysis* **14**, 449–470 (2010).
34. Murthy, V. N. *et al.* Cascaded deep decision networks for classification of endoscopic images. In *Medical Imaging 2017: Image Processing*, vol. 10133, 101332B (International Society for Optics and Photonics, 2017).
35. Latt, W. T. *et al.* A hand-held instrument to maintain steady tissue contact during probe-based confocal laser endomicroscopy. *IEEE transactions on biomedical engineering* **58**, 2694–2703 (2011).
36. Dorez, H. *et al.* Endoluminal high-resolution mr imaging protocol for colon walls analysis in a mouse model of colitis. *Magn. Reson. Mater. Physics, Biol. Medicine* **29**, 657–669 (2016).
37. Neumann, H. *et al.* Confocal laser endomicroscopy for *in vivo* diagnosis of clostridium difficile associated colitis—a pilot study. *PLoS One* **8**, e58753 (2013).
38. Liu, J. *et al.* Learning curve and interobserver agreement of confocal laser endomicroscopy for detecting precancerous or early-stage esophageal squamous cancer. *PLoS one* **9**, e99089 (2014).
39. Foersch, S. *et al.* Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors *in vivo*. *PLoS One* **7**, e41760 (2012).
40. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626 (2017).

Acknowledgements

This work was supported by the LABEX PRIMES (ANR-11-LABX- 0063) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR) as well as DORA plus (Estonian government programme).

Author contributions

Conceptualization, P.R. and C.W. and D.R.; Data curation, H.D. and R.S. and D.M.; Formal analysis, P.R. and D.R.; Methodology, D.R.; Software, P.R. and S.S.; Supervision, D.R.; Validation, P.R. and S.S. and D.R.; Visualization, P.R.; Writing - original draft, P.R. and D.R. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Article

Toward Joint Acquisition-Annotation of Images with Egocentric Devices for a Lower-Cost Machine Learning Application to Apple Detection

Salma Samiei ^{1,2} , Pejman Rasti ^{1,3} , Paul Richard ¹ , Gilles Galopin ² and David Rousseau ^{1,2,*}

¹ Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), Université d'Angers, 62 Avenue Notre Dame du Lac, 49035 Angers, France; salma.samiei@univ-angers.fr (S.S.); pejman.rasti@univ-angers.fr (P.R.); paul.richard@univ-angers.fr (P.R.)

² UMR 1345 Institut de Recherche en Horticulture et Semences (IRHS), INRAe, 42 Rue Georges Morel, 49071 Beaucouzé, France; gilles.galopin@agrocampus-ouest.fr

³ Department of Data Science, école D'ingénieur Informatique et Environnement (ESAIP), 49124 Angers, France

* Correspondence: david.rousseau@univ-angers.fr

Received: 9 May 2020; Accepted: 24 July 2020; Published: 27 July 2020



Abstract: Since most computer vision approaches are now driven by machine learning, the current bottleneck is the annotation of images. This time-consuming task is usually performed manually after the acquisition of images. In this article, we assess the value of various egocentric vision approaches in regard to performing joint acquisition and automatic image annotation rather than the conventional two-step process of acquisition followed by manual annotation. This approach is illustrated with apple detection in challenging field conditions. We demonstrate the possibility of high performance in automatic apple segmentation (Dice 0.85), apple counting (88 percent of probability of good detection, and 0.09 true-negative rate), and apple localization (a shift error of fewer than 3 pixels) with eye-tracking systems. This is obtained by simply applying the areas of interest captured by the egocentric devices to standard, non-supervised image segmentation. We especially stress the importance in terms of time of using such eye-tracking devices on head-mounted systems to jointly perform image acquisition and automatic annotation. A gain of time of over 10-fold by comparison with classical image acquisition followed by manual image annotation is demonstrated.

Keywords: egocentric vision; image annotation; apple detection; eye-tracking

1. Introduction

In the era of machine learning-driven image processing, unequalled performances are accessible with advanced algorithms, such as deep learning, which are highly used in computer vision for agriculture and plant phenotyping [1]. The bottleneck is no more the design of algorithms than the annotation of the images to be processed. When performed manually, this annotation can be very time consuming, and therefore very costly. Consequently, it is useful to investigate all possibilities to accelerate this process. Annotation time can be reduced via multiple approaches, which have all started to be investigated in the domain of bioimaging and especially plant imaging [2–9]. First, (i) annotation time can be reduced by parallelizing the task via online platforms [5]. Additionally, (ii) it can be reduced by using shallow machine learning algorithms that automatically select the most critical images or parts of the images to be annotated via active learning [4]. Transferring segmentation models (iii) learned over available datasets can significantly reduce the need for annotated data [10]. Another approach to reducing annotation time (iv) is to do the training on synthetic datasets that are automatically annotated [2,3,6,7,9,11]. At last, (v) annotation time can be reduced via the

use of ergonomic tools, which enable human annotators to accelerate the process without loss of annotation quality [8]. In this article, we contribute to the latter approach (v) to reduce annotation time. We introduce a novel use of egocentric devices in computer vision for plant phenotyping and assess their value to speed up image annotation.

The term “egocentric device” is used to designate all wearable imaging systems that record images from the first-person perspective. Images captured from egocentric devices are possibly of high value, since their field of view benefits from the attention of the person who wears the device and who is in charge of the targeted task to be done on the images. Reducing the field of view to a part of specific interest may reduce the complexity of the inspected scene and thus help the automatic processing of the acquired images. This is expected to be especially useful in complex scenes, such as those found outdoors in agriculture and phenotyping in the fields. Additionally, some egocentric devices, namely, head-mounted eye-trackers, can even include the capture of the ocular position of the annotator during the recording of the videos. This would, in theory, open up the possibility to annotate images directly, whereas acquisition and annotation are usually two separate steps. Such use of egocentric devices opens up the possibility to conduct these steps jointly and hence reduce annotation time. However, eye-trackers can never be perfectly calibrated, and their practical value in terms of both performance and time is still to be assessed in order to speed up annotation. That is what we propose here.

For the first application of egocentric devices to accelerate annotation, we considered as a proof of concept, a standard problem in computer vision for plant phenotyping. We chose the detection, i.e., segmentation, counting, and localization of apples in color images. This task has been addressed in many ways, including recently, with deep learning. This canonical problem is challenging for computer vision, since it includes self-occlusion of multiple instances, occlusion by the shoots of the apple trees, the variation of illumination, clutter from the self-similar background, variety in sizes and colors of fruits, etc. Additionally, this computer vision problem is significant for various agricultural applications, such as the design of automatic harvesters, automatic estimation of the fruit pack out, and variety testing. Most state-of-the-art methods developed for apple detection are currently working with supervised learning. Such methods require annotated images of apples to be efficient. In this article, we demonstrate how the use of egocentric devices can accelerate the annotation of apples in images. This acceleration in image annotation, illustrated here with apples, is of high value since it could benefit from reducing the annotation cost of any supervised learning segmentation method.

A visual abstract of the proposed original approach for a joint image acquisition-annotation process is illustrated with apple detection in Figure 1. For comparison, the conventional approach is also depicted in Figure 1 wherein a handy camera is used to acquire images, and after image transfer to a computer, images are manually annotated. We propose a single-step approach where hands-free, head-mounted cameras with embedded computational resources are jointly acquiring and annotating images. The article is organized as follows. After positioning our work with the most related work (Section 2), we present (Section 3) the egocentric devices used, the acquisition protocol, and the dataset created for this study. A classical algorithm adapted from the literature is described, as we use it to detect apples in color images (Section 4). The same algorithm is then applied to compare five different computational strategies, specially designed for this study, to reap benefits from egocentric vision (Section 5). We finally conclude on the best practice identified via this comparison.

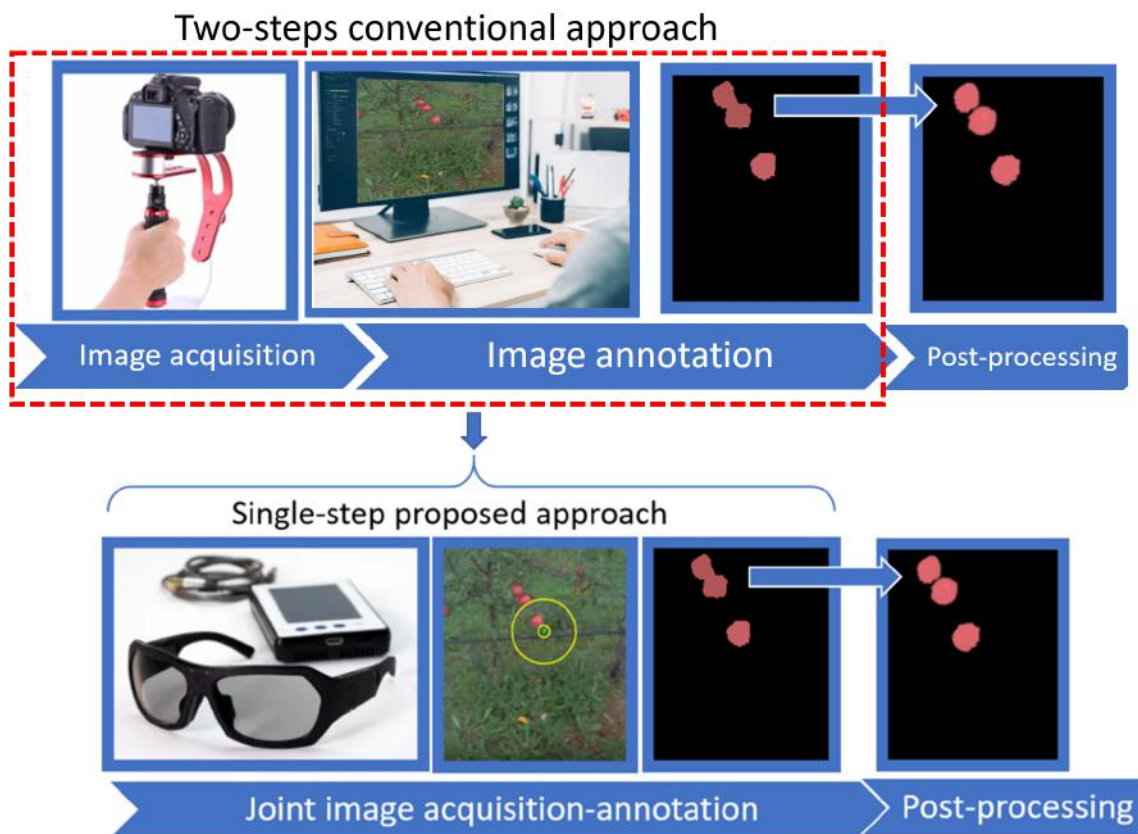


Figure 1. Visual abstract of the article. The red dotted-line encapsulates the conventional two steps of the acquisition and annotation process. We jointly perform image acquisition and image annotation by the use of a head-mounted egocentric device, which simultaneously captures images and the gaze of the person who wears the device and reaps benefits from both factors to annotate images automatically. It is to be noted that the post-processing step to separate touching annotated objects is not included here. It remains a step necessary in the conventional two-step approach and our proposed single-step approach.

2. Related Work

Egocentric (first-person) vision is a relatively new research topic in the field of computer vision which is increasingly attracting interest for understanding human activities [12–15], object detection [16,17], creation of models of the environment with different levels of precision [18,19], perception of social activity [20], user–machine interactions [21], driving assistance [22], and medical applications [23–25]. There are different types of egocentric systems, such as smart glasses, action cameras, and eye-trackers. Based on the processing capabilities, embedded sensors, such as the one used in this article, are now more and more utilized in conjunction with egocentric video analysis [21]. Features such as hand appearance and head motion give essential cues about the attention, behavior, and goals of the viewer [26–29]. In our case, we also used the fact that, usually, in egocentric vision, salient objects of interest tend to occur at the center of the image, since they attract the attention of the viewer [16,30]. In this article, we primarily used an eye-tracking system for egocentric vision to speed up image annotation. The use of eye-tracker to speed up image annotation has been proven useful for annotation with a screen-based system in [8,31,32]. Those studies demonstrated a possible gain of time for annotation of 30-fold (approximately) by comparison with manual annotation. Here, we use, for the first time to the best of our knowledge, an embedded eye-tracking system in the form of glasses (see Figure 1) to jointly conduct image acquisition and annotation and thus extend the results of [8,31,32]. Embedded eye-tracking systems are known to be less accurate than screen-based eye-tracking systems because they can move slightly on the head of the observer during acquisition. However, embedded

eye-tracking systems open the door for an accelerated procedure with joint acquisition and annotation, as illustrated in Figure 1. In this article we will compare the performances in terms of accuracy of apple detection and annotation time of both screen-based eye-tracking systems and embedded eye-tracking systems for image annotation.

Object detection in agricultural conditions has been investigated with a large panel of computer vision approaches [33–45]. In the early works, such as [33], methods were handcrafted both from the hardware side and the software side. Nowadays, it is more common practice to use standard RGB cameras, and base the detection of apples on supervised machine learning methods learned end-to-end via deep learning, as in [44,45]. Such modern methods, neural network-based, show high performances but require large amounts of annotated images. Manual pixel-wise annotation is, in general, a time-consuming operation, taking approximately 1.5 h per 100 images (308×202 pixels). In practice, apple detection is also challenging because of illumination conditions [46–48]. In this article, we will not provide a novel method to detect apples automatically. Instead, we will investigate the possibility of performing acquisition and annotation of apples in an orchard environment simultaneously by using head-mounted egocentric devices. Indeed, while there has been significant recent interest in fruit detection, segmentation, and counting in orchard environments, the cost of providing a unified annotated dataset of the fruit on trees makes it the bottleneck in the state-of-the-art literature [49].

The head-mounted egocentric camera provides areas of interest located in the vicinity of the targeted objects in the scene. Therefore, these areas of interest are less accurate than if a manual annotator was pointing at the object with a mouse. We propose in this article to test a standard image segmentation approach to detect the targeted object in the areas of interest provided by the head-mounted egocentric camera. As a consequence, the work relates to the literature on weakly or semi-supervised learning [50] with inexact supervision; that is, the training data are given with labels that are not as exact as desired. Different semi-supervised learning models have been introduced, such as iterative learning (self-training), generative models, graph-based methods, and vector-based techniques [51,52]. The color-based clustering technique for apple detection by using Gaussian Mixture Models was explained in [53]. In this approach, the SLIC superpixel was applied to the input image using the LAB color space. The superpixel's results were clustered into approximately 25 color classes. Finally, based on the KL-divergence between Gaussian Mixtures, each superpixel was classified into an apple or background [54], from hand-labeled classes. Our objective was not to design a novel semi-supervised algorithm. Instead, we revisited existing standard methods based on superpixels and assessed the value of the areas of interest extracted by the head-mounted egocentric camera for a given task of object detection.

3. Material and Method

3.1. Egocentric Vision Device

The egocentric imaging system used was VPS-16 head-mounted eye-tracking glasses equipped with stereoscopic cameras in the nose bridge, a front camera with a diagonal coverage of 88 degrees, and an audio microphone sampling at 10 kHz. The front camera was calibrated with the eye-tracker before acquisition. The visual task defined to the wearer was to find apples on the targeted trees. The acquisition time was nearly 90 s for the whole dataset (calibration time included). This acquisition time is quite similar to the time required with a digital camera fixed on a tripod or hand-held, the former of which would need to be located in different positions to cover all apples located on a tree. The distance of the viewer and the tree was set approximately to one and a half meters. The viewer was counting the number of apples as evidence of the ground-truth, which was recorded via the audio microphone. Fixation points were recorded by the eye-tracker to investigate how they could serve to automatically annotate apples on the trees.

3.2. Dataset

With the sensor described in the previous subsection, we generated a new dataset of 10 videos (25 fps) from 10 various apple trees in the orchard environment captured by the egocentric head-mounted glasses' eye-tracker. The total number of extracted images from the entire dataset was 24,618 (frames). A fundamental parameter of eye-tracking analysis depends on the definition of the fixation and the algorithm used to separate fixation from saccades [55]. Fixation refers to a person's point-of-focus as they look at a stationary target in a visual field. Although the mean duration of a single fixation may depend on the nature of the task [56], numerous studies have been done to measure the average duration for a single fixation [56–65]. The mean fixation duration for visual search is 275 ms, and for tasks that require hand-eye coordination, such as typing, the mean fixation can be 400 ms [56]. Among our dataset, the number of frames which received gazing of at least 275 ms was 419. The acquisitions were made on two days at midday with different weather conditions at the orchard of INRAE Angers, France. No difference was found in the results of the data coming from the two days. This dataset includes a variety of apple colors together with apple and foliage density, which are representative of the dataset found in the literature for apple detection [66–68]. Due to the complexity of each orchard tree, the illumination, and the environment itself, different natural colors were found in the images, including various shades of green, red, yellow, brown, or gray for the appearance of foliage, grass, apples, and tree trunks. Ground-truth was created by manual annotation of the raw color images at approximately 54 s per image by using the Image Segmenter application in Matlab 2017a. A sample of raw color images from different apple trees and their corresponding manual ground-truth are illustrated in Figure 2. For the whole dataset, which consists in 419 images, it roughly took 6 h to manually annotate all images. These manual annotations were generated for evaluation of the accuracy of the egocentric vision methods presented in the next section.

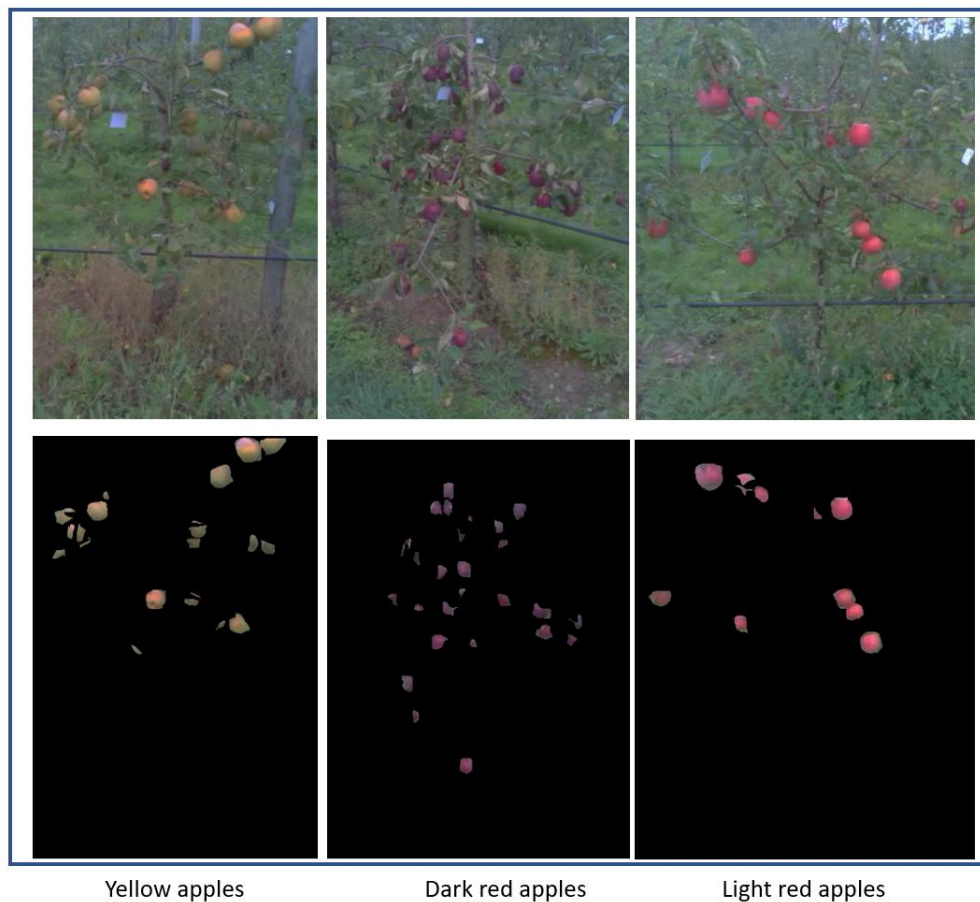


Figure 2. Example of RGB images of apple trees from our dataset and the corresponding ground-truth (manually annotated).

4. Image Processing Pipeline

In this section, we present the image processing pipeline developed to automatically annotate apples from the attention areas captured with egocentric vision. A global view of this pipeline is depicted in Figure 3 and includes three main steps: image pre-processing, segmentation, and performance evaluation.

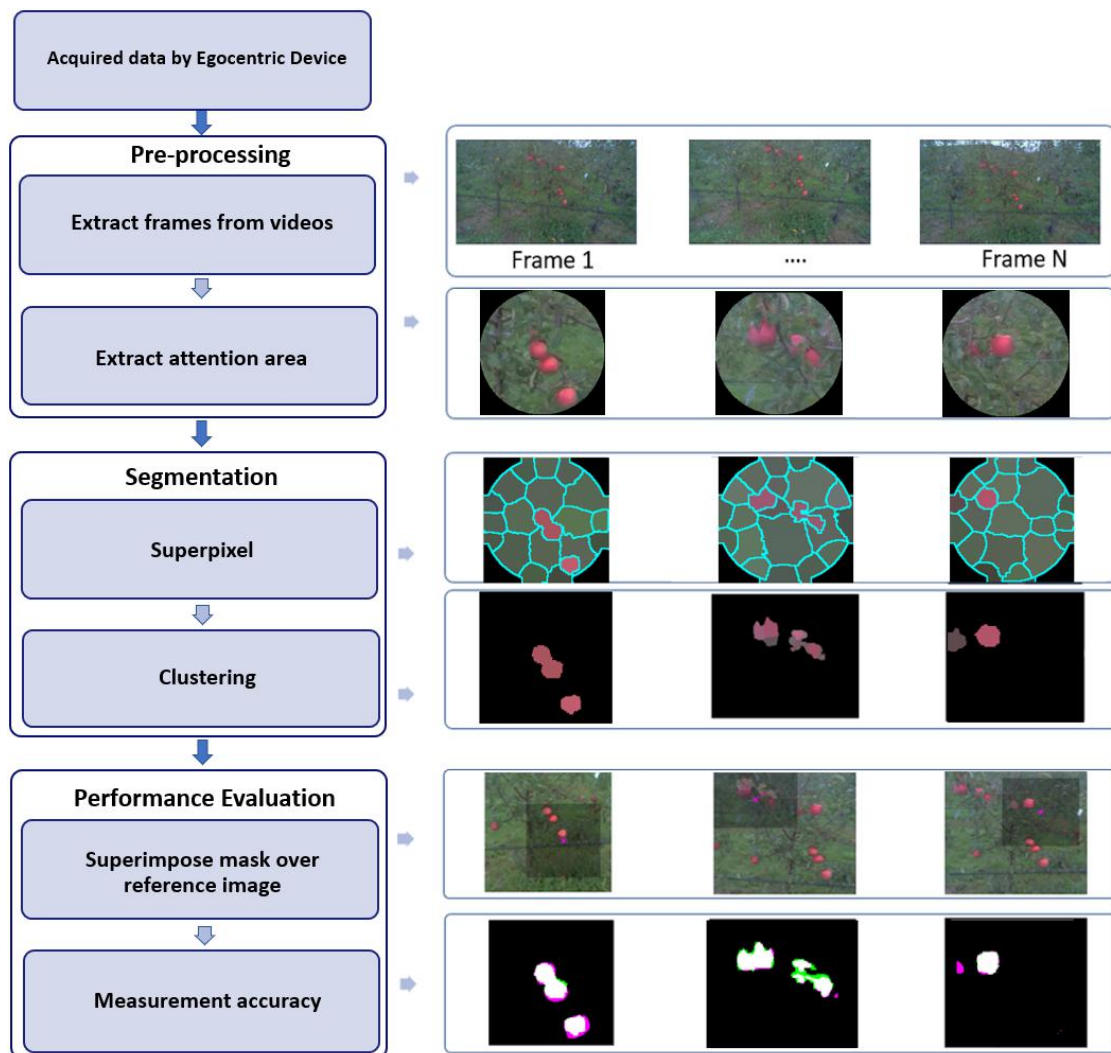


Figure 3. The three-step image processing pipeline proposed to automatically segment apples from the attention areas captured with egocentric devices.

The pre-processing started with the extraction of the frames with a resolution of 960×544 pixels from recorded videos. Next, an attention area was extracted from each frame based on egocentric priors. The extraction of this attention area constitutes the main contribution of the article. Several strategies have been tested and are presented in the next section. The pre-processed images were then segmented with a standard approach for apple detection similar to the one presented in [49,53,69–71]. A classical superpixel technique (SLIC) [72] was applied followed by a simple non-supervised clustering technique, K -means [73], to select superpixels corresponding to apples. To keep the size of superpixel independent of the size of the attention area, we defined the number of superpixels as the ratio of

$$N = \frac{A}{S}, \quad (1)$$

where A represents the size of the attention area, and S the size of an average apple, which is equal to 900 pixels in our dataset.

To simplify the images, the tree-labels (blue in our case) and sky parts were removed by applying color thresholding (optimized on a small dataset) in the RGB color domain on the superpixel segmented attention areas, as shown in Figure 4. The number of cluster K was found optimal for $K = 2$ and was applied to feature space composed of (R, G, B, H, S) respectively for red, green, brightness, hue, and saturation from each superpixel. The cluster with the smaller size was considered as the apple cluster based on the assumption that the background occupied the largest area in the attention area. Because blue parts were withdrawn and no green apples were present, the optimal value of $K = 2$ was reasonable for our use-case of apple detection in the orchard. Indeed, the local complexities in attention areas extracted from the egocentric devices were limited to objects on a background with a contrast of color. For other use-cases, where local contrast between the object and background could depend on other features (size, texture, shape, etc.), it would be necessary to adapt this segmentation.

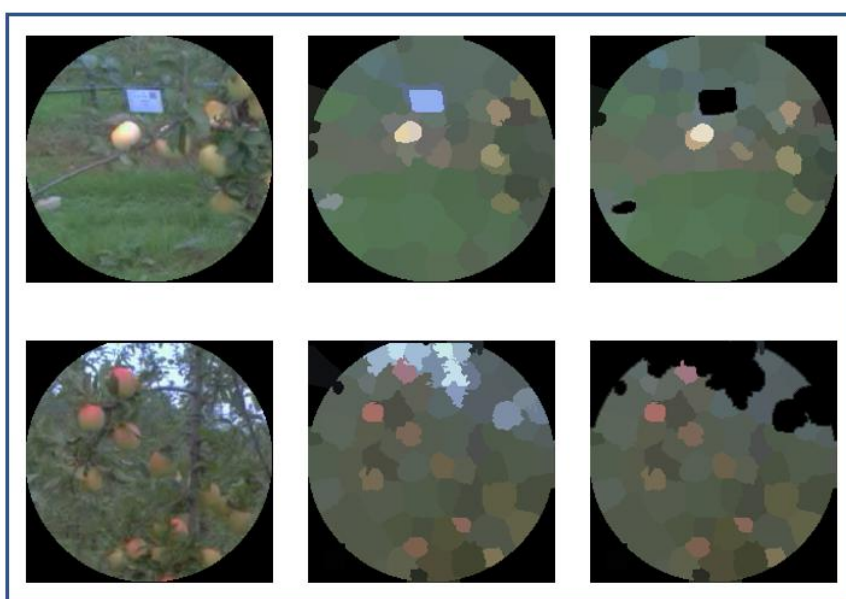


Figure 4. Color thresholding to remove blueish color belonging to the sky or blue tree-labels on superpixel segmented attention areas. Each row represents from left to right: the attention area, the superpixel segmented attention area, and the thresholded one, respectively.

Finally, the segmented apples were superimposed over the original image for qualitative assessment and localization, and compared with the manual binary ground-truth to compute the segmentation accuracy via the Dice $D_c(X, Y)$ and Jaccard index $J(X, Y)$ given by

$$D_c(X, Y) = \frac{2 * |X \cap Y|}{|X| + |Y|}, \quad (2)$$

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}, \quad (3)$$

where X and Y represent the segmented image and the ground-truth respectively.

In addition to the segmentation of apples, counting and localization were also computed in the following way. For object counting, we counted the number of connected components among detected

objects which shared sufficient overlaps with ground-truth. An empirical threshold of 75 percent was chosen for the overlap. The probability of good detection was computed as

$$PD = \frac{TP}{TP + FN}, \quad (4)$$

with TP number of true-positive objects and FN number of false-negative objects. We also computed the probability of true-negative rate as

$$TNR = \frac{TN}{TN + FP}, \quad (5)$$

with TN number of true-negative objects and FP number of false-positive objects.

In localization, the Euclidean distance between the centroid x_i of detected objects X_i and the centroid y_j of objects Y_j with a maximum intersection with ground-truth was computed as

$$d(x_i, y_j) = \sqrt{(u_{x_i} - v_{y_j})^2 + (u_{y_i} - v_{x_j})^2}, \quad (6)$$

with u and v , which stand for Cartesian coordinates in the images and

$$j = \arg \max_{j_0} |X_i \cap Y_{j_0}|. \quad (7)$$

The average distance

$$d = \frac{1}{N} \sum_{i=1}^N d(x_i, y_j), \quad (8)$$

was computed over all detected objects sharing sufficient overlap with ground-truth. Here again, a threshold of 75 percent of overlap was chosen. Distance d represents the average shift error of localization of apples with an egocentric device from manual ground-truth.

5. Strategies for Extracting Attention Area

In the following we mention different approaches for extracting attention area either using eye-tracking or not.

5.1. Attention Area from Eye-Tracking

In this section, we present strategies that we developed to extract attention areas from the eye-tracking devices to perform joint acquisition-annotation after passing these areas to the image processing pipeline of the previous section.

5.1.1. Selection by Eye-Tracking Glasses

The first approach extracted attention areas via the viewer fixation computed from the egocentric eye-tracking glasses. In order to fix a threshold, a gazing position was recorded when the same fixation position was observed during an interval of 6 frames, as calculated by

$$fi = Fps \times fd, \quad (9)$$

where fi is the frame interval, $Fps = 25$ is the number of frames per second, and fd is the average fixation duration, which was set as 275 ms. Despite careful calibration before the acquisition, small shift errors of alignment between the front camera of the device and the gazing point of the viewer can occur. Therefore, we extended the attention area around each gazing position with a given radius to compensate for the remaining small shift error of calibration of the eye-tracker. An illustration of the creation of an attention area around a fixation point is provided in Figure 5. A systematic

analysis of the evolution of the average segmentation accuracy as a function of the radius of the attention area around each gazing position was undertaken. It is shown in Figure 6 and demonstrates a non-monotonic evolution culminating at a value corresponding to triple the size of an average apple size in our dataset. Consistently, this optimal value was also found to be very close to the maximum shift error of calibration of the eye-tracker found in the whole dataset. For attention areas that are too small, due to the shift error, apples can be missed. For overly large attention areas, due to the complexity of the scene, the segmentation process fails to detect all apples correctly in the area.

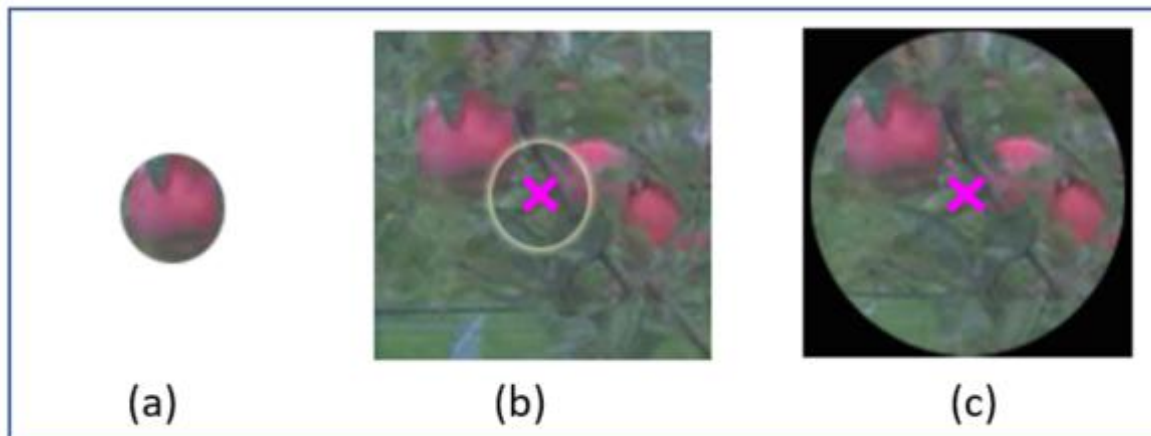


Figure 5. Construction of attention areas. (a) The average diameter of an average apple is 30 pixels in our dataset; (b) a cross indicates the center of the gaze of the annotator. There is a shift error from the apple of (a). The maximum distance of the gazing point with the center of the closest object was found at 169 pixels. (c) Chosen attention area with a size of 180×180 pixels.

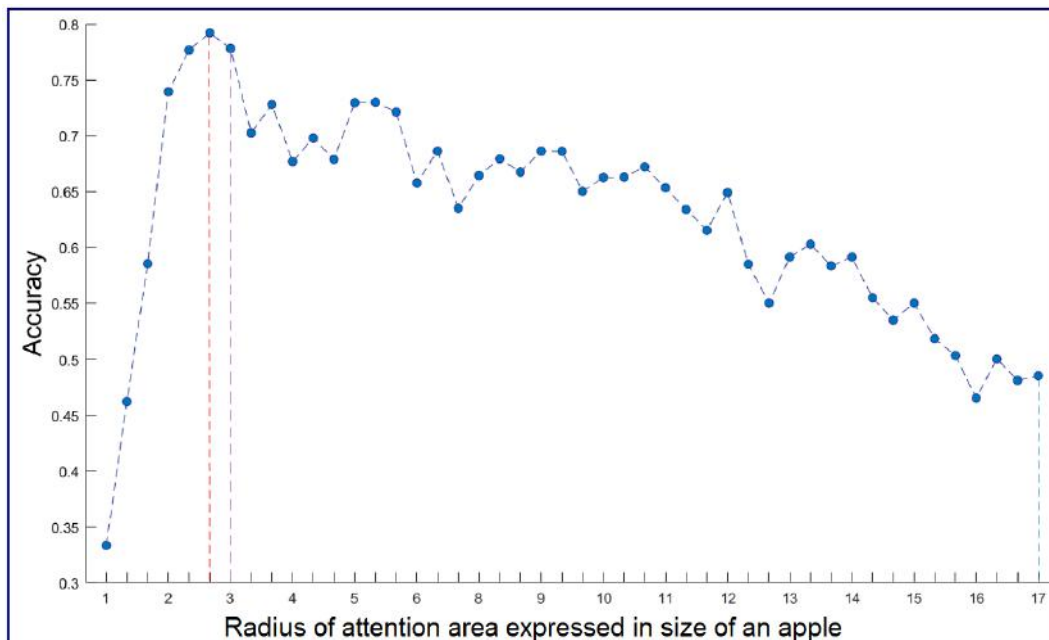


Figure 6. Apple segmentation accuracy as a function of the radius of attention area expressed in the size of apples taken as 30 pixels. Maximum accuracy achieved when the radius size of the attention map is equal to 80 (160×160 pixels) corresponding to the red dotted line. The purple dotted line corresponds to the maximum gaze shift error of (169 pixels) between eye-tracker and ground-truth when computed on the whole dataset.

5.1.2. Selection by Screen-Based Eye-Tracking

For comparison with the attention area created with the egocentric eye-tracker directly acquired in the orchard, we also generated an attention map from the gazing point recorded with a screen-based eye-tracker. Of course, this approach is less interesting for gain of time than the previous one with the head-mounted eye-tracker, since it does not allow a joint acquisition annotation. However, desktop eye-trackers are more accurate than head-mounted ones and thus are expected to constitute a reference serving as an upper bound in terms of quality of annotation with ego-centric vision. The experiment was performed on a screen with a resolution of 1920×1080 pixels while the eye movements of the viewer were recorded with an SMI binocular remote eye-tracker [74]. In this approach, for each apple tree, we peaked out one frame, which included all the apples.

The annotation protocol was the same as in the previous method. Each image was displayed to the viewer, who was asked to find the apples on the trees. The locations of the fixations of the viewer were recorded at 60 Hz. For a fair comparison, the attention area diameter around each recorded fixation was taken at the optimal value found for the eye-tracking systems embedded in glasses.

A comparison of the accuracy of the screen-based eye-tracking recording and the recording with eye-tracking embedded in glasses was conducted. Figure 7 shows that in the form of heatmap visualization of the attention of the viewer. The precision and accuracy of the produced gaze points with the screen-based eye-tracker were found to be higher than when using the head-mounted eye-tracker. The average shift error of Equation (8) was found to be 125 pixels less with the screen-based eye-tracker than with the head-mounted eye-tracker.

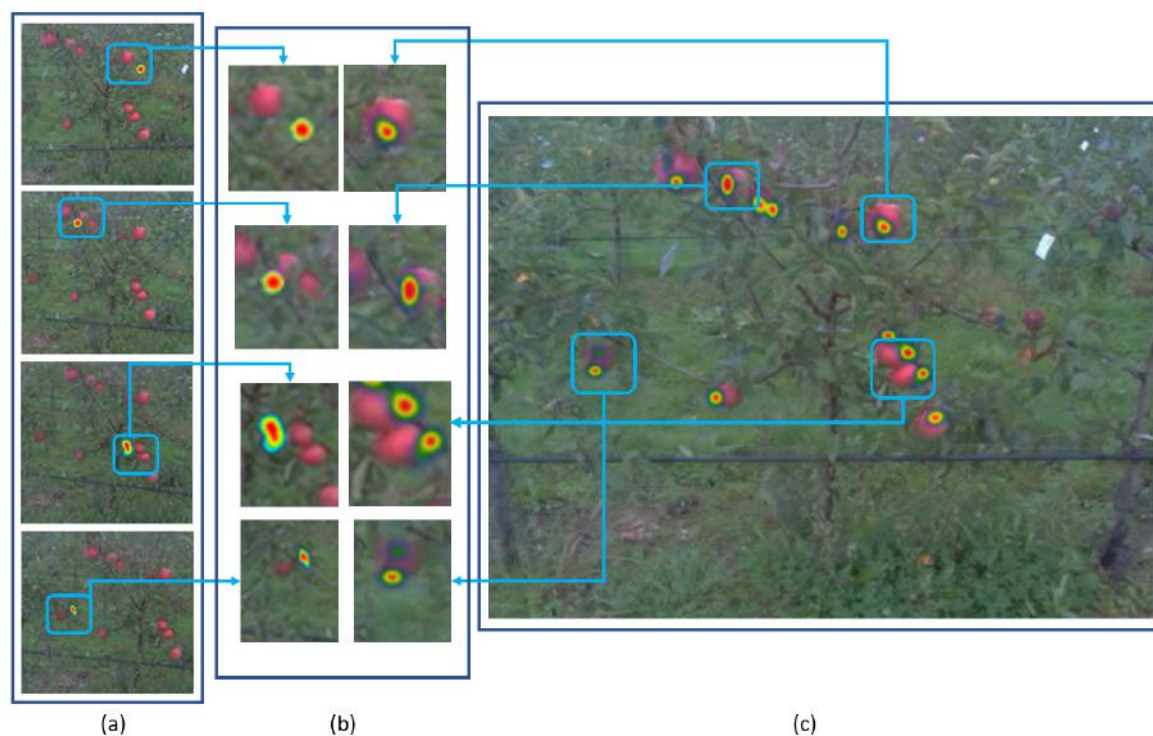


Figure 7. Heatmap visualization of the attention of the viewer captured by the head-mounted (glasses) eye-tracker (a) versus the screen-based eye-tracker (c). (b) Comparison of the heatmap generated by the glasses eye-tracker (left) vs. the heatmap generated by the screen-based eye-tracker (right).

5.2. Attention Area without Eye-Tracking

Other strategies were developed to extract attention areas for comparison with performances obtained with eye-tracking systems.

5.2.1. Full-Frame

In this approach, the attention map was considered as the full-frame recorded by the camera. Thus, in Figure 3, instead of a small patch of the entire original image, the full original image was directly transmitted to the superpixel segmentation. Such a choice assumes that the camera field of view is already a focus of the overall field of interest for the human annotator in charge of detecting apples.

5.2.2. Egocentric Prior

In this approach, we assumed, as is often done in egocentric vision [16], that the attention of the viewer was focused at the center of the frame. Therefore, we selected the attention area as a disk positioned at the center of the image with the size of 180×180 pixels for a fair comparison with the other approaches developed for eye-trackers.

5.2.3. Saliency Map

As the last method to compute an attention area, we turned toward a computational approach in charge of numerically identifying areas of interest. Such a concept has been developed in the computer vision literature under the name of the saliency map. Saliency acts as a local filter that enhances regions of the image which stand out relative to their adjacent parts in terms of orientation and/or gray level and/or color contrast [75]. Introduced in [76], saliency was inspired by the mechanisms of human visual attention and the fixation behavior of the observer. There are numerous computational models for salient object detection. In this study, for illustration and without any claim of optimality, we used the algorithm proposed by [77], which computes saliency map in images using low-level features and was proposed with codes included for reproducible science. Saliency maps were thresholded to binary masks following the fixed threshold procedure described in [77]. Each connected component of the binary saliency map served to produce an attention area. For a fair comparison with the other approaches, attention areas of size 180×180 pixels were chosen.

6. Results and Discussion

We are now ready to compare the results of the different approaches proposed for apple detection by extracting attention areas through egocentric vision in the perspective of a joint acquisition-annotation process. As shown in Table 1, we assessed the image annotation quality by the same image segmentation pipeline of Section 4 (depicted in Figure 3). Comparison is provided between the five different approaches presented in Section 5 for the extraction of attention areas from egocentric devices. In terms of segmentation, accuracy was estimated by the Dice Equation (2) and Jaccard Equation (3) indexes. The probability of good detection indicates the true counted apples computed by Equation (4). The true-negative rate Equation (5) represents the proportion of actual negatives that are correctly identified. The next column in Table 1 specifies the error of localization of detected apples computed by Equation (8). Time is the approximate consumed execution time (automatic annotation) acquired from each approach of the whole dataset. Finally, the time gain indicates the ratio of manual annotation time over the consumed execution time obtained from each automatic annotation approach. All these experimental results correspond to an average of over 10 different trees available in the dataset.

The best average performances (highlighted in bold in Table 1) in terms of segmentation accuracy of apples were obtained with the eye-tracking-based methods. Challenging images and resulting annotations with eye-tracking-based methods are provided in Figure 8 for qualitative assessment. Overall, the screen-based eye-tracker provided the best result but only slightly above the one obtained from the glasses eye-tracker. This embedded glasses eye-tracker, despite its substantial shift errors, had a high value since it enabled joint image acquisition and annotation. The saliency approach provided a result close to the one obtained with the baseline method (full-frame). This could certainly be improved with a systematic benchmark of other saliency methods of the literature. However,

a fundamental reason for the failure of the saliency approach, which would be common to all generic saliency maps, is that saliency is, so to say, attracted by contrasting objects which may not be apples (for example, stems, leaves, items in the background, a data matrix positioned in the field to identify trees). As a consequence, saliency creates many true-negatives in attention areas since the task of detecting apples does not specifically drive it. In contrast, human attention focuses on the apple as captured by eye-tracking systems.

Interestingly these results were consistent for the three tasks assessed: segmentation, counting, and localization. This demonstrates the robustness of the interest of eye-tracker devices for annotation. Eye-tracking systems, such as the two used in this study, can be considered as expensive devices (typically between 10,000 and 20,000 euros currently). It is interesting to see that the egocentric prior approach gave the third-best performance, and this could be accessible with any camera embedded on glasses (for 10 to 100 euros).

Table 1. Performance of apple detection with the five approaches developed for automatic apple annotation in the attention area captured by the egocentric devices. Each column corresponds to an average over the 10 trees of the dataset. Dice and Jaccard assess in percentage the quality of segmentation via Equations (2) and (3); good prediction and true-negative rate assess in percentage the quality of object detection via Equations (4) and (5); and the shift error of Equation (8) assesses in pixels the quality of good localization. The time corresponds to the approximate execution time for automatic annotation for the whole dataset in seconds. Time gain indicates the ratio of manual annotation time (6 h) over automatic annotation time obtained from each approach. Time was measured on a windows machine with an Intel Xeon CPU and 32.0 GB RAM by Matlab 2017a.

Method (Section)	Dice	Jaccard	Good Detection	True-Negative Rate	Shift Error	Time (Second)	Time Gain
Full-Frame (Section 5.2.1)	0.24 ± 0.22	0.21 ± 0.16	0.31 ± 0.20	0.17 ± 0.72	174.11 ± 34	880	24
Glasses Eye-tracker (Section 5.1)	0.78 ± 0.08	0.64 ± 0.08	0.84 ± 0.16	0.09 ± 0.07	15.97 ± 11	1960	11
Screen-based Eye-tracker (Section 5.1.2)	0.85 ± 0.09	0.77 ± 0.13	0.88 ± 0.12	0.09 ± 0.13	2.37 ± 1.86	3240	6
Egocentric Prior (Section 5.2.2)	0.46 ± 0.36	0.38 ± 0.31	0.54 ± 0.39	0.28 ± 0.23	84.82 ± 7.25	1960	11
Saliency (Section 5.2.3)	0.27 ± 0.13	0.16 ± 0.08	0.42 ± 0.45	0.51 ± 0.17	7.21 ± 8.28	2358	9

The values of the obtained results in terms of segmentation, counting, and localization were also assessed in terms of timing. As expressed in Section 3.1, acquisition time with an egocentric device is comparable with acquisition time with any standard camera. Therefore gains of time were compared regarding the annotation time only. This timing is provided in the last column of Table 1 for automatic annotation based on the image processing pipeline applied to extracted attention areas. Without any surprise, the full-frame approach, which requires no computation of attention map, is the fastest method. The second most rapid methods are the egocentric prior and glasses eye-tracker. The screen-based eye-tracker method, which gave the best performance in terms of apple detection, came with the slowest timing. However, these timings for automated annotation are to be compared with the timing requested by a human annotator to manually annotate all apples in the dataset. The estimated timing was 6 h for the 419 frames. The gain of time for all methods is presented in Table 1. Saliency, as presented here, could be criticized since many other variants of the saliency map could be tested and possibly provide better results. In terms of timing, however, we believe the performances are realistic, and it was worth mentioning them here. All in all, the glasses eye-tracker method appears to be a good trade-off between speed and annotation performance (as summarized in Table 2). For this head-mounted device, the gain in performance was about 11 times, which is smaller than what was found in the closest related work with desktop eye-trackers for object detection [8,31,32]. This difference may come from the fact that in this literature, the tasks targeted were relatively more straightforward and required less post-processing. Optimization of the code could thus increase the gain in time. We are currently investigating all those perspectives.

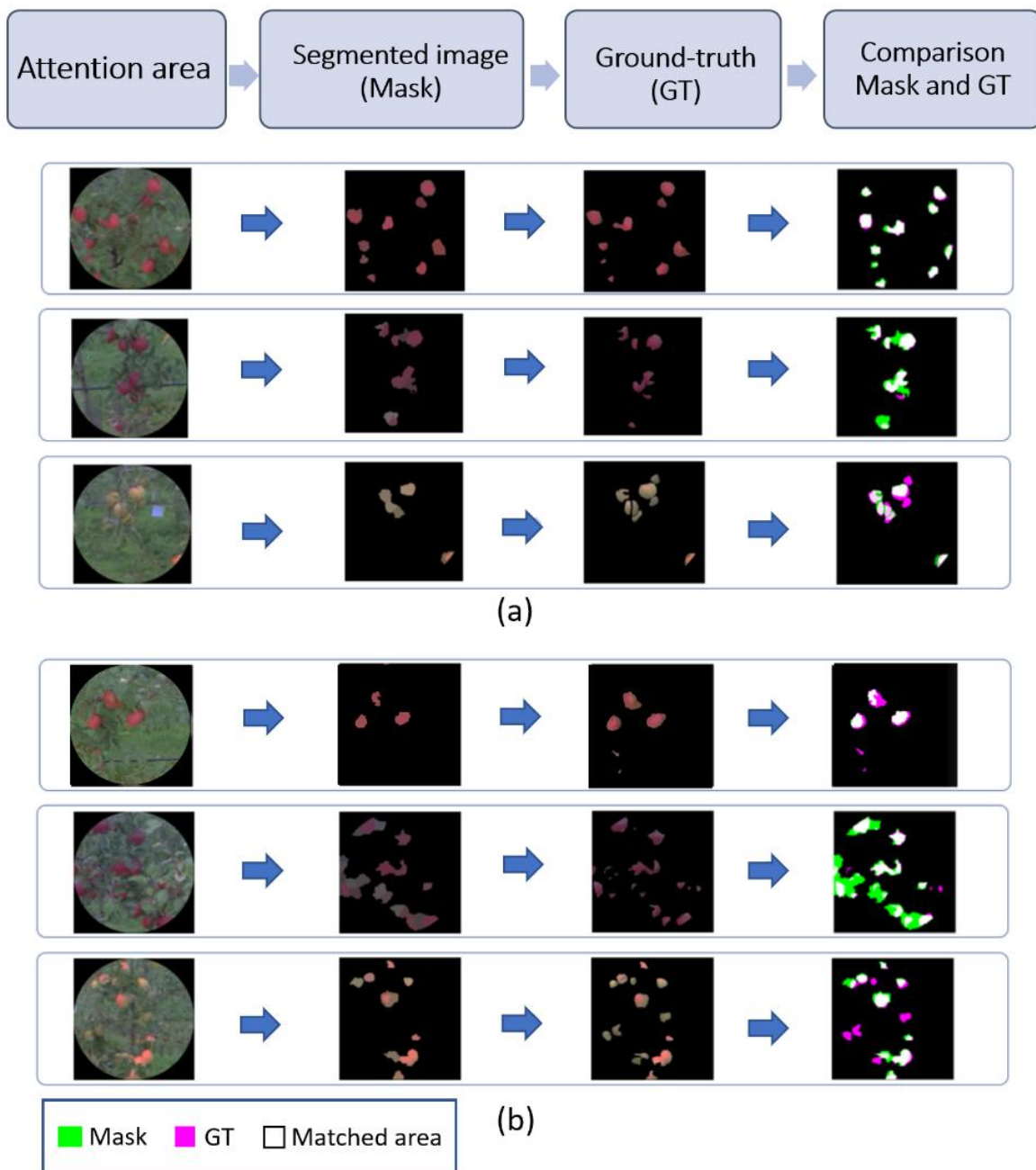


Figure 8. Qualitative assessment of results. From left to right, an example of the attention area captured by eye-tracking, automatic annotation obtained from the proposed image processing pipeline of Figure 3, ground-truth manually recorded, and comparison of manual ground-truth and automatic segmentation. (a) Examples of good performance; (b) Some challenging conditions wherein more errors were found (missed detection, false detection).

Table 2. Qualitative summary of the five uses of egocentric devices compared in this study.

Method	Joint Acquisition Annotation	Fastest Execution Time	Best Annotation	Best Counting	Best Localization
Full-Frame	+	+	-	-	-
Glasses Eye-tracker	+	-	+	+	-
Screen-based Eye-tracker	-	-	+	+	+
Egocentric Prior	+	-	-	-	-
Saliency	+	-	-	-	+

7. Conclusions

We have assessed the value of egocentric imaging devices to jointly perform acquisition and automatic image annotation. This was illustrated with apple detection in orchards, which is known to be a challenging task for computer vision applied to phenotyping or agriculture. Despite shift errors in the calibration of egocentric imaging devices, the performance of the detection of apples from the gazed recorded areas was found to be very close to the one obtained from the manual annotation. The compensation for these shift errors was obtained by applying a standard non-supervised segmentation algorithm only applied in attention areas centered on the gazing positions captured by the egocentric devices. Specific interest was shown for head-mounted eye-tracking systems with an estimated gain of time in comparison with manual annotation of 11 times with non-GPU-accelerated software.

This first use of egocentric vision to speed up image annotation opens up interesting perspectives, especially in plant phenotyping. The task here was focused on apples, but the approach is in fact generic. Thus, it would be interesting to extend the applicability to other phenotyping items of interest. The non-supervised image segmentation algorithm applied in gazed areas was purposely chosen simply in this article to demonstrate the value of the eye-tracking device. It is interesting to notice that performances obtained with this simple algorithm were already interesting quantitatively and qualitatively. The literature of non-supervised image segmentation with superpixels is huge [78,79], and it would be interesting to revisit more exhaustively this literature for the segmentation of gazed areas. Specific attention could focus on the methods addressing the limitation of superpixels [80], also observed in this article, with "leakage" of boundaries in the vicinity of the targeted objects [81]. To remain on the topic of apples, this could include the determination of flowering stages or the detection of diseases. Additional technological services from egocentric vision could be tested to speed up annotation. For instance, this includes the use of sound recording, which could be coupled to automatic speech recognition for later fusion with information extracted from the captured images. The pilot study presented here is promising. For a tool to be used by technicians and engineers in the field, it would be necessary to implement an ergonomic version of the software to experiment on a large network of users the method developed to accelerate image annotation with egocentric devices. Validation of the quality of the annotation was performed at various levels, including location, object detection, and pixel-wise segmentation. Another stage of validation of the quality of the annotation would be to train a machine learning algorithm on the annotated images and compare the performance with the manually annotated data.

Author Contributions: S.S. and D.R. conceived and implemented the work. S.S. and P.R. (Pejman Rasti) performed image acquisition. G.G. and P.R. (Paul Richard) contributed to the management and administration of the study. S.S. and D.R. wrote and revised the manuscript. All authors validated the final version of the manuscript.

Acknowledgments: Authors gratefully acknowledge the Région des Pays de la Loire for funding this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kamlaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [[CrossRef](#)]
2. Benoit, L.; Rousseau, D.; Belin, É.; Demilly, D.; Chapeau-Blondeau, F. Simulation of image acquisition in machine vision dedicated to seedling elongation to validate image processing root segmentation algorithms. *Comput. Electron. Agric.* **2014**, *104*, 84–92. [[CrossRef](#)]
3. Giuffrida, M.V.; Scharr, H.; Tsaftaris, S.A. ARIGAN: Synthetic arabidopsis plants using generative adversarial network. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW 2017), Venice, Italy, 22–29 October 2017; pp. 2064–2071. [[CrossRef](#)]

4. Peter, L.; Mateus, D.; Chatelain, P.; Declara, D.; Schworm, N.; Stangl, S.; Multhoff, G.; Navab, N. Assisting the examination of large histopathological slides with adaptive forests. *Med. Image Anal.* **2017**, *35*, 655–668. [[CrossRef](#)] [[PubMed](#)]
5. Giuffrida, M.V.; Chen, F.; Scharr, H.; Tsaftaris, S.A. Citizen crowds and experts: Observer variability in image-based plant phenotyping. *Plant Methods* **2018**, *14*, 12. [[CrossRef](#)] [[PubMed](#)]
6. Barth, R.; Ijsselmuiden, J.; Hemming, J.; Henten, E.V. Data synthesis methods for semantic segmentation in agriculture: A Capsicum annum dataset. *Comput. Electron. Agric.* **2018**, *144*, 284–296. [[CrossRef](#)]
7. Douarre, C.; Schielein, R.; Frindel, C.; Gerth, S.; Rousseau, D. Transfer learning from synthetic data applied to soil–root segmentation in X-ray tomography images. *J. Imaging* **2018**, *4*, 65. [[CrossRef](#)]
8. Samiei, S.; Ahmad, A.; Rasti, P.; Belin, E.; Rousseau, D. Low-cost image annotation for supervised machine learning. Application to the detection of weeds in dense culture. In *British Machine Vision Conference (BMVC), Computer Vision Problems in Plant Phenotyping (CVPPP)*; BMVA Press: Newcastle, UK, 2018; p. 1.
9. Douarre, C.; Crispim-Junior, C.F.; Gelibert, A.; Tougne, L.; Rousseau, D. Novel data augmentation strategies to boost supervised segmentation of plant disease. *Comput. Electron. Agric.* **2019**, *165*, 104967. [[CrossRef](#)]
10. Hung, C.; Nieto, J.; Taylor, Z.; Underwood, J.; Sukkarieh, S. Orchard fruit segmentation using multi-spectral feature learning. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 3–7 November 2013; pp. 5314–5320. [[CrossRef](#)]
11. Ubbens, J.; Cieslak, M.; Prusinkiewicz, P.; Stavness, I. The use of plant models in deep learning: An application to leaf counting in rosette plants. *Plant Methods* **2018**, *14*, 6. [[CrossRef](#)]
12. Fathi, A.; Farhadi, A.; Rehg, J.M. Understanding egocentric activities. In *Proceedings of the IEEE International Conference on Computer Vision*, Barcelona, Spain, 6–13 November 2011; pp. 407–414. [[CrossRef](#)]
13. Doherty, A.R.; Caprani, N.; Conaire, C.Ó.; Kalnikaite, V.; Gurrin, C.; Smeaton, A.F.; O'Connor, N.E. Passively recognising human activities through lifelogging. *Comput. Hum. Behav.* **2011**, *27*, 1948–1958. [[CrossRef](#)]
14. Pirsivash, H.; Ramanan, D. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 16–21 June 2012; pp. 2847–2854. [[CrossRef](#)]
15. Lu, Z.; Grauman, K. Story-driven summarization for egocentric video. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Portland, OR, USA, 23–28 June 2013; pp. 2714–2721. [[CrossRef](#)]
16. Fathi, A.; Ren, X.; Rehg, J.M. Learning to recognize objects in egocentric activities. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 20–25 June 2011; pp. 3281–3288. [[CrossRef](#)]
17. Erculiani, L.; Giunchiglia, F.; Passerini, A. Continual egocentric object recognition. *Comput. Vis. Pattern Recognit.* **2019**, arXiv:1912.05029v2.
18. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [[CrossRef](#)]
19. Rituerto, A. Modeling the environment with egocentric vision systems. *Electron. Lett. Comput. Vis. Image Anal.* **2015**, *14*, 49–51. [[CrossRef](#)]
20. Alletto, S.; Serra, G.; Calderara, S.; Cucchiara, R. Understanding social relationships in egocentric vision. *Pattern Recognit.* **2015**, *48*, 4082–4096. [[CrossRef](#)]
21. Betancourt, A.; Morerio, P.; Regazzoni, C.S.; Rauterberg, M. The Evolution of First Person Vision Methods: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 744–760. [[CrossRef](#)]
22. Liu, K.Y.; Hsu, S.C.; Huang, C.L. First-person-vision-based driver assistance system. In *Proceedings of the 2014 International Conference on Audio, Language and Image Processing*, Shanghai, China, 7–9 July 2014; pp. 239–244. [[CrossRef](#)]
23. Mayol, W.W.; Davison, A.J.; Tordoff, B.J.; Murray, D.W. Applying active vision and SLAM to wearables. In *Springer Tracts in Advanced Robotics*; Dario, P., Chatila, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; Volume 15, pp. 325–334. [[CrossRef](#)]
24. Karaman, S.; Benois-Pineau, J.; Mégret, R.; Dovgalecs, V.; Dartigues, J.F.; Gaëstel, Y. Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *Proceedings of the International Conference on Pattern Recognition*, Istanbul, Turkey, 23–26 August 2010; pp. 4113–4116. [[CrossRef](#)]

25. Doherty, A.R.; Hodges, S.E.; King, A.C.; Smeaton, A.F.; Berry, E.; Moulin, C.J.; Lindley, S.; Kelly, P.; Foster, C. Wearable cameras in health: The state of the art and future possibilities. *Am. J. Prev. Med.* **2013**, *44*, 320–323. [[PubMed](#)]
26. Li, Y.; Fathi, A.; Rehg, J.M. Learning to predict gaze in egocentric video. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3216–3223. [[CrossRef](#)]
27. Li, C.; Kitani, K.M. Pixel-level hand detection in ego-centric videos. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3570–3577. [[CrossRef](#)]
28. Bambach, S.; Lee, S.; Crandall, D.J.; Yu, C. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; Volume 2015, pp. 1949–1957. [[CrossRef](#)]
29. Ma, M.; Fan, H.; Kitani, K.M. Going Deeper into First-Person Activity Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1894–1903. [[CrossRef](#)]
30. Tatler, B.W.; Baddeley, R.J.; Gilchrist, I.D. Visual correlates of fixation selection: Effects of scale and time. *Vis. Res.* **2005**, *45*, 643–659. [[CrossRef](#)]
31. Walber, T. Making use of eye tracking information in image collection creation and region annotation. In Proceedings of the 20th ACM International Conference on Multimedia (MM 2012), Nara, Japan, 29 October–2 November 2012; ACM Press: New York, NY, USA, 2012; pp. 1405–1408. [[CrossRef](#)]
32. Lucas, A.; Wang, K.; Santillan, C.; Hsiao, A.; Sirlin, C.B.; Murphy, P.M. Image Annotation by Eye Tracking: Accuracy and Precision of Centerlines of Obstructed Small-Bowel Segments Placed Using Eye Trackers. *J. Digit. Imaging* **2019**, *32*, 855–864. [[CrossRef](#)]
33. Parrish, E.A.; Goksel, A.K. Pictorial Pattern Recognition Applied To Fruit Harvesting. *Trans. Am. Soc. Agric. Eng.* **1977**, *20*, 822–827. [[CrossRef](#)]
34. D’Grand, E.; Rabatel, A.G.; Pellenc, R.; Journeau, A.; Aldon, M.J. Magali: A self-propelled robot to pick apples. *Am. Soc. Agric. Eng. Pap.* **1987**, *46*, 353–358.
35. Whittaker, A.D.; Miles, G.E.; Mitchell, O.R.; Gaultney, L.D. Fruit Location in a Partially Occluded Image. *Trans. Am. Soc. Agric. Eng.* **1987**, *30*, 591–596. [[CrossRef](#)]
36. Slaughter, D.C.; Harrell, R.C. Color vision in robotic fruit harvesting. *Trans. ASAE* **1987**, *30*, 1144–1148. [[CrossRef](#)]
37. Sites, P.W.; Delwiche, M.J. Computer Vision To Locate Fruit on a Tree. *Trans. Am. Soc. Agric. Eng.* **1988**, *31*, 257–263, 272. [[CrossRef](#)]
38. Rabatel, G. A vision system for Magali, the fruit picking robot. In Proceedings of the International Conference on Agricultural Engineering, Paris, France, 2–5 March 1988.
39. Kassay, L. Hungarian robotic apple harvester. In Proceedings of the ASAE Annual Meeting Papers, Charlotte, NC, USA, 21–24 June 1992.
40. Ceres, R.; Pons, J.; Jimenez, A.; Martin, J.; Calderon, L. Agrirobot : A Robot for Aided Fruit Harvesting. *Ind. Robot.* **1998**, *25*, 337–46. [[CrossRef](#)]
41. Jiménez, A.R.; Ceres, R.; Pons, J.L.; Jimenez, A.R.; Ceres, R.; Pons, J.L. A survey of computer vision methods for locating fruit on trees. *Trans. Am. Soc. Agric. Eng.* **2000**, *43*, 1911–1920. [[CrossRef](#)]
42. Zhou, R.; Damerow, L.; Sun, Y.; Blanke, M.M. Using colour features of cv. ‘Gala’ apple fruits in an orchard in image processing to predict yield. *Precis. Agric.* **2012**, *13*, 568–580. [[CrossRef](#)]
43. Song, Y.; Glasbey, C.A.; Horgan, G.W.; Polder, G.; Dieleman, J.A.; van der Heijden, G.W. Automatic fruit recognition and counting from multiple images. *Biosyst. Eng.* **2014**, *118*, 203–215. [[CrossRef](#)]
44. Sa, I.; Ge, Z.; Dayoub, F.; Upcroft, B.; Perez, T.; McCool, C. Deepfruits: A fruit detection system using deep neural networks. *Sensors* **2016**, *16*, 1222. [[CrossRef](#)]
45. Boogaard, F.P.; Rongen, K.S.; Kootstra, G.W. Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging. *Biosyst. Eng.* **2020**, *192*, 117–132. [[CrossRef](#)]
46. Wang, Q.; Nuske, S.; Bergerman, M.; Singh, S. Automated Crop Yield Estimation for Apple Orchards. In *Experimental Robotics*; Springer Tracts in Advanced Robotics; Springer: Berlin/Heidelberg, Germany, 2013; pp. 745–758. [[CrossRef](#)]

47. Hung, C.; Underwood, J.; Nieto, J.; Sukkarieh, S. A feature learning based approach for automated fruit yield estimation. In *Springer Tracts in Advanced Robotics*; Mejias, L., Corke, P., Roberts, J., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 105, pp. 485–498. [CrossRef]
48. Bargoti, S.; Underwood, J. Image classification with orchard metadata. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; Volume 2016; pp. 5164–5170. [CrossRef]
49. Bargoti, S.; Underwood, J.P. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. *J. Field Robot.* **2017**, *34*, 1039–1060. [CrossRef]
50. Zhong, G.; Huang, K. *Semi-Supervised Learning: Background, Applications and Future Directions*; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2018.
51. Pise, N.N.; Kulkarni, P. A Survey of Semi-Supervised Learning Methods. In Proceedings of the 2008 International Conference on Computational Intelligence and Security, Suzhou, China, 13–17 December 2008; Volume 2, pp. 30–34.
52. Zhu, X.J. *Semi-Supervised Learning Literature Survey*; Technical Report; University of Wisconsin-Madison Department of Computer Sciences: Madison, WI, USA, 2005.
53. Roy, P.; Kislak, A.; Plonski, P.A.; Luby, J.; Isler, V. Vision-based preharvest yield mapping for apple orchards. *Comput. Electron. Agric.* **2019**, *164*, 104897. [CrossRef]
54. Goldberger, J.; Gordon, S.; Greenspan, H. An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures. In Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 1, pp. 487–493.
55. Blihnaut, P. Fixation identification: The optimum threshold for a dispersion algorithm. *Atten. Percept. Psychophys.* **2009**, *71*, 881–895. [CrossRef]
56. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef]
57. Jacob, R.J.K. What You Look at is What You Get: Eye Movement-Based Interaction Techniques. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 90), Seattle, WA, USA, 1–5 April 1990; Association for Computing Machinery: New York, NY, USA, 1990; pp. 11–18. [CrossRef]
58. Irwin, D.E. *Eye Movements and Visual Cognition: Scene Perception and Reading*; Springer: New York, NY, USA, 1992; pp. 146–165. [CrossRef]
59. Jacob, R.J.K. Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. *Adv. Hum. Comput. Interact.* **2003**, *4*, 151–190.
60. Salvucci, D.D.; Goldberg, J.H. Identifying Fixations and Saccades in Eye-Tracking Protocols. In Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA'00), Palm Beach Gardens, FL, USA, 6–8 November 2000; Association for Computing Machinery: New York, NY, USA, 2000; pp. 71–78. doi:10.1145/355017.355028. [CrossRef]
61. Manor, B.R.; Gordon, E. Defining the temporal threshold for ocular fixation in free-viewing visuo-cognitive tasks. *J. Neurosci. Methods* **2003**, *128*, 85–93. [CrossRef]
62. Duchowski, A. *Eye Tracking Methodology*; Springer: London, UK, 2007. [CrossRef]
63. Shic, F.; Scassellati, B.; Chawarska, K. The incomplete fixation measure. In Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, Savannah, GA, USA, 26–28 March 2008; ACM Press: New York, NY, USA, 2008; p. 111. [CrossRef]
64. Spakov, O.; Miniotas, D. Application of Clustering Algorithms in Eye Gaze Visualizations. Available online: <https://pdfs.semanticscholar.org/b016/02b60a1fcb1ca06f6af0d4273a6336119bae.pdf> (accessed on 21 June 2020).
65. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2016**, arXiv:1612.03928.
66. Safren, O.; Alchanatis, V.; Ostrovsky, V.; Levi, O. Detection of green apples in hyperspectral images of apple-tree foliage using machine vision. *Trans. ASABE* **2007**, *50*, 2303–2313. [CrossRef]
67. Gené-Mola, J.; Vilaplana, V.; Rosell-Polo, J.R.; Morros, J.R.; Ruiz-Hidalgo, J.; Gregorio, E. KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data. *Data Brief* **2019**, *25*, 104289. [CrossRef] [PubMed]

68. Hani, N.; Roy, P.; Isler, V.; Hani, N.; Roy, P.; Isler, V. MinneApple: A Benchmark Dataset for Apple Detection and Segmentation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 852–858. [[CrossRef](#)]
69. Kang, H.; Chen, C. Fruit detection and segmentation for apple harvesting using visual sensor in orchards. *Sensors* **2019**, *19*, 4599. [[CrossRef](#)]
70. Liu, X.; Jia, W.; Ruan, C.; Zhao, D.; Gu, Y.; Chen, W. The recognition of apple fruits in plastic bags based on block classification. *Precis. Agric.* **2018**, *19*, 735–749. [[CrossRef](#)]
71. Liu, X.; Zhao, D.; Jia, W.; Ji, W.; Sun, Y. A Detection Method for Apple Fruits Based on Color and Shape Features. *IEEE Access* **2019**, *7*, 67923–67933. [[CrossRef](#)]
72. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2281. [[PubMed](#)]
73. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28*, 100. [[CrossRef](#)]
74. Sahin, A. SensoMotoric Instruments launches SMI Eye Tracking. Available online: https://en.wikipedia.org/wiki/SensoMotoric_Instruments (accessed on 21 June 2020).
75. Achanta, R.; Estrada, F.; Wils, P.; Süsstrunk, S. Salient region detection and segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; ICVS 2008; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5008; pp. 66–75. [[CrossRef](#)]
76. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [[CrossRef](#)]
77. Achanta, R.; Hemami, S.; Estrada, F.; Süsstrunk, S. Frequency-tuned salient region detection. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604. [[CrossRef](#)]
78. Wang, C.; Chen, J.; Li, W. Review on superpixel segmentation algorithms. *Appl. Res. Comput.* **2014**, *31*, 6–12.
79. Wang, M.; Liu, X.; Gao, Y.; Ma, X.; Soomro, N.Q. Superpixel segmentation: A benchmark. *Signal Process. Image Commun.* **2017**, *56*, 28–39. [[CrossRef](#)]
80. Stutz, D.; Hermans, A.; Leibe, B. Superpixels: An evaluation of the state-of-the-art. *Comput. Vis. Image Underst.* **2018**, *166*, 1–27. [[CrossRef](#)]
81. Levinshtein, A.; Stere, A.; Kutulakos, K.N.; Fleet, D.J.; Dickinson, S.J.; Siddiqi, K. Turbopixels: Fast superpixels using geometric flows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2290–2297. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: <http://www.elsevier.com/locate/complbiomed>

Simulated perfusion MRI data to boost training of convolutional neural networks for lesion fate prediction in acute stroke

Noëlie Debs^a, Pejman Rasti^b, Léon Victor^a, Tae-Hee Cho^a, Carole Frindel^a, David Rousseau^{b,*}

^a CREATIS, CNRS UMR-5220, INSERM U1206, Université Lyon 1, INSA Lyon Bât, Blaise Pascal, 7 Avenue Jean Capelle, 69621, Villeurbanne, France

^b Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRA IRHS, Université d'Angers, 62 Avenue Notre Dame du Lac, 49000 Angers, France

ARTICLE INFO

Keywords:

Stroke
Lesion prediction
Perfusion magnetic resonance imaging
Arterial input function
Simulation
Convolutional neural network

ABSTRACT

The problem of final tissue outcome prediction of acute ischemic stroke is assessed from physically realistic simulated perfusion magnetic resonance images. Different types of simulations with a focus on the arterial input function are discussed. These simulated perfusion magnetic resonance images are fed to convolutional neural network to predict real patients. Performances close to the state-of-the-art performances are obtained with a patient specific approach. This approach consists in training a model only from simulated images tuned to the arterial input function of a tested real patient. This demonstrates the added value of physically realistic simulated images to predict the final infarct from perfusion.

1. Introduction

Stroke is a major cause of mortality and disability in the world [1]. Stroke is divided into ischemic (85%) and hemorrhagic types (15%) [2]. Ischemia occurs when a cerebral artery is occluded [3]. Neuroimaging in acute stroke aims to obtain rapid information on tissue and vessel status to aid acute stroke intervention [4]. Diagnosis obtained from modern neuroimaging modalities enables efficient management of ischemic stroke and decide whether patient may benefit from intravenous thrombolysis or mechanical thrombectomy [5,6]. The most common neuroimaging, due to its widespread immediate availability, is computed tomography (CT), which is used in the initial diagnosis to determine the type of stroke (ischemic or hemorrhagic) [7]. Magnetic resonance imaging (MRI) may be substituted for CT as it becomes more readily available and it provides greater physiological information on soft tissues. MRI imaging for acute stroke include diffusion-weighted imaging (DWI) and perfusion-weighted imaging (PWI) [4]. DWI allows early detection of an infarcted lesion within minutes of a stroke by quantifying motion of water molecules: restriction in the diffusional movement of water is subsequent to cytotoxic edema. PWI is a functional brain imaging modality which requires the administration of an intravenous bolus of a contrast agent and provides information on the hemodynamic status of the tissues. The combination of DWI and PWI is commonly used in clinical practice to evaluate the extent of irreversible tissue damage (). In the treatment decision context, this information

helps physicians to identify acute stroke patients that could benefit from reperfusion therapies.

Still, developing automated methods to predict the extent of the stroke lesion from MRI scans remains an open challenge [8]. This prediction has been mainly addressed so far with thresholded hemodynamic biomarkers based on kinetic models [9–11]. However, given the high dimensionality of PWI, machine learning techniques [12] have also been successfully proposed in recent years [13–21]. A limitation to the use of supervised machine learning is the limited amount of data. This lack of data is mainly due to the poor quality of the clinical datasets (corrupted or missing images), the insufficient amount of labelled data (current datasets limited to a few hundred patients) and the imbalance between classes (more pixels healthy than pathological pixels). This can be considered as a bottleneck specially when using highly discriminating algorithms depending on a large number of parameters. A way to circumvent this limitation is to generate more data from simulation and image synthesis model [22–24]. Data augmentation is also a way to improve regularization and reduce overfitting by injecting more prior information into the training dataset [25].

In this article, we assess the interest of simulation for the prediction of the fate of acute stroke lesion. This prediction is undertaken here with deep learning on convolutional neural networks (CNN) since they are known to have, by comparison with the classical shallow learning techniques (support vector machines, random forests ...), higher amount of parameters to be tuned and can produce the best performances but

* Corresponding author.

E-mail address: david.rousseau@univ-angers.fr (D. Rousseau).

also are the most likely to benefit from data augmentation.

Deep learning has been applied in stroke in different contexts including prediction from perfusion imaging or with other MRI modalities, for whole tissue segmentation or voxel-based classification and also with more or less complex neural network architectures [14–18]. In our case, we apply a standard architecture of CNN learning at the voxel level for binary classification of the fate of the tissue (*i.e.* tissue expansion or regression) from spatio-temporal data. Neural networks usually require some data augmentation which is conventionally done in deep learning with spatial distortions likely to occur in nature [26]. In our case, due to the temporal dimension of the data, usual spatial distortions would not correspond to realistic transformation which differentiate a patient from another. Instead, we propose to use a 3D plus time simulator recently developed for perfusion MRI [27]. In the use case of Giacalone et al. [27] the simulator served as a ground truth generator to evaluate the robustness of deconvolution algorithms [28]. We propose to extend here the use of the physical simulator of Giacalone et al. [27] to another problematic of more clinical importance in acute ischemic stroke management, with the prediction of the fate of the tissue from perfusion imaging.

As main novelty of our work, we demonstrate the possibility to boost the performance of final stroke prediction with help of synthetic perfusion MRI images produced by the physically and physiologically relevant simulator of Giacalone et al. [27]. Additionally, we further enrich this simulation by focusing on arterial input function which was stressed as the limiting factor of the simulator in Giacalone et al. [27]. In the other deep learning studies applied to stroke [14–18] the training dataset was based on a cohort of patients. By contrast here, we demonstrate the possibility, thanks to the use of simulation, to train on the perfusion MRI data of a single patient in acute stroke to predict the final infarct of this specific patient. Closest related work regarding synthetic data has been used to learn perfusion parameters from a relatively small number of training samples in CT perfusion [29] with classical data augmentation techniques. By contrast, we predict the final fate of the tissue from raw (*i.e.* non deconvolved) data without help of perfusion parameters, we work on MRI perfusion images and we develop synthetic data from an MRI simulator. Recent studies have shown the benefit of learning from raw perfusion data of training cohorts for ischemic lesion prediction [30–32]. Here, as additional novelty, we show how synthetic data simulated from raw MRI perfusion data can be used to directly learn the final infarct of a given specific patient, without the need of a training cohort.

2. Material and method

2.1. Clinical MRI data

We used clinical MRI data from the European I-Know multicenter database [33]. All patients from the study gave their informed consent and the imaging protocol was approved by the regional ethics committee. In total, we had a cohort of 76 patients with acute ischemic hemispheric stroke at our disposal, including 40 patients who received a thrombolytic treatment while the remaining 36 patients received no treatment. None of the patients reperfused after stroke.

All patients underwent the following MRI protocol on admission: diffusion-weighted-imaging (DWI; repetition time 6000 ms, field of view 24 cm, matrix 128 × 128, slice thickness 5 mm), fluid-attenuated-inversion-recovery (FLAIR; repetition time, 8690 ms; echo time, 109 ms; inversion time 2500 ms; flip angle, 150; field of view, 21 cm; matrix, 224 × 256; 24 sections; section thickness, 5 mm), T2-weighted gradient echo, MR-angiography and dynamic susceptibility-contrast perfusion imaging (DSC-PWI; echo time 40 ms, repetition time 1500 ms, field of view 24 cm, matrix 128 × 128, 18 slices, slice thickness 5 mm; gadolinium contrast at 0.1 mmol/kg injected with a power injector). From DSC-PWI, we extracted the commonly used hemodynamic maps such as: the 3D maps of the cerebral blood flow (CBF), the cerebral blood volume

(CBV), the mean transit time (MTT), the time to maximum (TMAX) and the time to peak (TTP). A follow-up FLAIR-MRI was performed at 1-month after admission time. Raw perfusion MRI were registered, for each slice, using the first time point as reference for all the other time points, with a maximum mutual information approach. Final lesion was segmented for each patient on the one-month follow-up FLAIR-MRI by 3 experts. The FLAIR-MRI were first co-registered to DSC-PWI by computing the average temporal signal before contrast-agent arrival. Raw perfusion MRI were registered, for each slice, using the first time point as reference for all the other time points, with a maximum mutual information approach. This was done by registering each temporal point (n+1) on its previous temporal point (n) and by then applying recursively the transformation matrices obtained until all time points were aligned with the first time point. All registrations were done using Elastix software [34]. The transformation matrix obtained was then used to register the final ischemic lesion mask. After registration, final lesion was rebinarized by applying a 50% threshold correction to avoid possible partial volume effects introduced by registration.

2.2. MRI simulator

Simulated data were generated with the DSC-MRI perfusion simulator of Giacalone et al. [27] that is able to simulate contrast-agent concentration images. We briefly recall the parameters of this simulator which includes realistic brain and lesion shapes, distinct classes of tissues (healthy, infarcted, gray and white matter) and their associated hemodynamic parameters as well as arterial input function (AIF). In Giacalone et al. [27]; the simulator was used to test the robustness of AIF deconvolution. The sensitivity to all parameters was systematically tested and the uncertainty of the AIF itself was demonstrated, as also shown in Calamante et al. [35]; to be the limiting factor. In this study, to extend the value of the simulator of Giacalone et al. [27] to machine learning, we decided to limit the investigation on the choice of the simulated AIF. Haemodynamic parameters were set to their default values presented in Giacalone et al. [27]; that we recall in Table 1. These haemodynamic parameters correspond to average values from the values reported in the literature. Acquisition parameters were set to 200 a. u. for the baseline value, 60 s for the acquisition time, 0.030s s for the time echo and 21 dB for the SNR.

The AIF is modeled as a gamma distribution that can be expressed using the simplified formulation proposed by Ref. [36]:

$$f(t) = \begin{cases} 0, & \text{if } t \leq d \\ y_{max} \cdot \left(\frac{t-d}{t_{max}}\right)^\alpha \cdot \exp\left(-\alpha\left(1 - \frac{t-d}{t_{max}}\right)\right), & \text{if } t \geq d \end{cases} \quad (1)$$

where y_{max} and t_{max} respectively correspond to the magnitude and the position of the maximum of the arterial input function, d is the arrival time of the contrast agent and α corresponds to the shape parameter of the gamma function.

AIF extraction. We characterized the AIF of each patients with a multiple AIFs selection method. To do so, AIFs were extracted for each patient from voxels located in the main cerebral arteries on their DSC-PWI [37]. The voxel selection was performed with a manual ROI

Table 1

Simulator default parameters for tissue variability of hemodynamic parameters Cerebral Blood Flow (CBF, in mL/g/s) and Mean Transit Time (MTT, in s). The distribution of each parameter is modeled by a Gaussian of average μ , standard deviation σ . We considered 3 tissue classes: Healthy Gray Tissues (HGT), Healthy White Tissues (HWT), and Lesional Tissues (LT). Background (BG) has a null distribution.

Hemodynamic Parameters	BG	HGT	HWT	LT
CBF ($\mu \pm \sigma$)	0 ± 0	60 ± 9	25 ± 2.1	10 ± 4.3
MTT ($\mu \pm \sigma$)	0 ± 0	4 ± 2.2	4.8 ± 3.2	10 ± 5

method by three different operators. Then raw perfusion signals for all selected voxels of each patient were averaged to produce mean contrast-agent concentration signal. The contrast agent curves were then fitted by a gamma function defined in Eq. (1).

AIF characterization. The estimation of AIF parameters are known to be critical for the prediction of the ischemic lesion fate [27,35]. There are different possible origins of AIF variability. First, in a multicentric study the duration of the perfusion protocol may differ. In our case the duration was standardized to 1 min for all centers. Also, the delay between beginning of acquisition and the injection may differ from one patient to another producing a temporal shift. However because CNN are translation invariant they are not sensitive to this possible time shift. Intra-patient variability accounts for the AIF variation depending on the location of the selected voxel. Inter-patient variability is attributed to the amount of blood coming to the brain that may of course vary from one patient to another. Both these biological sources of variability were present in our dataset. Intra-patient variability was reduced in our study by averaging AIF for each patient after arterial selection by three distinct experts. Concerning inter-patient variability, as observed in Meijs et al. [38] and in our dataset, most AIFs present a narrow and high distribution represented in blue in Fig. 1 (the transit time of the contrast agent is between 10 and 15 seconds). However, few patients have larger AIF (represented in red in Fig. 1), hence they are underrepresented in the cohort. These observations motivated the choice to investigate various approaches to simulate AIF in this study along the various datasets described in the following.

2.3. Training and testing datasets

Three training datasets were created to predict ischemic lesion fate of real patients of the cohort as described in this subsection. The first training dataset corresponds to real patients from the cohort, while the two remaining training datasets are pure simulated data.

2.3.1. Training datasets

Training dataset with a selection of real patients. First, a training dataset of 6 patients from the cohort of Section 2.1 was created: all presented narrow AIFs as shown in blue in Fig. 1. This approach enables to test the predictive value of a biased training dataset of real patients presenting a very low AIF variability regarding our tested patients. This dataset from real patients is obviously very small. It will serve as reference to compare with the prediction result of dataset generated from simulated patients.

Training dataset with simulation from theoretical AIF found in the literature (dataset A). A simulated dataset, noted A, was created with concentration images generated with the AIF default settings of the simulator, which correspond to average AIF parameters from the

literature [39]. In this configuration, AIF simulation parameters were set to a unique value, that is to say: $y_{max} = 0.61$, $t_{max} = 4.5$, $d = 3$, $\alpha = 3$. This approach enables to test the predictive value of synthetic perfusion images simulated from a theoretical AIF not tuned to the values of our tested patients.

Training dataset with simulation from patient-specific AIF (dataset B). A second simulated dataset, noted B, was created by setting the AIF input parameters to the clinical AIF fitting parameters extracted from each tested patient (see chosen values in Table 2). This approach enables to test the predictive added value of synthetic perfusion images simulated from an AIF tuned to the values of our tested patients.

2.3.2. Testing dataset

We chose 8 patients from the cohort of Section 2.1 to build the testing dataset. These were selected to cover the diversity of AIF shape observed in the cohort. Only 2 of these 8 patients received a thrombolitic treatment but none of them reperused on their own. These 8 chosen patients, in addition to their representative AIF shapes, have been selected with sufficiently large final lesions. Indeed when final lesions are very small (typically smaller than some mLs), DSC are calculated on very few voxels, and each poorly predicted voxel very quickly penalizes the patient DSC. Also, lesions exploded in multiple non connected sub-lesions are also more difficult to predict as demonstrated in Frindel et al. [40]. This choice is justified to guarantee a controlled evolution of the lesion (stable or enlarged lesions) in order to focus on the AIF and the variability resulting from this parameter. As shown in Fig. 2, patients 1, 2, 3 and 6 present large AIF, whereas the other patients present relatively narrow AIF shapes. AIF of each individual patient to be tested were extracted. The fitting parameters of each of the tested patients are presented in Table 2 and depicted in Fig. 2. In dataset B variants of the AIF of Table 2 were simulated.

Table 2

AIF parameters (y_{max} , t_{max} , d , α) values of the tested patients obtained from their mean contrast agent gamma curve.

Patient	y_{max}	t_{max}	d	α
1	0.90	6.32	13.14	4.23
2	0.86	10.53	11.99	8.54
3	0.60	3.41	18.67	0.99
4	0.92	5.26	19.64	0.94
5	0.74	2.35	11.62	1.01
6	0.60	4.58	14.01	2.30
7	0.77	3.45	9.40	2.48
8	0.85	7.08	8.79	8.03

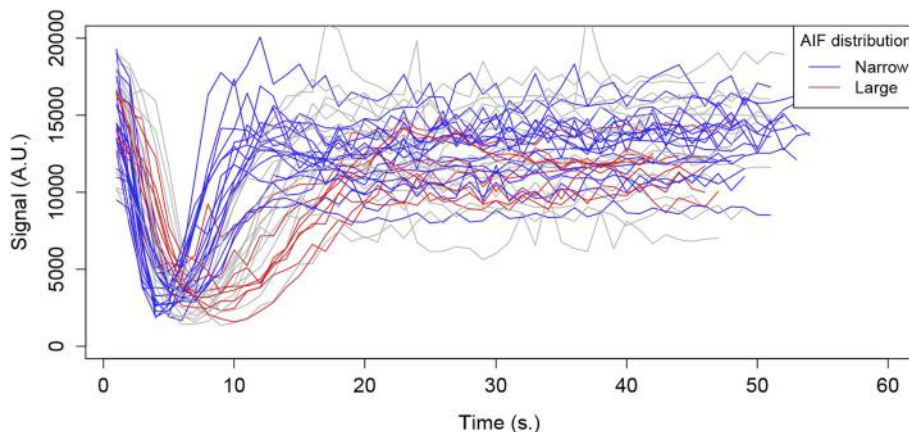


Fig. 1. Mean AIF curves extracted on each of the 76 patients in the cohort. The blue and red lines correspond to AIFs with respectively narrow and large distributions. The gray lines correspond to the remaining AIFs. It shows the wide variability and inhomogeneity of AIF shapes.

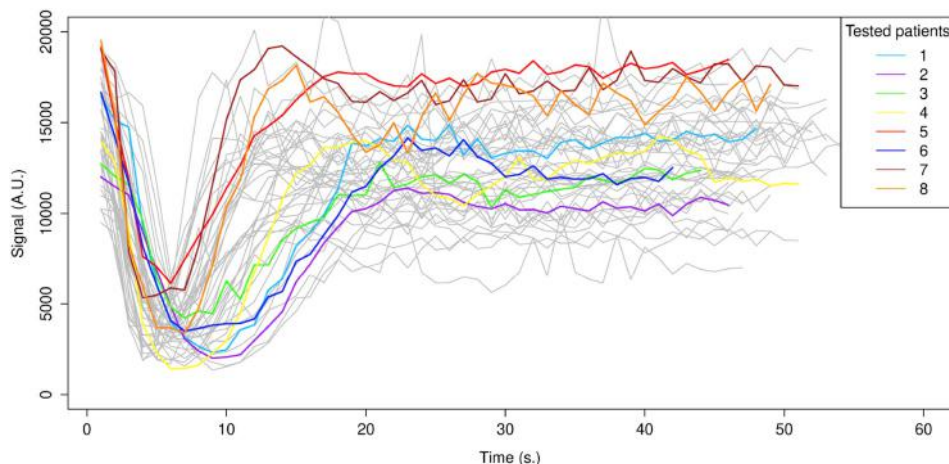


Fig. 2. AIF curves from the 8 testing patients, shown in color, among all the AIF of the cohort. In order to better visualize them, we did not show the delay d . The chosen patients present wide AIF shape differences.

2.4. Voxel fate prediction models

The prediction of cerebral tissue fate from perfusion images was

done after a spatio temporal encoding of the voxel environment fed to convolutional neural networks, as described in this subsection. Fig. 3 shows the proposed pipeline when learning from simulated data.

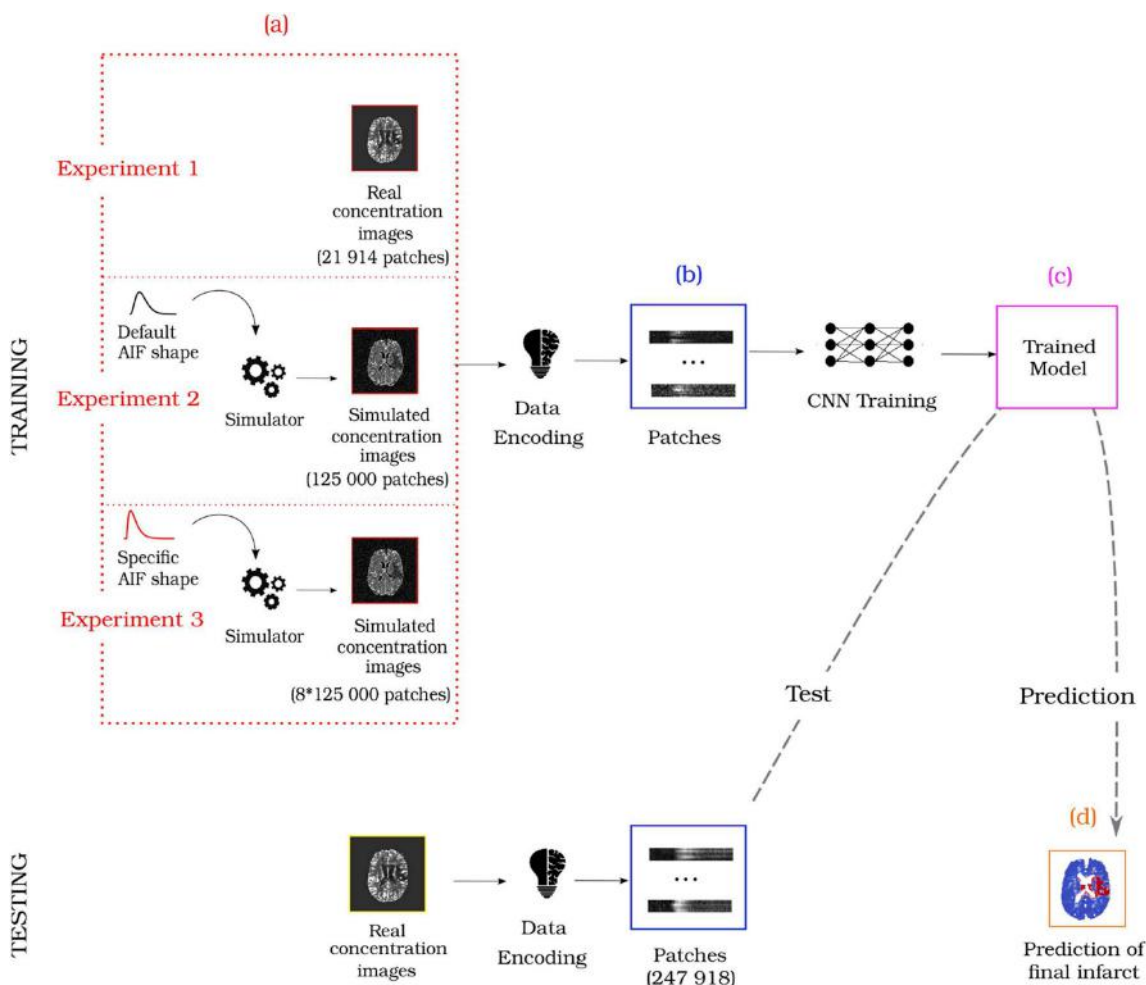


Fig. 3. Overview of the proposed prediction pipeline. (a) The initial images are contrast-agent concentration images. In experiment 1, the training dataset consists in patches from real concentration images, whereas in experiment 2 and 3, the training dataset consists in synthetic patches obtained from the simulator. In experiment 2 AIF input parameter is set to a default value, and in experiment 3, AIF input parameter are the ones of the tested patient. (b) Concentration images are encoded into spatio-temporal patches. (c) A Convolutional Neural Network (CNN) model is trained from patches of the concentration images. (d) Each voxel from the tested concentration images is classified as healthy or infarcted.

2.4.1. Encoding of perfusion images

Perfusion images were converted into contrast-agent concentration images after a logarithmic transformation, under the assumption of a linear relationship between the contrast agent concentration and the change in transverse relaxation rate [41,42]. This transformation makes it possible to standardize the images between patients since the baseline is dispensed with. Then, concentration images were encoded in local spatio-temporal patches as recently described in Giacalone et al. [30]. Shortly, the time signal of a voxel of interest is deployed along a spatial direction. Its 8 voxels in the Moore neighborhood of order 1 are also deployed in the same direction, stacking the time signature of each neighboring voxel below each other. Thereby a patch of size 9 by Nt is created for each voxel, where Nt is the number of temporal acquisition points in the perfusion imaging sequence. In order to obtain patches independent of each other, we did not consider a Moore neighborhood of higher order.

It has been shown that patches for injured voxels present different patterns than patches for healthy voxels which can be discriminated in terms of texture [30]. To go further than Giacalone et al. [30]; we can notice from Fig. 4 that this discriminability strongly depends on AIF. For instance in Fig. 4, patient 8, who had a narrow AIF, has a contrast-agent transit time much shorter than patient 2, who had a larger AIF. This supports the need for specific patient learning, taking into account the AIF of each patient to better predict their pathological voxels.

2.4.2. CNN classifier

A CNN was designed to directly take as input the spatio-temporal patches of dimension (9,60) to make a voxel-based prediction, as one patch represent the spatio-local environment of one voxel. The output for each patch was the predicted probability to belong to two classes (healthy tissue or infarcted tissue). We chose CNN as they are known to be translation invariant [43,44]. This property was a key point in learning, as we only wish to learn about the transit of the contrast agent in tissues and not about its arrival time. Thus, the network would not be sensitive to the delay d , but only to the white pattern of the spatio-temporal signature. A unique architecture described in Table 3 was designed. This architecture present a limited number of convolutional layers in order to avoid the patch size reduction and overfitting. Network weights were randomly initialized at the beginning of the training. Rectifier linear units (ReLU) function was used as activation function, known to perform better and faster than the sigmoid or hyperbolic tangent functions [45,46]. In the last fully connected layer, we used softmax, with 2 output units as our task approach a binary segmentation problem. As long as the patches input have small dimensions (9,60) and that convolution tends to reduce the output image dimension,

Table 3

Convolutional neural network architecture proposed. The type of the two first layers are 2D Convolutional layers (Conv2D) and the two last layers are fully connected layers (FC).

Layer	Type	Filter Size	Stride	# filters	FC units	Output Shape
1	Conv2D	2*2	1*1	16	-	(8, 59, 16)
2	Conv2D	2*2	1*1	32	-	(7, 58, 32)
3	FC	-	-	-	15	(15)
4	FC	-	-	-	2	(2)

we decided not to use any max-pooling to avoid further size reduction. We used dropout [47,48] in the fully connected layers in order to avoid overfitting. We used the categorical cross-entropy function as a loss function and a stochastic gradient descent to optimize the model. For all experiments, the total number of weights to train was 197 087, the dropout was set to 0.5, the number of epochs was set to 30, and the batch size to 32.

Each model was trained 10 times in order to have an overview of their global performance and not only the best metric shot. To get a balanced training dataset we ensured that half of the patches contain voxels classified as lesion on the follow-up FLAIR-MRI. All CNNs were trained using Keras 2.1.3 with Python 3.6.3 interface. The training of the networks took globally less than 15 min on a standard work station with an NVIDIA GeForce GTX 1080 GPU with 8 GB memory.

2.4.3. Evaluation of the classification

We assessed our results using the Dice Similarity Coefficient (DSC) [49] and the Hausdorff Distance (HD) [50], as both were used for the international Ischemic Stroke Lesion Segmentation challenge of MICCAI [51]. These metrics were computed between the predicted infarcted voxels and the mask of the final lesion provided by FLAIR-MRI. For comparison the prediction of the perfusion lesion was also computed from a TMAX perfusion map thresholded at 6 s with approach proposed by Frindel et al. [28]. This procedure is a standard approach in clinical research [52,53]. We computed DSC and HD between the voxels above this threshold and the mask of the final lesion as a clinical reference. Using synthetic data for training enables to produce data on demand. We evaluated the minimum number of simulated patches required to obtain stable learning for the described CNN architecture. For each patient, we had 125 000 initial patches that we divided into 8 subsets by simple random sampling with replacement of different size: respectively 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of the initial number of patches. For each patient, each set of patches was trained 10 times.

2.5. Experiment details

We classified the patches of the testing dataset through three experiments, each using a different training dataset. We give experimental details in the following subsection.

Experiment 1: training from a selection of real patients. In this experiment, the training dataset consisted of patches from 6 patients of the cohort presenting narrow AIF (see Section 2.3.1). From these 6 patients, we were able to get 21 914 patches with half of it healthy, and the other half infarcted. The validation dataset consisted of 17 766 patches from 3 independent patients presenting a different AIF shape. In this experiment, learning rate was set to 0.01.

Experiment 2: training with simulation from theoretical AIF. In this experiment, the training dataset consisted of 125 000 patches from the simulated images of dataset A with half of it healthy, and the other half infarcted. The validation dataset consisted of a set of 125 000 patches from other images obtained with the same simulation parameters. In this experiment, learning rate was set to 0.0001.

Experiment 3: training with simulation from patient-specific AIF. In this experiment, we learned from a synthetic specific patient. Dataset B was separated into several subsets, each set representing simulated images of

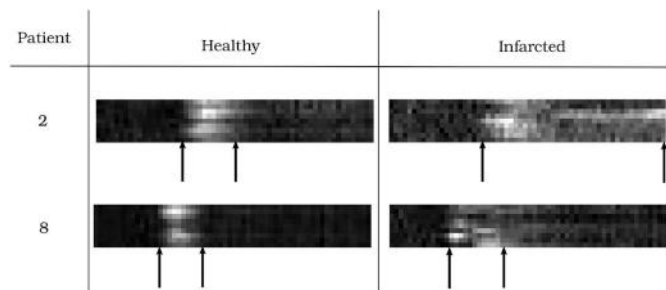


Fig. 4. Illustration of a healthy patch (left) and a pathological patch (right) from concentration image for the tested patients 1 and 8. Healthy voxels exhibits a narrow hyperintensity segment resulting from the quick contrast-agent bolus passage. Pathological voxels exhibits a spread out hyperintensity segment, noisy and low contrasted, resulting from the difficult passage of the contrast-agent bolus. The two patients have different hemodynamic characteristics: the transit time of the contrast-agent, represented by the space between the two arrows, is faster for patient 8 than for patient 2. Patch intensities were converted into grayscale image for the representation.

a specific tested patient. Therefore, 8 different trainings were done. For each training, the corresponding training set consisted of 125 000 patches with half of it healthy, and the other half infarcted, and the corresponding validation dataset consisted of a set of 125 000 patches from other images obtained with the same patient specific simulation parameters. In this experiment, learning rate was set to 0.0001.

3. Results

Table 4 reports the results obtained by the three conducted experiments in terms of mean DSC and HD values and their standard deviation for each tested patient. The DSC and HD values between the thresholded TMAX and the final lesion is also shown, as a clinical reference.

The results from experiment 1 show the impact of a mismatch between training and testing in terms of AIF. Discrimination between healthy and pathological voxels was only trained from narrow AIF patients: it turns out to be impossible to correctly predict the voxels of the tested patients with different AIF shapes such as patient 1, 2, 3 and 6 who have mean DSC below 0.13. In contrast, tested patients with AIFs close to those of the learned patients, such as patients 4, 5, 7 and 8 have mean DSC greater than 0.49.

In experiment 2, training CNN with synthetic data obtained from a theoretical AIF, corresponding to average AIF parameters from the literature without any patient specific tuning, provides very poor results. Almost all voxels are predicted infarcted, so that DSC is very low and the Hausdorff distance high. This illustrates that a theoretical AIF is not able to capture the variability that exists between patients. AIF differs from subject to subject and since we work on raw perfusion imaging data, it is important to incorporate the specific AIF of each considered patient.

The variability brought by the AIF from one patient to another is well-known [38]. This is the reason why most classical approaches use deconvolution in order to compensate for this variability. In experiment 3, we did not solve this ill-positioned inverse problem and rather incorporate the specific arterial input function in the direct problem through simulation of the perfusion signals with a set of realistic arterial input functions. In experiment 3, the training dataset contains systematically simulated images where the AIF is tuned to those of the patient to be predicted (dataset B). It clearly appears that the adjustment of the AIF-related parameters in the simulator has a considerable impact on the learning performance. The average DSC in experiment 3 is 0.40 (± 0.19), compared with 0.14 (± 0.074) in experiment 2 and 0.30 (± 0.22) in experiment 1. Learning from specific AIF in experiment 3 clearly improved the median DSC, which is 0.48 in experiment 3, compared with 0.27 in experiment 1 and 0.15 in experiment 2. It appears also that learning from raw data in experiment 3 gives better results than thresholding the deconvolved TMAX which present an average DSC of 0.32 (± 0.14). As another reference, it is interesting to note that the best

models so far in the stroke prediction ISLES 2017 challenge had an average DSC of 0.38 (± 0.22) and an average Hausdorff distance of 29.21 (± 15.04). The mean performance scores in experiment 3 is in the same order of magnitude. However an absolute comparison is not strictly possible because the two datasets are different. To test the transferability of our proposed method, we have run 10 times experiment 3 for patients with large lesions, patients number 7 and 20, from the ISLES dataset. The average DSC was 0.39 (± 0.30) and the average Hausdorff distance was 46.9 (± 1.80).

Regarding our described CNN architecture in experiment 3, we also investigated the minimum patches required for stable training. Results are shown in Fig. 5. It appears that between 25 000 and 100 500 training patches, DSC increases almost linearly depending on the number of training patches. Beyond 100 000 patches, the curve seems to reach a plateau: all the diversity of information provided by the simulated images has been learned, and the supply of new images is redundant. Also the standard deviation of the DSC values is lower after training more than 100 000 patches. These observations indicate that given our CNN architecture and our dataset, the minimum number of training patches to obtain stable and optimal results is around 100 000.

As additional experiments, we compared experiment 1, 2 and 3 with other neural network architectures and metrics of merit. The results of this comparison are provided in the additional material section. The added value of the simulation used in experiment 3 over experiments 2 and 1 is robustly obtained.

4. Discussion

The previous results have demonstrated the value of perfusion imaging simulation based on AIF for the prediction of lesion fate in stroke. In this section, we go beyond the sole observation of average performance and now discuss the limit of our experiments.

Faced with the problem of AIF representativeness in the training set underlined in experiments 1 and 2, we proposed through experiment 3 to learn directly from the AIF of the patient to be tested. Thanks to this type of learning, each patient was correctly predicted individually, even for patients whose AIF is under-represented in the overall cohort. It should be noted that for experiment 3, we simulated images from a single averaged AIF of the tested patient. However, some studies show that it may be beneficial to take into account the intra-patient AIF variability [54]. Indeed, in experiment 3, the extracted AIF seems relevant for patients 2, 3, 4, 5, 6 and 7, but not enough for patients 1 and 8 as they show better performances in the experiment 1. Probably, these two patients cannot be summarized in one single AIF as they might present a large AIF intra-variability. This encourages us for further work to simulate images with several AIFs according to the intra-AIF variability of the tested patient, and therefore potentially better represent their

Table 4

Hausdorff distance (HD) and similarity metrics (DSC) after performing 10 times experiment 1 (1st column), 2 (2nd column), and 3 (3rd column). All metrics are averaged over the 10 times, and shown for each tested patient (average \pm standard deviation). DSC and HD between TMAX $\geq 6s$ and final lesion is shown (4 rth column). We showed in bold when experiment 3 gave the best performance at the patient scale. For the three experiments, HD standard deviation is low or even zero because some outlier voxels were systematically mis-predicted. The metrics are also averaged over the test dataset (last row).

Patient	Experiment 1		Experiment 2		Experiment 3		TMAX thr	
	HD	DSC	HD	DSC	HD	DSC	HD	DSC
1	50.4 \pm 0.0	0.17 \pm 0.031	50.2 \pm 0.0	0.12 \pm 0.00	50.2 \pm 0.0	0.086 \pm 0.003	49.7	0.11
2	51.1 \pm 0.0	0.18 \pm 0.002	51.1 \pm 0.0	0.18 \pm 0.00	48.8 \pm 0.829	0.62 \pm 0.008	43.2	0.51
3	44.1 \pm 0.0	0.054 \pm 0.001	43.8 \pm 0.0	0.053 \pm 0.00	43.8 \pm 0.0	0.23 \pm 0.010	43.4	0.14
4	44.4 \pm 0.0	0.47 \pm 0.041	44.9 \pm 0.0	0.17 \pm 0.00	44.8 \pm 0.085	0.48 \pm 0.008	42.2	0.47
5	43.8 \pm 1.08	0.36 \pm 0.069	45.4 \pm 0.0	0.11 \pm 0.00	44.4 \pm 0.0	0.47 \pm 0.005	46.1	0.31
6	43.0 \pm 0.0	0.11 $\pm 10^{-5}$	43.0 \pm 0.0	0.11 \pm 0.00	42.5 \pm 0.0	0.20 \pm 0.006	44.7	0.28
7	44.0 \pm 0.36	0.47 \pm 0.031	45.0 \pm 0.0	0.26 \pm 0.00	45.0 \pm 0.0	0.53 \pm 0.004	46.3	0.36
8	40.1 \pm 3.07	0.64 \pm 0.026	47.1 \pm 0.0	0.17 \pm 0.00	44.0 $\pm 10^{-6}$	0.48 \pm 0.017	45.8	0.40
Total mean	45.1 \pm 3.73	0.30 \pm 0.22	46.3 \pm 2.94	0.14 \pm 0.074	45.4 \pm 2.65	0.40 \pm 0.19	45.18 \pm 2.36	0.32 \pm 0.14

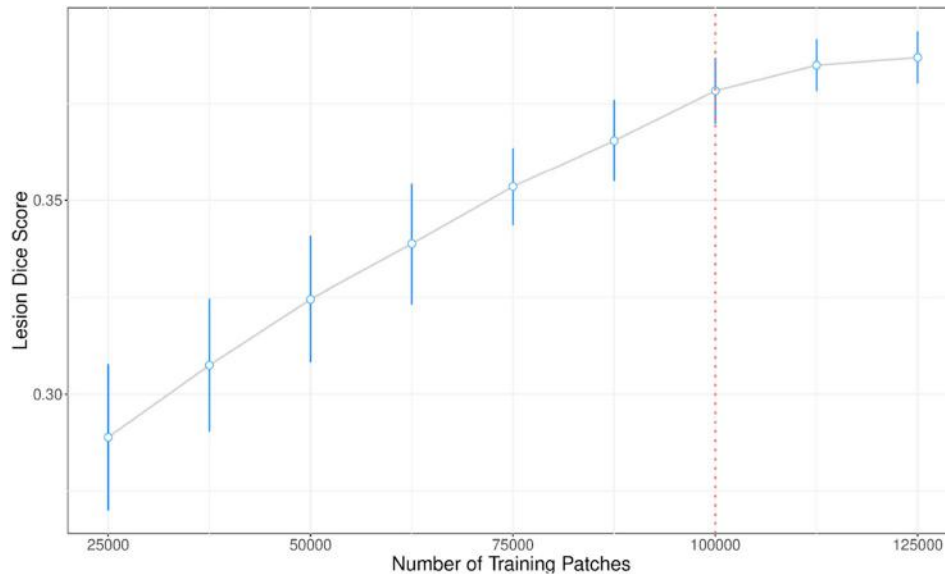


Fig. 5. Resulting DSC according to the number of training patches in experiment 3. Each point corresponds to the mean DSC of 10 repeated experiments on each tested patient and its standard deviation. The red dotted line indicates the optimal number of training patches.

hemodynamic characteristics.

At the voxel level, Fig. 6 shows the position of the badly classified voxels resulting from experiments 1, 2 and 3. We can see that many of our errors are in the ventricles. As contrast-agent does not pass into these areas, the voxels have a particularly noisy signal. Thus the model detects cerebral blood flow disturbance and directly links it with the final ischemic stroke. These errors could be easily filtered by limiting mask segmentation errors.

The current state-of-the-art in the prediction of lesion fate in stroke classically works on the deconvolved parametric maps (CBV, CBF, Tmax, TTP and MTT) learned on cohorts of patient. With our approach it is actually not possible to learn directly from these maps since our encoding is based on the spatio-temporal signature of perfusion data. A possible comparison would be to compare learning from raw perfusion signals, encoded with our spatio-temporal approach, and learning from deconvolved signals (the one used to produce standard parametric maps CBV, CBF, Tmax, TTP, and MTT). Such a comparison was shown in Giacalone et al. [30] and demonstrated the interest of the proposed spatio-temporal encoding of raw perfusion signal.

In this work, the CNN architecture exploited consists of a low number of layers. The minimum number of patches required for a specific patient learning was therefore limited to 100 000 (which is approximately 25 simulations needed to represent a patient with 2000 infarcted voxels and corresponds to a computation time of 3 min on the work station of Section 2.4.2). However, with other encoding methods and more complex models, the number of data required for learning could be higher. The simulator we used in this paper [27] would be able to overcome this problem, by allowing to produce a theoretically unlimited amount of labelled data.

Finally, it should be noted that this approach, although no instantaneous, seems fully compatible with the real-time management of stroke patient in clinical routine although including learning and simulation. Indeed, for performances compared with the state-of-the-art, the overall computation time (simulation and learning) for a patient-specific approach as developed in experiment 3 is about 20 min. This remains of the same order of magnitude as the time announced for a deconvolution approach as in Frindel et al. [28].

5. Conclusion

In this article, we demonstrated how the simulation of

hemodynamical signals can be used to increase the amount of data and boost the performance of convolutional neural networks for the prediction of the lesion evolution in acute ischemic stroke from DSC-PWI. This new demonstration of the value of simulation to train machine learning techniques in medical imaging enabled to obtain performances close to the ones of the literature for this important stroke problem.

Several simulation approaches have been tested including the simulation from average AIF values found in the literature or simulation from the AIF of the specific tested patient. In this patient specific scenario, we have shown that the performance of prediction was higher to results from the state of the art methods applied on cohorts while learning here on a specific patient instead of a cohort. A limitation in machine learning based prediction for biomedical imaging is the limited size of cohorts. This is a priori especially true with highly expressive models such as the one based on convolutional neural networks. We have demonstrated on the specific disease of stroke that in fact it is possible to predict on an extremely limited cohort of a single patient from convolutional neural networks with help of simulation.

This work could be extended in several directions. For stroke, several works propose to predict the final infarction simply through the perfusion modality, in the framework of the challenge ISLES for example [8] or outside this challenge [17,30,31]. However, acute DWI is known to be a highly predictive image for the final stroke lesion [55,56]. Incorporating diffusion in the model could highly improve performance classification. This would require to integrate a diffusion simulator [57,58] and redesign the architecture of the convolutional neural network used in this study. Also the simulation was here limited to a pixel-wise simulator. More spatial context could be integrated in the realism of the simulator by adding a physical layer at a spatial level of the DSC-MRI simulator. As another extension of this work for stroke, one can notice that the simulation of brain tissue was here based on global stationary binary class model (healthy versus infarcted) while from a clinical perspective, stroke appears as a more complex problem with importance on the spatial localisation of the tissue. It is however obvious that whatever the level of simulation complexity, some bias between real and simulated images will remain. A way to learn this bias and thus improve the performance of simple simulation-based training approaches, such as the one introduced here, could be to add some domain adaptation techniques just before the last decision layers of the convolutional neural network used here [59]. Finally, the use of simulation to train machine learning based model is of generic value and may be extended

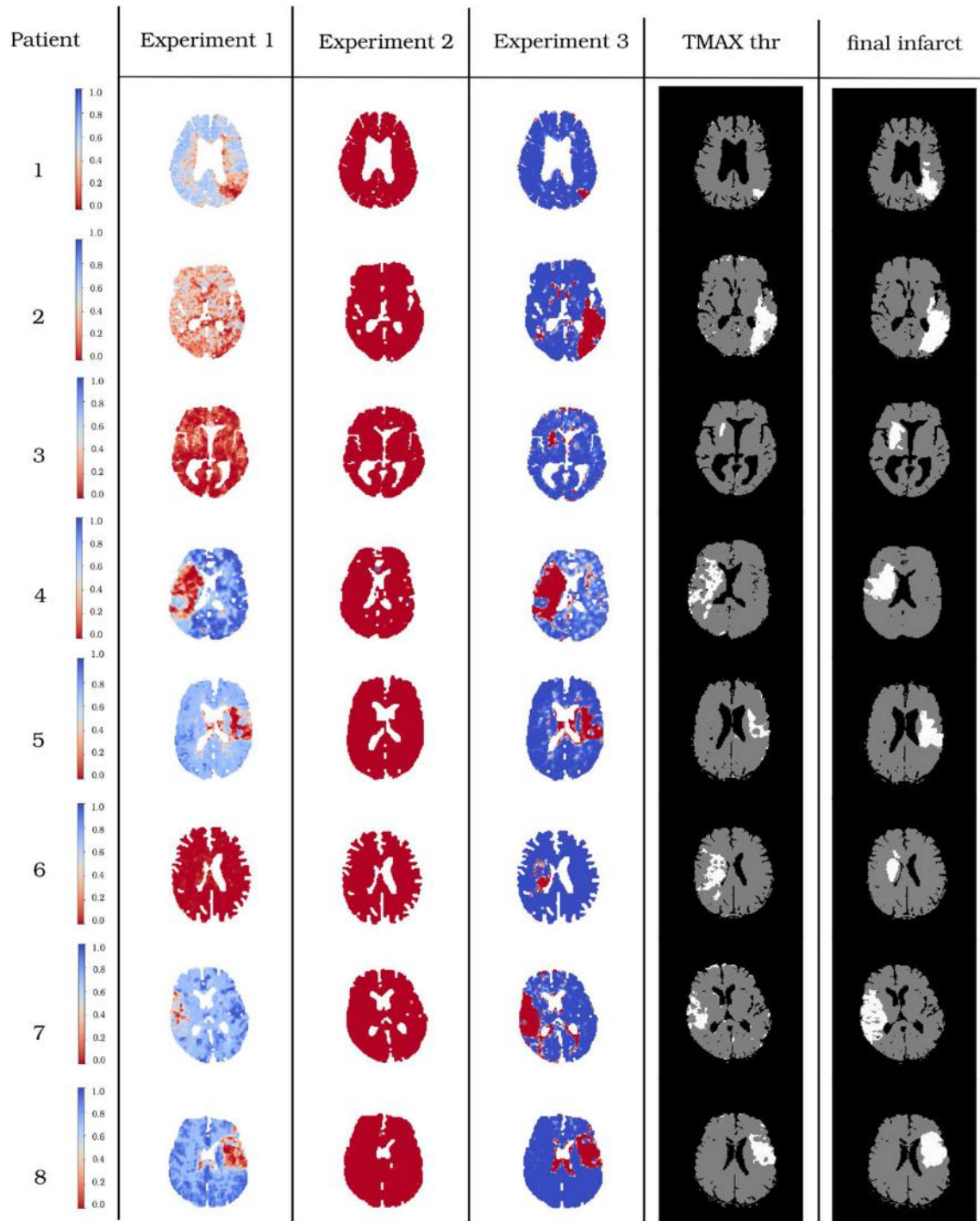


Fig. 6. Output predictions of experiment 1 (1st column), experiment 2 (2nd column) and experiment 3 (3rd column) for the testing patients. The colorbar presents the probability for each voxel to be healthy. Voxels in blue shades were predicted healthy and voxels in red shades were predicted infarcted. The classical biomarker TMAX thresholded at 6 s is shown (4th column). Columns 1 to 4 should be compared to the final flair (5th column).

to any other disease and imaging modality for which some simulator are already available [60].

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.compbimed.2019.103579>.

References

- [1] C.J. Murray, R.M. Barber, K.J. Foreman, A.A. Ozgoren, F. Abd-Allah, S.F. Abera, V. Aboyans, J.P. Abraham, I. Abubakar, L.J. Abu-Raddad, et al., Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990-2013: quantifying the epidemiological transition, *The Lancet* 386 (2015) 2145–2191.
- [2] L.R. Caplan, *Caplan's Stroke*, Cambridge University Press, 2016.
- [3] J. Park, *Acute Ischemic Stroke*, Springer, 2017.
- [4] K. Sartor, *Magnetic Resonance Imaging in Ischemic Stroke*, Springer Science & Business Media, 2006.

- [5] M. Goyal, A.M. Demchuk, B.K. Menon, M. Eesa, J.L. Rempel, J. Thornton, D. Roy, T.G. Jovin, R.A. Willinsky, B.L. Sapkota, et al., Randomized assessment of rapid endovascular treatment of ischemic stroke, *N. Engl. J. Med.* 372 (2015) 1019–1030.
- [6] G.W. Albers, M.P. Marks, S. Kemp, S. Christensen, J.P. Tsai, S. Ortega-Gutierrez, R. A. McTaggart, M.T. Torbey, M. Kim-Tenser, T. Leslie-Mazwi, et al., Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging, *N. Engl. J. Med.* 378 (2018) 708–718.
- [7] C. Zerna, J. Hegehdus, M.D. Hill, Evolving treatments for acute ischemic stroke, *Circ. Res.* 118 (2016) 1425–1442.
- [8] O. Maier, B.H. Menze, J. von der Gabelntz, L. Häni, M.P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al., ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI, *Med. Image Anal.* 35 (2017) 250–269.
- [9] S. Christensen, M.G. Lansberg, CT perfusion in acute stroke: practical guidance for implementation in clinical practice, *J. Cereb. Blood Flow Metab.* 39 (9) (2018).
- [10] G.W. Albers, Use of imaging to select patients for late window endovascular therapy, *Stroke* 49 (2018) 2256–2260.
- [11] M. Najm, F.S. Al-Ajlan, M.E. Boesen, L. Hur, C.K. Kim, E. Fainardi, M.D. Hill, A. M. Demchuk, M. Goyal, T.Y. Lee, et al., Defining CT perfusion thresholds for infarction in the golden hour and with ultra-early reperfusion, *Can. J. Neurol. Sci.* 45 (2018) 339–342.
- [12] Z. Zhang, E. Sejdic, Radiological images and machine learning: trends, perspectives, and prospects, *Comput. Biol. Med.* 108 (2019) 354–370.
- [13] O. Maier, C. Schröder, N.D. Forkert, T. Martinetz, H. Handels, Classifiers for ischemic stroke lesion segmentation: a comparison study, *PLoS One* 10 (2015) 1–16.
- [14] N. Stier, N. Vincent, D. Liebeskind, F. Scalzo, Deep learning of tissue fate features in acute ischemic stroke, in: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2015, pp. 1316–1321.
- [15] A. Nielsen, M.B. Hansen, A. Tietze, K. Mouridsen, Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning, *Stroke* , STROKEAHA– 117 (2018).
- [16] R. Zhang, L. Zhao, W. Lou, J.M. Abrigo, V.C. Mok, W.C. Chu, D. Wang, L. Shi, Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional densenets, *IEEE Trans. Med. Imaging* 37 (2018) 2149–2160.
- [17] C. Lucas, A. Kemmling, A.M. Mamlouk, M.P. Heinrich, Multi-scale neural network for automatic segmentation of ischemic strokes on acute perfusion images, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1118–1121.
- [18] S. Pedemonte, B. Bizzo, S. Pomerantz, N. Tenenholtz, B. Wright, M. Walters, S. Doyle, A. McCarthy, R.R. De Almeida, K. Andriole, et al., Detection and delineation of acute cerebral infarct on DWI using weakly supervised machine learning, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Springer, 2018, pp. 81–88.
- [19] M. Livne, J.K. Boldsen, I.K. Mikkelsen, J.B. Fiebach, J. Sobesky, K. Mouridsen, Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke, *Stroke* 49 (2018) 912–918.
- [20] A. Subudhi, U.R. Acharya, M. Dash, S. Jena, S. Sabut, Automated approach for detection of ischemic stroke using delaunay triangulation in brain mri images, *Comput. Biol. Med.* 103 (2018) 116–129.
- [21] G. Praveen, A. Agrawal, P. Sundaram, S. Sardesai, Ischemic stroke lesion segmentation using stacked sparse autoencoder, *Comput. Biol. Med.* 99 (2018) 38–52.
- [22] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017*, pp. 2107–2116.
- [23] F. Mahmood, R. Chen, N.J. Durr, Unsupervised reverse domain adaptation for synthetic medical images via adversarial training, *IEEE Trans. Med. Imaging* 37 (2018) 2572–2581.
- [24] H.C. Shin, N.A. Tenenholtz, J.K. Rogers, C.G. Schwarz, M.L. Senjem, J.L. Gunter, K. P. Andriole, M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in: *Simulation and Synthesis in Medical Imaging*, Springer, 2018, pp. 1–11.
- [25] P.Y. Simard, D. Steinkraus, J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: *Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE, 2003*, pp. 958–962.
- [26] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, *Journal of Big Data* 6 (2019) 60.
- [27] M. Giacalone, C. Frindel, M. Robini, F. Cervenansky, E. Grenier, D. Rousseau, Robustness of spatio-temporal regularization in perfusion MRI deconvolution: an application to acute ischemic stroke, *Magn. Reson. Med.* 78 (2017) 1981–1990.
- [28] C. Frindel, M.C. Robini, D. Rousseau, A 3-D spatio-temporal deconvolution approach for MR perfusion in the brain, *Med. Image Anal.* 18 (2014) 144–160.
- [29] D. Robben, P. Suetens, Perfusion parameter estimation using neural networks and data augmentation, in: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer, 2018, pp. 439–446.
- [30] M. Giacalone, P. Rasti, N. Debs, C. Frindel, T.H. Cho, E. Grenier, D. Rousseau, Local spatio-temporal encoding of raw perfusion MRI for the prediction of final lesion in stroke, *Med. Image Anal.* 50 (2018) 117–126.
- [31] A. Pinto, S. Pereira, R. Meier, V. Alves, R. Wiest, C.A. Silva, M. Reyes, Enhancing clinical MRI perfusion maps with data-driven maps of complementary nature for lesion outcome prediction, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Springer, 2018, pp. 107–115.
- [32] D. Robben, A.M. Boers, H.A. Marquering, L.L. Langezaal, Y.B. Roos, R.J. van Oostenbrugge, W.H. van Zwam, D.W. Dippel, C.B. Majorie, A. van der Lugt, et al., Prediction of Final Infarct Volume from Native CT Perfusion and Treatment Parameters Using Deep Learning, 2018 arXiv preprint arXiv:1812.02496.
- [33] L. Hermitte, T.H. Cho, B. Ozenne, N. Nighoghossian, I.K. Mikkelsen, L. Ribe, J. C. Baron, L. Østergaard, L. Derex, N. Hjort, et al., Very low cerebral blood volume predicts parenchymal hematoma in acute ischemic stroke, *Stroke*, STROKEAHA 113 (2013).
- [34] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P. Pluim, Elastix: a toolbox for intensity-based medical image registration, *IEEE Trans. Med. Imaging* 29 (2009) 196–205.
- [35] F. Calamante, D.G. Gadian, A. Connelly, Delay and dispersion effects in dynamic susceptibility contrast MRI: simulations using singular value decomposition, *Magn. Reson. Med.* 44 (2000) 466–473.
- [36] M.T. Madsen, A simplified formulation of the gamma variate function, *Phys. Med. Biol.* 37 (1992) 1597.
- [37] A. Waaijjer, I. Van der Schaaf, B. Velthuis, M. Quist, M. Van Osch, E. Vonken, M. Van Leeuwen, M. Prokop, Reproducibility of quantitative CT brain perfusion measurements in patients with symptomatic unilateral carotid artery stenosis, *Am. J. Neuroradiol.* 28 (2007) 927–932.
- [38] M. Meijs, S. Christensen, M.G. Lansberg, G.W. Albers, F. Calamante, Analysis of perfusion MRI in stroke: to deconvolve, or not to deconvolve, *Magn. Reson. Med.* 76 (2016) 1282–1290.
- [39] E. Kellner, I. Mader, M. Mix, D.N. Splitthoff, M. Reisert, K. Foerster, T. Nguyen-Thanh, P. Gall, V.G. Kiselev, Arterial input function measurements for bolus tracking perfusion imaging in the brain, *Magn. Reson. Med.* 69 (2013) 771–780.
- [40] C. Frindel, A. Rouanet, M. Giacalone, T.H. Cho, L. Østergaard, J. Fiehler, S. Pedraza, J.C. Baron, M. Wiart, Y. Berthezène, et al., Validity of shape as a predictive biomarker of final infarct volume in acute ischemic stroke, *Stroke*, STROKEAHA 114 (2015).
- [41] L. Østergaard, R.M. Weisskoff, D.A. Chesler, C. Gyldensted, B.R. Rosen, High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. Part I: mathematical approach and statistical analysis, *Magn. Reson. Med.* 36 (1996) 715–725.
- [42] A. Villringer, B.R. Rosen, J.W. Belliveau, J.L. Ackerman, R.B. Lauffer, R.B. Buxton, Y.S. Chao, V.J. Wedeenand, T.J. Brady, Dynamic imaging with lanthanide chelates in normal brain: contrast due to magnetic susceptibility effects, *Magn. Reson. Med.* 6 (1988) 164–174.
- [43] M.D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks, in: *Computer Vision – ECCV 2014*, Springer, 2014, pp. 818–833.
- [44] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biol. Cybern.* 36 (1980) 193–202.
- [45] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [46] K. Jarrett, K. Kavukcuoglu, Y. LeCun, et al., What is the best multi-stage architecture for object recognition?, in: *2009 IEEE 12th International Conference on Computer Vision IEEE, 2009*, pp. 2146–2153.
- [47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout : a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* 15 (2014) 1929–1958.
- [48] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, 2012. 1207.0580.
- [49] L.R. Dice, Measures of the amount of ecologic association between species, *Ecology* 26 (1945) 297–302.
- [50] J. Henrikson, Completeness and total boundedness of the Hausdorff metric, *MIT Undergraduate Journal of Mathematics* 1 (1999) 69–80.
- [51] S. Winzeck, A. Hakim, R. McKinnis, J.A. Pinto, V. Alves, C. Silva, M. Pisov, E. Krivov, M. Belyaev, M. Monteiro, et al., ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI, *Front. Neurol.* 9 (2018) 679.
- [52] J.M. Olivot, M. Mlynash, V.N. Thijs, S. Kemp, M.G. Lansberg, L. Wechsler, R. Bammer, M.P. Marks, G.W. Albers, Optimal tmax threshold for predicting penumbral tissue in acute stroke, *Stroke* 40 (2009) 469–475.
- [53] T.H. Cho, N. Nighoghossian, I.K. Mikkelsen, L. Derex, M. Hermier, S. Pedraza, J. Fiehler, L. Østergaard, Y. Berthezène, J.C. Baron, Reperfusion within 6 hours outperforms recanalization in predicting penumbra salvage, lesion growth, final infarct, and clinical outcome, *Stroke* 46 (2015) 1582–1589.
- [54] M. Livne, V.I. Madai, P. Brunecker, O. Zaro-Weber, W. Moeller-Hartmann, W. D. Heiss, K. Mouridsen, J. Sobesky, A PET-guided framework supports a multiple arterial input functions approach in DSC-MRI in acute stroke, *J. Neuroimaging* 27 (2017) 486–492.
- [55] L. Røhl, L. Østergaard, C.Z. Simonsen, P. Vestergaard-Poulsen, G. Andersen, M. Sakoh, D. Le Bihan, C. Gyldensted, Viability thresholds of ischemic penumbra of hyperacute stroke defined by perfusion-weighted MRI and apparent diffusion coefficient, *Stroke* 32 (2001) 1140–1146.
- [56] M. Giacalone, C. Frindel, E. Grenier, D. Rousseau, Multicomponent and longitudinal imaging seen as a communication channel - an application to stroke, *Entropy* 19 (2017) 187.
- [57] M.S. Graham, I. Drobnjak, H. Zhang, Realistic simulation of artefacts in diffusion MRI for validating post-processing correction techniques, *Neuroimage* 125 (2016) 1079–1094.

- [58] H.b. Du, L.H. Wang, W.Y. Liu, F. Yang, Z. Li, Y.M. Zhu, Diffusion mri simulation for human brain based on the atlas, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 898–902.
- [59] N. Courty, R. Flamary, D. Tuia, A. Rakotomamonjy, Optimal transport for domain adaptation, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (2017) 1853–1865.
- [60] T. Glatard, C. Lartizien, B. Gibaud, R.F. Da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard, S. Camarasu-Pop, et al., A virtual imaging platform for multi-modality medical image simulation, *IEEE Trans. Med. Imaging* 32 (2013) 110–118.

Data augmentation from RGB to chlorophyll fluorescence imaging Application to leaf segmentation of *Arabidopsis thaliana* from top view images

Natalia Sapoukhina

INRA, UMR1345 Institut de Recherche en Horticulture et Semences IRHS,
SFR 4207, PRES UNAM,
49071 Beaucouzé, France

natalia.sapoukhina@inra.fr

Salma Samiei

salma.samiei@univ-angers.fr

Pejman Rasti

spejman.rasti@univ-angers.fr

David Rousseau

david.rousseau@univ-angers.fr

Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS),
UMR INRA IRHS, Université d'Angers,
62 avenue Notre Dame du Lac, 49000 Angers, France

Abstract

In this report we investigate various strategies to boost the performance for leaf segmentation of Arabidopsis thaliana in chlorophyll fluorescent imaging without any manual annotation. Direct conversion of RGB images to gray levels picked from CVPPP challenge or from a virtual Arabidopsis thaliana simulator are tested together with synthetic noisy versions of these. Segmentation performed with a state of the art U-Net convolutional neural network is shown to benefit from these approaches with a Dice coefficient between 0.95 and 0.97 on the segmentation of the border of the leaves. A new annotated dataset of fluorescent images is made available.

1. Introduction

Due to heavy occlusion, variability in terms of size and shape, leaf segmentation is a challenging task from the computer vision perspective [16]. One strategy to simplify the segmentation is to reduce the biological variability and focus on a limited amount of plant species of specific interest. This has been undertaken in the CVPPP challenge since 2014 with a focus on a few species including *Arabidopsis thaliana* which serves as a reference for a number of fundamental biological questions. The effort to provide finely annotated data [14] has enabled great improvement of the state

of the art on segmentation performance. An open question is now how to transfer this knowledge obtained from RGB images on annotated plants either to other species or other modalities of imaging. In this work, we focus on the transfer of the knowledge gained from annotated leaves of *Arabidopsis thaliana* in RGB to images of the same plant in chlorophyll fluorescence imaging.

2. Related Work

Segmentation of *Arabidopsis thaliana* leaves in RGB images has been highly studied since the introduction of the CVPPP challenge. In 2014 and 2015 the contributions of this challenge proposed methods based on models [20, 27, 21], most of the participants have so far mainly tackled the challenge with deep neural network [29, 26, 31]. In this work we did not propose any innovation on this side and rather work on a standard neural network architecture but applied it for the first time on another imaging modality. We used the U-Net architecture [23] which had been mainly employed for the pixel-wise segmentation of separation boundaries in medical [34] and satellite images [13]. Here, we applied U-Net for the first time to the best of our knowledge on leaf segmentation of *Arabidopsis thaliana* in chlorophyll fluorescence imaging.

Chlorophyll fluorescence analysis is a non-destructive technique which has been developed to probe plant physiology [6]. Among all the chlorophyll fluorescence param-

eters that can be estimated, the maximum quantum yield of photosystem II (PSII) photochemistry ($F_v/F_m = (F_m - F_0)/F_m$) [9] is an indicator of plant stress [22]. Fluorescence chlorophyll by image analysis on whole plant has been widely studied [24, 4, 17]. So far, to the best of our knowledge analysis on individual leaves has not been tackled in top view images of *Arabidopsis thaliana*.

Image simulation to boost machine learning received an increasing interest in plant imaging [35, 10, 1, 8]. This can include standard data augmentation, sophisticated infography or generative models from convolutional networks. In this communication we generated the images from one imaging modality to learn on another imaging modality. This topic has been demonstrated possible for instance for life science applications in the medical domain [12] in a cross modal image synthesis and also in microscopy in a superresolution problem [19]. We considered for the first time data augmentation from the synthesis of images from RGB imaging modality to chlorophyll fluorescence imaging in plant sciences.

3. Method

3.1. Datasets

Three datasets coined *CVPPP*, *CSIRO* and *Real-Fluo* are considered in this study. They are described in the following lines.

CVPPP. We used the dataset provided in the Leaf Segmentation Challenge held as part of the computer vision problems in plant phenotyping *CVPPP* workshop [14]. *CVPPP* dataset consists in 27 RGB images of tobacco plants and 783 RGB images of *Arabidopsis* wild and mutant plants. We considered only the *Arabidopsis* dataset in this study. All images are hand labelled to obtain ground truth masks for each leaf in the scene (as described in [14]). These masks are image files encoded in PNG where each segmented leaf is identified with a unique integer value, starting from 1, where 0 is background.

CSIRO. To extend the *CVPPP* dataset we also used synthetic images of top down view renders of *Arabidopsis* generated with the simulator described in [30, 32]. The *CSIRO* dataset contains 10000 synthetic images (width x height: 550 x 550 pixels). Similarly to *CVPPP* dataset, each RGB image has a corresponding leaf instance segmentation annotation: each leaf in an image is uniquely identified by a single color value, starting from 1, where 0 is background. All images are stored in PNG format.

Real-Fluo. For model testing we used 38 real gray-scale fluorescent images of *Arabidopsis*. The PSI Open Fluor-Cam FC 800-O (PSI, Brno, Czech Republic) was used to capture chlorophyll fluorescence images and to estimate the maximum quantum yield of PSII (F_v/F_m) on wild type control of *Arabidopsis thaliana*. The system sensor is a

CCD camera with a pixel resolution of 512 by 512 and a 12-bit dynamic range. The system includes 4 LED panels divided into 2 pairs. One pair provides an orange actinic light with a wavelength of around 618 nm, with an intensity that can vary from 200 to 400 $\mu\text{mol}/\text{m}^2/\text{s}$. It provides a 2s pulse that allows the measurement of the initial fluorescent state (F_0). The other pair provides a saturating pulse during 1s in blue wavelength, typically 455 nm, with an intensity of up to 3000 $\mu\text{mol}/\text{m}^2/\text{s}$. The saturating pulse allows collecting of the maximum fluorescence (F_m). Fluorescence chlorophyll imaging was used in a dark adapted mode after a dark period of 45 min to produce maps with the fluorescent quantum efficiency $F_v/F_m = (F_m - F_0)/F_m$. All these 38 images were manually annotated using Phenotiki image analysis software [15] and are made available to the reader (see the web link at the end of the article).

3.2. U-Net Model

The segmentation of the leaves was considered to be a pixel-wise classification where the pixel of the leaf contour should be detected among the other pixels of the image. Picking out leaf contours allowed separating leaves and thereby performing leaf segmentation, for example with help of a watershed transform. Each pixel was therefore classified among three mutually exclusive classes: mask without contours, leaf contours and background. It means that every pixel was labeled by a three-component one-hot vector.

The U-Net model [23] was used for the pixel-wise classification. As shown in Figure 1 U-Net architecture is separated in 3 parts: the contracting/downsampling path, the bottleneck, the expanding/upsampling path. The encoder-decoder type architecture with skipped connections allows combining low-level feature maps with higher-level ones, and enables precise pixel classification. A large number of feature channels in upsampling part allows propagating context information to higher resolution layers. The output of the model was a three-channel label that indicated the class of every pixel as shown in Figure 2. All activation functions in the convolutional layers were rectified linear units, ReLU [11]. The last layer before the prediction was a softmax activation with 3 classes. Images and labels from all datasets were resized to width x height: 128 x 128 pixels. Using ground truth (GT) labels, we created three-channel labels as shown in Figure 2. To reinforce the learning of the contour class, which was highly unbalanced, we replaced the encoder by a ResNet152 backbone pre-trained on ImageNet [33]. The decoder was not changed from the original description [23]. We empirically found that the best performances were obtained when all skipped connections were kept which was in accordance with the intrinsic multiscale nature of plants [25]. The resulting U-Net neural network had a total 1,942,275 trainable parameters.



Figure 1: U-Net architecture. Each blue box corresponds to a multi-channel feature map. Gray arrows indicate the merging of the context and localization information that was done by concatenating the features from the contracting path with the corresponding ones in the expansion path. Input image has 128x128 pixels, the output of the model is a three-channel binary image: mask without contours, leaf contours and background.

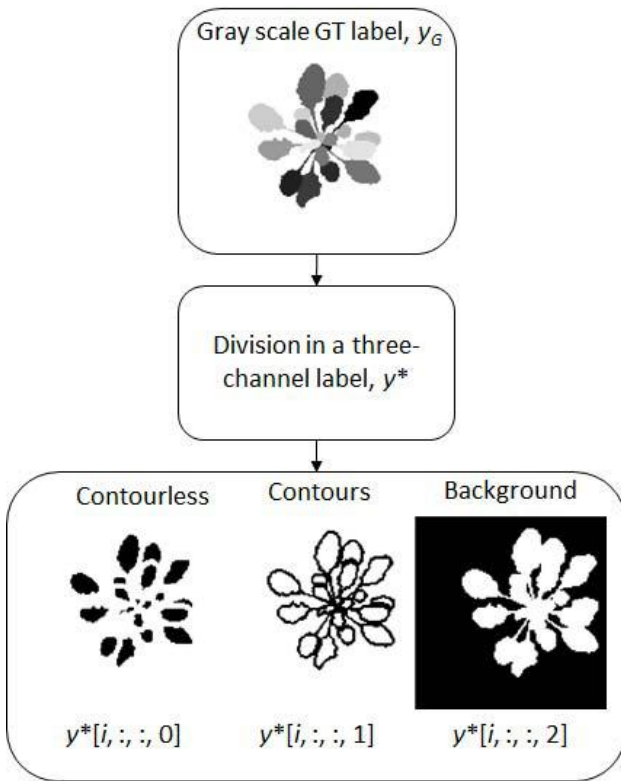


Figure 2: Production of the three-channel binary labels from ground truth (GT) labels: the first channel contains mask without leaf contours, the second channel - leaf contours and the third one - background.

3.3. Data augmentation

Several strategies of data augmentation were investigated from *CVPPP* and *CSIRO* datasets to train our U-Net in order to perform leaf segmentation on the *Real-Fluo* dataset.

In a first simplest strategy, we converted *CVPPP* and *CSIRO* directly from RGB to gray levels along the simple CIE formulae $Gray = .299 * Red + .587 * Green + .114 * Blue$. In a second strategy, we considered binary images such as the ones in Figure 3 column (b) and mapped on them a noisy texture learned from the real fluorescence images, *Real-Fluo*, shown in Figure 3 column (a). A copy of the original binary image for each plant was also kept so as to produce the associated GT. For a first trial of transfer from RGB images to fluorescence images, we propose to test an extremely simple model for the noisy texture which is estimated as an additive Gaussian white noise process independent and identically distributed for a given leaf. This choice was first driven by an Occam's Razor simplicity spirit. Indeed with such a model the simulated leaves have no spatial structures such as vascular veins. Leaves are therefore expected to be distinguished in real images only from their first order statistics. Also, as an additional motivation to test this simple fluorescence chlorophyll simulator, the noise in real fluorescence images is expected to be mostly thermal noise on the camera which will control the standard deviation of the noise. The leaves themselves, if considered to have homogeneous tissue, may have a variety of average values in fluorescence emission depending on their physiological state.

To estimate the parameters of these Gaussian processes, we analyzed the distribution of the gray levels among a small set of images of real plants. In order to ensure that this small set of chlorophyll fluorescence images was representative from the rest of the images we considered one image of plant at each developmental stage represented in the test dataset. Average value and standard deviation of the gray levels inside the plants for both considered chlorophyll fluorescence parameters F_0 and F_m are given in Table 1. The order of magnitude of the average value and standard deviation of the chlorophyll fluorescence parameters F_0 and F_m remained in the same range in our experiment.

Synthetic chlorophyll fluorescent images were then simply produced by adding Gaussian noises with mean μ and variance σ^2 for each fluorescence map $(\mu_{F_0}, \sigma_{F_0}^2)$, and $(\mu_{F_m}, \sigma_{F_m}^2)$, randomly sampled in Table 1. A different realization of these noises was applied for each individual leaf of gray scale GT labels in *CVPPP* and *CSIRO* datasets so as to produce a synthetic fluorescent example x_F given by

$$x_F = 1 - \frac{\sum_l (y_g^{(l)} + \mathcal{N}(\mu_{F_0}, \sigma_{F_0}^2))}{\sum_l (y_g^{(l)} + \mathcal{N}(\mu_{F_m}, \sigma_{F_m}^2))}, \quad (1)$$

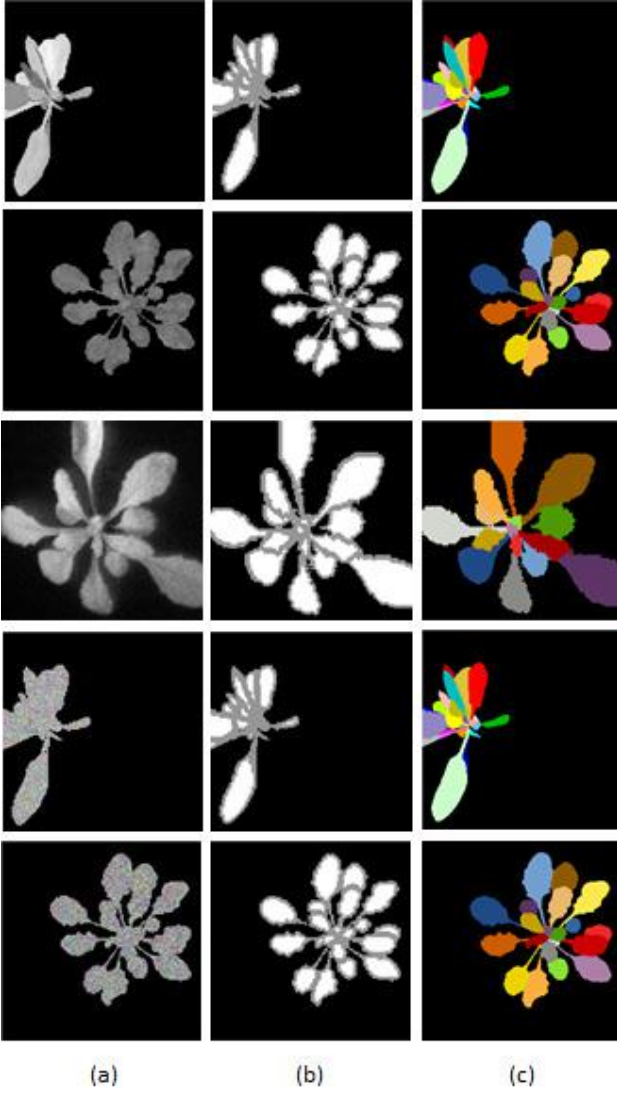


Figure 3: Examples from datasets used for model training and its evaluation. (a) Plant image examples. (b) three-channel labels for pixel-wise classification. (c) Ground truth labels with leaf segmentation. First line: *CSIRO* dataset, 783 examples. Second line: *CVPPP* dataset, 783 examples. Third line: *Real-Fluo* dataset, 38 examples. Forth line: *CSIRO-Fluo* dataset, 5481 examples. Fifth line: *CVPPP-Fluo* dataset, 5481 examples. Number of examples in datasets are given before application of the standard data augmentation.

where $y_g^{(l)}$ is l th binary leaf from a gray scale GT label and $\mathcal{N}(\mu_{F_0}, \sigma_{F_0}^2)$ is a Gaussian noise realization. For every GT label we produced 7 synthetic fluorescent examples x_F by drawing random values for μ_{F_0}, σ_{F_0} and μ_{F_m}, σ_{F_m} from Table 1. The pipeline of data augmentation is shown

in Figure 4.

As a result, in addition to *CVPPP* and *CSIRO*, we obtained new datasets, *CVPPP-Fluo* and *CSIRO-Fluo*, containing $5481 = 783 * 7$ and $70000 = 10000 * 7$ synthetic fluorescent images (width x height: 128×128 pixels), respectively. Now, our objective is to compare the added value of all these datasets for leaf segmentation in *Real-Fluo* dataset with the U-Net model presented in previous section.

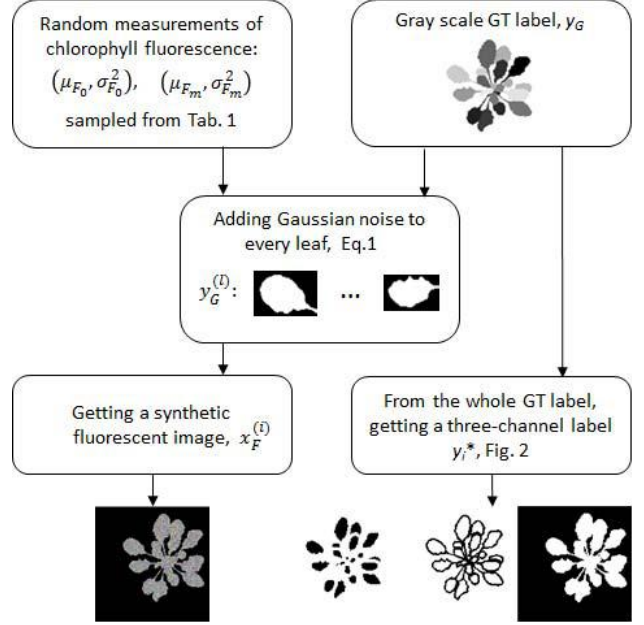


Figure 4: Data augmentation using synthetic fluorescent training data. For each gray-scale GT label from *CVPPP* or *CSIRO* datasets we produced fluorescent images and associated three-channel labels.

Time	μ_{F_0}	σ_{F_0}	μ_{F_m}	σ_{F_m}
Day 1	167.83	34.88	180.77	24.68
Day 5	165.81	33.1	180.00	22.36
Day 6	164.48	30.87	177.9	20.8
Day 7	158.16	31.45	174.73	21.1
Day 8	165.24	32.31	181.14	21.36
Day 9	168.3	28.03	184.36	17.86
Day 12	173.06	28.01	189.96	17.15

Table 1: Mean, μ , and standard deviation, σ for chlorophyll fluorescence F_0, F_m estimated on a single plant from *Real-Fluo* dataset at different dates after emergence of first leaves (cotyledons).

3.4. Watershed Post-Processing

To segment leaves with use of estimated three-channel labels, we applied the classical marker-controlled watershed segmentation [3, 2]. The markers were generated with a contourless mask from output three-channel label and then, to segment leaves, we flooded marked “basins” within the bounds of mask.

4. Experiment and Results

4.1. Training

On top of the data augmentation techniques that we generated from *CVPPP* and *CSIRO* datasets as described in section 3.3, we apply a standard data augmentation strategy in order to further reduce overfitting and improve generalization. For this data augmentation we used Albumentations library [5]. While the data augmentation strategies of section 3.3 focused on contrast and noise distribution, here we generated geometrical transformation such as horizontal flip, vertical flip, random rotate at 90 degree and random half-sized crop and applied them to shuffled training dataset.

It was shown that for high level of imbalance, loss functions based on overlap measures appeared to be more robust [28]. Through all of our experiments, we minimized weighted combination of multi-class cross entropy and dice losses

$$L(y, y^*) = w_0 C(y, y^*) + w_1 (1 - D(y[\dots, 0], y^*[\dots, 0])) + w_2 (1 - D(y[\dots, 1], y^*[\dots, 1])). \quad (2)$$

$C(y, y^*)$ is the categorical cross entropy defined as

$$C(y, y^*) = - \sum_{ij} y_{ij} \log y_{ij}^* \quad (3)$$

and $D(y, y^*)$ is the Dice coefficient

$$D(y, y^*) = \frac{2 \sum_{ij} y_{ij} y_{ij}^* + \epsilon}{\sum_{ij} y_{ij} + \sum_{ij} y_{ij}^* + \epsilon}, \quad (4)$$

where y is a model prediction with values y_{ij} , y^* is a ground truth label with values y_{ij}^* and $\epsilon = 0.001$ is used here to ensure the coefficient stability by avoiding the numerical issue of dividing by 0. The weight ratios (w_0, w_1, w_2) used to correct the class imbalance was respectively 0.4, 0.1, 0.5 for cross entropy, contourless mask and contours. Adam optimizer was used with default parameters $l_r = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Our training procedure consisted of splitting the data into 80% and 20% training and cross validation respectively. We shuffled the dataset examples at the beginning of each epoch and used a batch size of 16 examples. We also implemented batch normalization before each activation.

Leaf segmentation in fluorescence images was done with different data augmentation strategies for the training based on the datasets of Figure 3 and their combinations. A base line consisted in training directly on the *CVPPP* or *CSIRO* RGB to gray images. The learning from the simulated fluorescence dataset either generated along Eq. (1) from *CVPPP* labels and/or *CSIRO* labels was tested for comparison. The previous strategies were tested also when small amount of real fluorescence images were added in the training. The eight different tested training strategies are summarized in Table 2.

4.2. Results

To assess the quality of segmentation, we used the soft Dice coefficient, Eq. (4), that was computed separately for all pixels and for leaf contours. Furthermore, the pixel-wise accuracy was evaluated in order to get a general idea of the model performance. It was computed as the ratio between correctly classified pixels and the total number of pixels in the test sample. To assess the performance of leaf contour detection we computed additional metrics. True positives (TP) are contour pixels present in both prediction and GT mask. False positives (FP) are contour pixels present in prediction but absent in GT mask. False negatives (FN) are contour pixels absent in prediction but present in GT mask. Knowing these numbers we can estimate true positive rate

$$TPR = \frac{TP}{TP + FN}, \quad (5)$$

that describes the fraction of correctly classified contour pixels in comparison of the total number of contour pixels in GT mask. Moreover, positive predictive value

$$PPV = \frac{TP}{TP + FP}, \quad (6)$$

gives us the fraction of correctly classified contour pixels among all predicted contour pixels.

Table 2 displays the model performance on the *Real-Fluo* dataset for eight model training experiments. A first global observation is that the performance of training on *CVPPP* alone was rather high. This demonstrates a high similarity of RGB reflectance images converted to gray levels and the fluorescence images despite the physical differences in the mechanism of their production. Training on *CVPPP-Fluo* and *CSIRO-Fluo* alone or combined did not provide better performances than *CVPPP* alone. The best model Dice score was 97% obtained for extended *CVPPP* and *CVPPP-Fluo* datasets with 10 examples from *Real Fluo* dataset. The use of small quantity of real fluorescent images among images with modeled fluorescence resulted in Dice score gain of 2-3% in comparison with *CVPPP* and *CVPPP-Fluo* datasets. The same positive effect of the injection of 10 real fluorescent images on the model performance

Training dataset	Accuracy	L_{train}	D_{train}	L_{test}	D_{test}	D_c	TPR	PPV
<i>CVPPP</i>	0.96	0.03	0.98	0.19	0.95	0.67	0.74	0.62
<i>CVPPP-Fluo</i>	0.96	0.01	0.99	0.22	0.94	0.68	0.82	0.58
<i>CSIRO-Fluo</i>	0.94	0.02	0.99	0.27	0.92	0.46	0.48	0.46
<i>CVPPP-Fluo + CSIRO Fluo</i>	0.95	0.01	0.99	0.3	0.93	0.62	0.63	0.62
<i>CVPPP + 10ex</i>	0.98	0.04	0.98	0.05	0.97	0.84	0.8	0.87
<i>CVPPP-Fluo + 10ex</i>	0.98	0.02	0.99	0.05	0.97	0.84	0.83	0.85
<i>CSIRO-Fluo + 10ex</i>	0.98	0.03	0.98	0.06	0.96	0.84	0.85	0.82
<i>CVPPP-Fluo + CSIRO Fluo + 10ex</i>	0.98	0.04	0.97	0.06	0.96	0.8	0.8	0.82

Table 2: Performance metrics on *Real-Fluo* samples for training strategies with the various strategies tested of data augmentation. To estimate the overall model performance, we used pixel-wise Accuracy. To assess overfitting, we calculated loss function, Eq. (2), and Dice coefficient, Eq. (4), for the training dataset, L_{train} , D_{train} , and for the test dataset, L_{test} , D_{test} , respectively. The accuracy of contour pixels detection was evaluated by means of Dice coefficient, D_c , true positive rate, TPR , and positive predictive value, PPV . See Eq. (5, 6) for the last two estimates.

was observed for the other datasets as well, *CSIRO-Fluo + 10ex* and *CVPPP-Fluo + CSIRO Fluo + 10ex*. It increased overall pixel-wise accuracy to 0.98 and the quality of contour detection became quite high as well: $D_c \in [0.8, 0.84]$, $TPR \in [0.8, 0.83]$, $PPV \in [0.82, 0.87]$. In more details, the comparison of model performance from training on *CVPPP* and *CVPPP-Fluo* showed that the imitation of fluorescence by modelling increased TRP by 9% and decreased PPV by 6%. It means that the modelled fluorescence allowed us to detect a little bit better leaf contours on real fluorescent plant images but at the same time it had the tendency to classify surplus pixels as contour pixels. Training on *CSIRO-Fluo* had the lowest values of metrics in comparison with the other training strategies. However these can be considered as interesting results if one keeps in mind that in this case the network was trained only on purely synthetic datasets. Probably, there is a deficiency of important information of leaf texture in synthetic plants from *CSIRO-Fluo* that prevents the simulation of fluorescence in a sufficient realistic way.

As shown in Figure 5, with the worst example from the best training strategy, most errors of pixel classification occurred for occluded leaves, i.e. for really difficult cases. Another source of discrepancies was an inaccurate annotation of some contour pixels. It means that some pixels were correctly classified as contour but since they were present in GT label with displacement they were not counted in true positive rate. However, this type of errors did not prevent the correct segmentation as it is shown in the upper line of Figure 6. Only two cases of occlusive leaves were not segmented. These kind of discrepancies could potentially be solved using a more advanced post-processing method. Overall, the segmentation performance was higher for young small plants where there was not a lot of leaf occlusion as it is shown in the lower line of Figure 6.

5. Conclusion and Discussion

In this paper, we studied the transfer of knowledge for leaf segmentation learned from RGB imaging to fluorescence imaging. Various data augmentation strategies were tested with real images of plants or on pure synthetic plants and from RGB to gray conversion up to a physical modelling of noise in fluorescence.

This was illustrated on *Arabidopsis thaliana* which is one of the most studied plant for fundamental biology and with the U-Net neural network architecture applied for the first time in this context. We have demonstrated that existing annotated datasets in RGB could be used to learn to segment leaves in fluorescence images by a simple RGB to gray conversion. Also, good performances (although not the best) of segmentation could be obtained by learning on purely synthetic datasets automatically annotated and mapped with a first order statistics physical modelling of noise in fluorescence. Segmentation performance were found higher when some real images were also introduced in the training process.

These results could be extended in various promising directions. First, one could try to improve the segmentation result presented here. Other neural network architectures could for instance be tested such as the one recently introduced to consider segmentation as a regression [18]. Also performances on training from simulated datasets could benefit from domain adaptation [7] to compensate for the necessarily non perfect match between simulation and reality. Other plant imaging modalities could finally be also investigated in the same way as in this communication. One could for instance think to thermal imaging or Tera hertz imaging which are also used to assess the physiological state of leaves. There are currently no annotated datasets for these images and it would therefore be interesting to explore if data augmentation from other imaging in which annotated

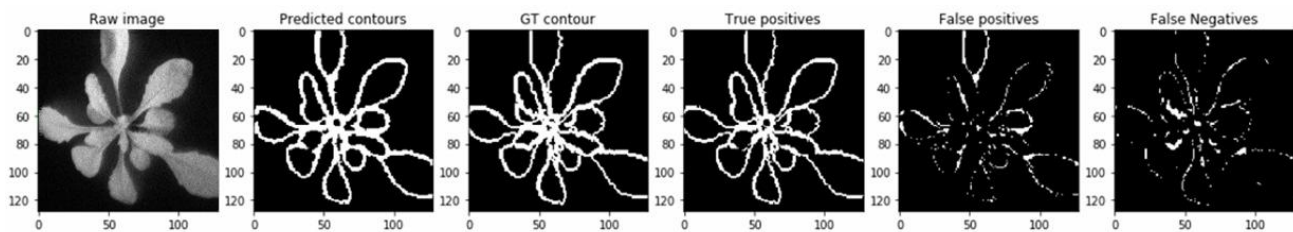


Figure 5: Example of leaf contours detection in a *Arabidopsis* fluorescent image. In this case the model was trained on *CVPPP-Fluo + 10ex* dataset including 5481 images with imitated fluorescence and 10 real fluorescent images. True positives show well predicted contour pixels existing in GT label. False positives show surplus pixels that were classified as the contour but did not exist in GT label. False negatives are pixels that had to be classified as contour pixels since they were presented in GT label.

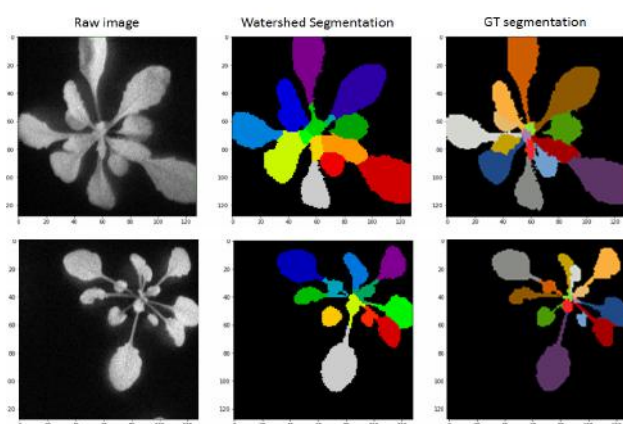


Figure 6: Examples of watershed segmentation produced by the model trained on *CVPPP-Fluo + 10ex*. Upper line: $D_c = 0.8$, $Accuracy = 0.96$, $TPR = 0.82$, $PPV = 0.78$. Lower line: $D_c = 0.84$, $Accuracy = 0.98$, $TPR = 0.83$, $PPV = 0.86$.

datasets are available could be helpful.

To contribute to reproducible science, we have opened, as an additional result from our study, access to our annotated dataset of *Arabidopsis thaliana* in fluorescence imaging (<https://uabox.univ-angers.fr/index.php/s/BglUZgoE5EWK4MM>).

6. Acknowledgements

Authors thank Etienne Belin and Tristan Boureau from Platform PHENOTIC part of PHENOME phenotyping network for the production of *Arabidopsis* and associated fluorescence images. Salma Samiei gratefully acknowledges Angers Loir Metropole for the funding of her PhD Grant.

References

- [1] R. Barth, J. IJsselmuiden, J. Hemming, and E. J. Van Henten. Data synthesis methods for semantic segmentation in agriculture: A capsicum annum dataset. *Computers and Electronics in Agriculture*, 144:284–296, 2018.
- [2] S. Beucher. The watershed transformation applied to image segmentation. *Scanning Microscopy International*, 6(1):299–314, 1991.
- [3] S. Beucher and C. Lantuejoul. Use of watersheds in contour detection. In *Proc. Int. Workshop Image Processing, Real-Time edge and Motion Detection/Estimation*, 1976.
- [4] J. Bresson, F. Vasseur, M. Dauzat, G. Koch, C. Granier, and D. Vile. Quantifying spatial heterogeneity of chlorophyll fluorescence during plant growth and in response to water stress. *Plant Methods*, 11(1):23, 2015.
- [5] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin. Albuementations: fast and flexible image augmentations. *ArXiv e-prints*, 2018.
- [6] W. L. Butler. Energy distribution in the photochemical apparatus of photosynthesis. *Annual Review of Plant Physiology*, 29(1):345–378, 1978.
- [7] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.
- [8] M. Di Cicco, C. Potena, G. Grisetti, and A. Pretto. Automatic model based dataset generation for fast and accurate crop and weeds detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5188–5195. IEEE, 2017.
- [9] B. Genty and S. Meyer. Quantitative mapping of leaf photosynthesis using chlorophyll fluorescence imaging. *Functional Plant Biology*, 22(2):277–284, 1995.
- [10] M. V. Giuffrida, H. Scharf, and S. A. Tsafaris. Arigan: Synthetic arabidopsis plants using generative adversarial network. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy*, pages 22–29, 2017.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proc. IEEE*, 2015.

- [12] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato. Cross-modality image synthesis from unpaired data using cycle-gan. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 31–41. Springer, 2018.
- [13] V. Iglovikov, S. Seferbekov, A. Buslaev, and A. Shvets. Ternausnetv2: Fully convolutional network for instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [14] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsafaris. Finely-grained annotated datasets for image-based plant phenotyping. pages 1–10, 2015.
- [15] M. Minervini, M. V. Giuffrida, and S. A. Tsafaris. An interactive tool for semi-automated leaf annotation. *Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP) Workshop*.
- [16] M. Minervini, H. Scharr, and S. A. Tsafaris. Image analysis: the new bottleneck in plant phenotyping [applications corner]. volume 32, pages 126–131. IEEE, 2015.
- [17] A. M. Mutka, S. J. Fentress, J. W. Sher, J. C. Berry, C. Pretz, D. A. Nusinow, and R. Bart. Quantitative, image-based phenotyping methods provide insight into spatial and temporal dimensions of plant disease. *Plant physiology*, 172(2):650–660, 2016.
- [18] P. Naylor, M. Laé, F. Reyat, and T. Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2):448–459, 2019.
- [19] C. Ounkomol, S. Seshamani, M. M. Maleckar, F. Collman, and G. R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature methods*, 15(11):917, 2018.
- [20] J.-M. Pape and C. Klukas. 3-d histogram-based segmentation and leaf detection for rosette plants. In *European Conference on Computer Vision*, pages 61–74. Springer, 2014.
- [21] J.-M. Pape and C. Klukas. Utilizing machine learning approaches to improve the prediction of leaf counts and individual leaf segmentation of rosette plant images. *Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP)*, pages 1–12, 2015.
- [22] S. A. Rolfe and J. D. Scholes. Chlorophyll fluorescence imaging of plant–pathogen interactions. *Protoplasma*, 247(3-4):163–175, 2010.
- [23] O. Ronneberger, P. Fischer, and T. Brox. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.
- [24] C. Rousseau, E. Belin, E. Bove, D. Rousseau, F. Fabre, R. Berruyer, J. Guillaumès, C. Manceau, M.-A. Jacques, and T. Boureau. High throughput quantitative phenotyping of plant resistance using chlorophyll fluorescence image analysis. *Plant methods*, 9(1):17, 2013.
- [25] D. Rousseau, Y. Chéné, E. Belin, G. Semaan, G. Trigui, K. Boudehri, F. Franconi, and F. Chapeau-Blondeau. Multi-scale imaging of plants: current approaches and challenges. *Plant methods*, 11(1):6, 2015.
- [26] H. Scharr, T. P. Pridmore, and S. A. Tsafaris. Computer vision problems in plant phenotyping, cvppp 2017–introduction to the cvppp 2017 workshop papers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2021, 2017.
- [27] K. Simek and K. Barnard. Gaussian process shape models for bayesian segmentation of plant leaves. *Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP)*, pages, pages 4–1, 2015.
- [28] H. C. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *ArXiv e-prints*, 2017.
- [29] S. A. Tsafaris, M. Minervini, and H. Scharr. Machine learning for plant phenotyping needs image processing. *Trends in plant science*, 21(12):989–991, 2016.
- [30] D. Ward and P. Moghadam. Synthetic arabidopsis dataset. *CSIRO. Data Collection.*, 2018.
- [31] D. Ward, P. Moghadam, and N. Hudson. Deep leaf segmentation using synthetic data. *arXiv preprint arXiv:1807.10931*, 2018.
- [32] D. Ward, P. Moghadam, and N. Hudson. Deep leaf segmentation using synthetic data. In *Proceedings of the British Machine Vision Conference (BMVC) Workshop on Computer Vision Problems in Plant Phenotyping (CVPPP)*, 2018.
- [33] P. Yakubovskiy. Segmentation models. https://github.com/qubvel/segmentation_models, 2019.
- [34] A. Zaimi, M. Wabartha, P. L. A. V. Herman, C. S. Perone, and J. Cohen-Adad. Axondeepseg: automatic axon and myelin segmentation from microscopy data using convolutional neural networks. *Scientific Reports*, 8(1):38216, 2018.
- [35] Y. Zhu, M. Aoun, M. Krijn, J. Vanschoren, and H. T. Campus. Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants. *Computer Vision Problems in Plant Phenotyping (CVPPP2018)*, 2018.

