Problèmes de segmentation d'images et contribution à la morphologie mathématique

Robin Alais

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

Image Segmentation Problems and Contribution to Mathematical Morphology
Problèmes de segmentation d'image et contribution à la morphologie mathématique

Soutenue par
Robin ALAIS
Le 4 janvier 2021

École doctorale n°621
Ingénierie des systèmes, matériaux, mécanique, énergétique

Spécialité
Morphologie mathématique

Composition du jury :

Antoine MANZANERA Professeur, ENSTA Paris	*Président*
Florence ROSSANT Professeur, Institut Supérieur d'Electronique de Paris	*Rapporteure*
Valery NARANJO Professeur, Universitat Politècnica de València	*Rapporteure*
Etienne DECENCIERE Directeur de recherche, MINES ParisTech	*Directeur de thèse*
Petr DOKLADAL Maître de recherche, MINES ParisTech	*Co-directeur de thèse*
Bruno FIGLIUZZI Maître Assistant, MINES ParisTech	*Examinateur*

MINES ParisTech | PSL

Image Segmentation Problems and Contribution to Mathematical Morphology

Robin Alais

Remerciements

This is the end. La fin de la rédaction de ce manuscrit pour moi, et probablement la fin de la lecture pour la majorité des lecteurs, qui viennent simplement vérifier ici si je les remercie à leur juste valeur, voire espèrent trouver quelques traits d'humour assez courants dans cet exercice. Essayons de ne pas trop décevoir ceux-ci ; quant à celles et ceux qui auraient ouvert cette thèse pour son contenu scientifique, je leur souhaite une agréable lecture, même si j'ai bien conscience qu'employer le pluriel dans cette phrase relève de l'optimisme, sinon de la science-fiction.

Pour commencer, je tiens à remercier les membres de mon jury de thèse d'avoir accepté d'évaluer mon travail. Merci aux rapporteures Valery Naranjo et Florence Rossant pour leur lecture attentive de ce manuscrit, leurs remarques et suggestions, à l'écrit comme à l'oral pendant la soutenance. Merci également à Antoine Manzanera d'avoir accepté de présider le jury.

Il est difficile d'exprimer en quelques lignes ma gratitude envers mes encadrants, Etienne, Petr et Bruno, qui ont su m'accompagner dans mes pérégrinations scientifiques, de la morphologie mathématique aux réseaux de neurones, entre thermogrammes et images rétiniennes en passant par les valeurs d'extinction, plus théoriques mais très riches et qui ont au moins le mérite de justifier la présence de belles photos de montagne dans ce manuscrit. Au-delà de l'encadrement scientifique, l'aspect humain est tout autant — sinon plus — important pour réussir à mener à bien une thèse, inévitablement parsemée de moments plus difficiles. "Ca y est, ça fonctionne" le lundi, suivi de "j'ai refait des tests, en fait ça ne fonctionne pas du tout" le mardi étant un exemple typique de ce qui peut arriver avec les réseaux de neurones ; "ça ne fonctionne pas du tout" du lundi au jeudi, suivi de "écoute, je n'ai aucune idée de comment, mais ça a l'air de fonctionner" le vendredi étant un exemple typique de ce qui a pu arriver lorsque des partenaires hostiles à des formats civilisés me demandaient de rédiger une documentation avec Microsoft Word. Le monde extérieur n'ayant pas eu la courtoisie de s'arrêter entre le début de ma thèse et l'écriture de ces lignes, il faudrait également évoquer des difficultés plus personnelles pour être exhaustif, aussi je ne le serai pas et me concentrerai plutôt sur les bons souvenirs. Des moments passés à chercher de nouvelles idées, comprendre les limitations de la précédente ou pourquoi elle n'a pas marché, à recouvrir le tableau de dessins à la recherche d'une propriété ou d'un contre-exemple. Des articles enfin acceptés ("Dear reviewer 2, thank you very much for your suggestion to entirely redo everything. However, ..."). Des moments de détente autour du repas du midi, autour d'un café (l'emploi du mot café au singulier est un mensonge) ... Pour tout ce que vous avez fait pendant ces années, un grand merci. J'ai eu beaucoup de chance de vous avoir comme encadrants.

Je tiens ensuite à remercier tout particulièrement Emilie. Le battement d'ailes de papillon qui aboutit des années plus tard (entre autres) à ce manuscrit, c'est le fait de t'avoir rencontrée. Merci pour tes conseils avisés, ton goût partagé pour la rigueur aussi bien mathématique que syntaxique, et pour ton amitié précieuse.

J'ai eu la chance de faire cette thèse au Centre de Morphologie Mathématique et d'y faire pléthore de rencontres enrichissantes. Merci à Michel pour son écoute et ses conseils, dans le cadre professionnel comme autour d'une partie de billard ; merci à Jesús qui apportait également une vision complémentaire, notamment en tant que membre du comité de suivi de thèse. Merci à Beatriz pour les

nombreux échanges, parfois mathématiques, parfois non, autour d'un café, et qui m'a sauvé plusieurs fois lorsqu'il s'agissait de retrouver des articles classiques aussi incontournables qu'inaccessibles. Merci à François de nous avoir montré d'autres applications, plus axées matériaux que proprement traitement d'images. Merci à Fernand d'avoir partagé sa culture en morphologie mathématique. Merci à José pour sa disponibilité et ses précieux conseils. Merci à Santiago pour sa bonne humeur, et à Samy pour son soutien. Un immense merci à l'indispensable Anne-Marie, que ce soit pour sa maîtrise des arcanes administratives ou sa grande gentillesse, dans le cadre professionnel et en dehors.

Merci à mes co-bureaux successifs, Serge, Jean-Charles, Sara, Théo, Andres, fans parfois malgré eux de guitare, de café et d'espagnol approximatif, ainsi qu'aux voisins Bassam, Enguerrand, Haisheng, Amin, Eric, Leonardo, François. Merci à Seb, organisateur infatigable de soirées jeux mais aussi de groupes de travail Machine Learning entre doctorants. Merci à Vaïa d'avoir essayé de nous apprendre la salsa avec plus ou moins de succès, mais surtout pour une amitié qui perdure ; merci à Albane - entre autres choses - de ne pas avoir essayé de m'apprendre la salsa et pour ces soirées passées à rendre notre culture cinématographique mutuelle moins désastreuse. Merci à JB, jamais à court de sujets de conversation, pour le meilleur et pour le pire. Merci à Elodie, sportive émérite sur le terrain de basket comme en parcourant les 25 bosses avec aisance et légèreté. Merci également à Kaïwen, ce fut toujours un plaisir d'échanger avec toi.

Hors du CMM, il est aussi possible de faire des rencontres dont l'écrasante majorité a un domaine de compétences commençant par "géo" quelque chose. Une notable exception est Pierre, informaticien et néanmoins ami, et partner in crime de multiples soirées et projets. Merci à Aurélien, jeu-de-sociéteur et randonneur infatigable voire franchement énervant quand tu entames ton huitième litre d'eau par 40 degrés et qu'il rajuste son écharpe. Merci à son meilleur ami Nico, patient professeur de bridge et amateur de jeux de mots que la décence m'empêche d'expliciter ici. Merci également à Simona, Bob, Maëlle, Thibaut, Bibiche, Martin, Léo, joueuses et joueurs de tarot de talent variable mais à la compagnie toujours agréable. Merci à Arezki, joueur de tarot audacieux et avec qui j'ai eu la chance de collaborer pour un bel article de géologie. La meilleure pour la fin (de ce paragraphe), ma partenaire de pauses et de concerts d'Iron Maiden, toujours là dans les bons moments comme les plus délicats, merci Angélique.

Dans des étages peuplés de gaussiennes et de variogrammes, merci aux géostat ; Anna, Jihane, Ricardo, Léa, Mathieu, Mike, Alan. Merci à Jean, camarade de pauses, de randonnées et de projets plus ou moins improbables. Merci à ma thésarde Laure pour ces heures de maths, de raling, de musique et ce bel article auquel j'ai eu la chance de contribuer. Merci à Lantu pour nos échanges mathématiques et musicaux, toujours enrichissants. Merci Marine, la meilleure pour tout sauf pour déceler le sarcasme (ce n'est pas du sarcasme).

Merci les sguibs d'être la meilleure bande d'amis dont on puisse rêver. Vous avez des goûts musicaux discutables mais je vous aime quand même.

Merci enfin à ma famille.

Contents

Introduction

Cette thèse s'articule autour de deux projets : ATHENA, pour Active Thermography for Nondestructive inspection Automation, et RetinOptic. Ces deux projets sont liés à la même problématique de traitement d'image, à savoir la détection d'objet et la segmentation. Ces dernières années, les réseaux de neurones convolutionnels se sont avérés très efficaces pour diverses tâches de vision par ordinateur, et constituent à présent l'état de l'art pour plusieurs problèmes de détection d'objet et de segmentation. Les principaux désavantages des techniques d'apprentissage profond sont le besoin d'une base de données suffisamment grande et représentative, ainsi que le manque d'interprétabilité des modèles obtenus. Malgré les performances impressionnantes des algorithmes d'apprentissage profond pour des tâches complexes, lorsque le nombre d'images est petit ou qu'une annotation fiable n'est pas disponible, il est pertinent de proposer des techniques de segmentation plus classiques, ne reposant pas sur l'apprentissage supervisé, en particulier si l'interprétabilité est nécessaire. C'est le cas dans le projet ATHENA; a contrario, dans le projet Retinoptic, nous avons pu constituer progressivement une base de données de plusieurs milliers d'images, et entraîner des réseaux de convolution.

Dans la première partie, les objects à détecter peuvent être décrits sommairement comme "une zone claire immédiatement à gauche d'une zone sombre" et "une zone claire à droite d'une zone sombre, potentiellement séparées par une zone grise plate". Le nombre d'images à disposition est assez petit, et l'interprétabilité du modèle requise, ce qui exclut l'utilisation de techniques avancées d'apprentissage automatique. En plus de fournir un algorithme de segmentation efficace, une partie du travail consiste à donner des définitions de concepts jusqu'ici mal définis ou dépendants de l'opérateur, comme la symétrie du signal et le rapport signal/bruit. Il s'avère que notre stratégie, à la fois pour segmenter les défauts et les caractériser, repose sur l'étude de certains extrema. Dans les deux premiers chapitres, cette analyse concerne des signaux 1D et reste assez basique, ce qui est suffisant pour la tâche à accomplir. Une réflexion plus poussée est développée dans le chapitre suivant afin de caractériser et quantifier les extrema dans un cadre plus général : nous étendons la notion classique de valeurs d'extinctions pour définir de nouveaux attributs pour les maxima, ainsi que de nouvelles décompositions morphologiques et de nouveaux opérateurs morphologiques. Ce chapitre, théorique, peut être lu indépendamment des autres.

Dans la seconde partie, le but est d'estimer si la région de la macula d'une image rétinienne est de suffisamment bonne qualité, et de localiser la macula le cas échéant. La solution proposée doit également être rapide, ainsi que suffisamment légère pour tourner sur des systèmes embarqués. Parmi les quelque 100.000 images à notre disposition, nous avons progressivement annoté un peu plus de 6000, indiquant pour chacune si la macula était clairement visible et entièrement à l'intérieur de l'image, ainsi que sa position lorsque tel était le cas. Nous avons étudié différents types de réseaux de neurones, à la fois pour la classification (la macula est-elle visible ?) et la régression (pour prédire les coordonnées de la fovéa). Nous avons aussi essayé de fournir un autre type d'entrée aux réseaux, basée sur une décomposition morphologique qui, intuitivement, était susceptible de rendre la tâche plus facile en supprimant ou en atténuant les artefacts d'illumination, et nous avons essayé de comprendre et de corriger les comportements inattendus de nos premiers réseaux de régression. Les résultats décrits dans cette partie, négatifs comme positifs, doivent être interprétés avec prudence. Le but n'est pas

ici de proposer des règles de conduite universelles pour l'apprentissage profond, mais les situations rencontrées et les solutions proposées — qu'elles fonctionnent ou non — sont des exemples des obstacles rencontrés lorsque l'on utilise l'apprentissage profond sur des données réelles.

La macula est le "centre de la vision" : la fovéa, au centre, est la partie de la rétine qui concentre le plus de cônes, qui sont les cellules permettant la vision précise et en couleur dans de bonnes conditions de lumière. Ainsi, les lésions dans la région de la macula peuvent rapidement altérer la vision centrale, et pour qu'un diagnostic soit fait sur une image rétinienne — que ce soit par un médecin ou par un algorithme — il est crucial que la qualité soit suffisante dans cette région. Dans un réseau de télémédecine, il n'est malheureusement pas rare (environ 10% des examens) que la qualité de l'image soit insuffisante pour effectuer un diagnostic. Un des objectifs du projet Retinoptic consistait à concevoir un nouveau rétinographe portable, permettant un dépistage plus large de la population. Cependant, les systèmes portables fournissent généralement des images de moins bonne qualité que les rétinographes sur table, ce qui constitue la motivation principale de notre estimateur de qualité.

Plan de la thèse

Ce document comporte deux parties, concernant respectivement les projets ATHENA et Retinoptic. Le chapitre 1 présente la thermographie active, le cas particulier du projet ATHENA, et présente le problème traité dans cette partie. Le chapitre 2 détaille la solution développée, basée sur l'analyse de certains extrema d'intérêt, et fournit des exemples de segmentation.

Le chapitre 3 constitue la contribution de cette thèse à la morphologie mathématique : nous étendons l'analyse des extrema par valeurs d'extinction, et nous définissons plusieurs décompositions ainsi que plusieurs nouveaux opérateurs morphologiques, et étudions leurs propriétés.

La partie II concerne le projet RetinOptic : au chapitre 4, nous présentons le contexte du projet et motivons notre objectif de détecter la macula. Le chapitre 5 présente des réseaux de classification permettant de distinguer les images où la macula est visible de celles où elle ne l'est pas. Au chapitre 6, nous étudions la localisation de la macula par des réseaux de régression, sous l'hypothèse que nous disposerions déjà d'une bonne classification. Nous étudions la possibilité d'utiliser en entrée des réseaux une autre représentation de nos images, utilisant une décomposition morphologique choisie de manière à rendre la tâche plus simple. Nous proposons également plusieurs solutions pour résoudre un problème dû à l'anisotropie de notre base de données : les réseaux semblent ne prédire correctement qu'une des deux coordonnées de la fovéa. Enfin, au chapitre 7, nous présentons une autre approche du problème, en utilisant un réseau de segmentation entièrement convolutionnel, capable de fournir à la fois classification et régression.

Introduction

This thesis is articulated around two distinct projects: ATHENA, standing for Active Thermography for Nondestructive inspection Automation, and RetinOptic. Both projects are related to the same problematic in image processing, namely object detection and segmentation. In the past few years, convolutional neural networks have been proven very efficient for addressing a variety of computer vision tasks, and have now become the state-of-the-art approaches for several object detection and segmentation problems. The main drawbacks of deep learning techniques are the need for a large, and representative enough dataset, and the lack of interpretability of the resulting models. Despite the impressive performances of deep learning algorithms on complicated tasks, when working on a small number of images and in the absence of a reliable ground-truth annotation, developing more classic image segmentation techniques, not based on supervised learning, is still relevant, especially in the case where interpretability is required. This is the case of the ATHENA project; in contrast, in the Retinoptic project, we were able to progressively constitute a database of several thousand images, and to train convolutional networks.

In the first part, the objects to detect can be roughly described as "a bright zone immediately to the left of a dark zone" and "a bright zone to the right of a dark zone, possibly with a flat gray zone in between". The number of images available was quite small, and model understandability was a requirement, both of which excluded the use of advanced machine learning techniques. In addition to providing an efficient segmentation algorithm, part of the work consisted in giving proper definitions to so far ill-defined or operator-dependent concepts, such as signal symmetry and signal/noise ratio. It turns out that our strategy, both for segmenting defaults and characterizing them, relies on the study of certain extrema. In the first two chapters, this analysis concerns one-dimensional signals and remains very basic, which is sufficient for the task at hand. We next delve further into the reflexion of how to characterize and quantify extrema: we expand the classic idea of extinction values to define new features of maxima, as well as morphological decompositions and new morphological operators. Although the idea stemmed from the work on the ATHENA project, this chapter is self-contained and can be read independently.

In the second part, the aim is to estimate whether the macular region of a retinal image is of good enough quality, and to locate it if it is. Additionally, the proposed solution must be fast, and light enough to run on embedded systems. Out of the more than 100,000 images at our disposal, we progressively annotated a little more than 6,000 of them, indicating whether the macula was clearly visible and entirely within the image, and its position when it was. We investigated different kinds of neural networks, both for classification (is the macula visible?) and regression (for predicting the fovea's coordinates). We also tried feeding the networks a different kind of input, based on a morphological decomposition that, intuitively, might have made the task easier by suppressing or attenuating illumination artifacts, and tried to understand and correct odd behaviors of our first regression networks. The results we describe in this part, both negative and positive, must be interpreted with caution. Our goal is not to provide universal guidelines or rules for deep learning, but the situations encountered and the solutions proposed — whether they actually work or not — are examples of obstacles or shortcomings one might have to face when applying deep learning to real-life data.

The macula is the 'center of vision': the fovea, located in its center, is the part of the retina with the highest concentration of cones, which are the cells responsible for high-resolution, color vision in good light. As such, lesions in the macular region can quickly impair central vision, and it is crucial for a diagnosis to be made on a retinal image — be it by a practician or by an automatic diagnosis algorithm — that the quality of the image is sufficient in this region. Unfortunately, in telemedicine networks, it is not uncommon (around 10% of all examinations) that image quality is insufficient for a diagnosis to be made. The RetinOptic project aimed at devising a new, portable retinograph, which may enable a wider screening of the population, but with the downside that portable devices typically provide images of lower quality than tabletop ones, which was the main motivation of our image quality estimator, which is detailed in Chapter 7.

Outline of this thesis

This document is split in two parts, regarding the ATHENA and RetinOptic projects, respectively. Chapter 1 introduces active thermography, the particular case of the ATHENA project and presents the problem addressed in this part. Chapter 2 details the solution we developed, based on the analysis of certain extrema of interest, and provides segmentation examples.

Chapter 3 is the contribution of this thesis to mathematical morphology: we expand the analysis of extrema based on extinction values, and define several decompositions as well as several new morphological operators, and study their properties.

Part II concerns the RetinOptic project: in Chapter 4, we introduce the context of the project, and motivate our goal of detecting the macula. Chapter 5 presents convolutional neural network classifiers for discriminating between images where the macula is visible and images where it is not. Chapter 6 focuses on the localization of the macula by regression neural networks, assuming we already dispose of a good classifier. We investigate the possibility of feeding the networks a different input, consisting in a morphological decomposition designed so as to make the task easier. We also propose several solutions to overcome an issue caused by the structure of our dataset ground truth positions: at first, the network seem to correctly predict only one of the two coordinates of the fovea. Finally, in Chapter 7, we present another way of approaching the problem, using a fully-convolution segmentation network, which is able to act as both a classifier and a regression model, after a simple post-processing.

Publications

Active Thermography

Robin Alais, Petr Dokládal, Etienne Decencière and Bruno Figliuzzi. "Automatic Detection of Cracks and Delaminations in Thermal Images". *IXth Workshop "NDT in Progress"*, 2017.

Morphological Decompositions

Robin Alais, Petr Dokládal, Etienne Decencière and Bruno Figliuzzi. "Function Decomposition in Main and Lesser Peaks". *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, 2017.

Retinal Image Understanding

Robin Alais, Petr Dokládal, Ali Erginay, Bruno Figliuzzi and Etienne Decencière. "Fast Macula Detection and Application to Retinal Image Quality Assessment", *Biomedical Signal Processing and Control*, 2020.

As co-author:

Geology

Arezki Chabani, Caroline Mehl, Isabelle Cojan, Robin Alais and Dominique Bruel."Semi-automated component identification of a complex fracture network using a mixture of von Mises distributions: Application to the Ardeche margin (South-East France)", *Computers & Geosciences*, 2020.

Geostatistics

Laure Pizzella, Robin Alais, Simon Lopez, Xavier Freulon and Jacques Rivoirard. "Taking better advantage of fold axis data to characterize anisotropy of complex folded structures in the implicit modeling framework", *Mathematical Geosciences*, 2021.

Part I

ATHENA

Chapter 1

Problem Presentation

Ce chapitre présente les bases de la thermographie active ainsi que le projet ATHENA, sur lequel porte la première partie de ce manuscrit.

Après avoir rappelé le principe de la thermographie active et introduit le contexte du projet et ses objectifs, nous détaillons la méthode dite de flying-spot utilisée pour acquérir les thermogrammes, et présentons les différents défauts à détecter, ainsi que leur signature à une dimension.

1.1 Active Thermography Principle

In recent years, active thermography for condition monitoring has gained interest as an alternative to other non-destructive techniques such as liquid penetrant or magnetic particle testing. It has the advantage of being contactless and can be performed in situ more easily. As compared to dye penetrant testing, it is able to detect not only surface-breaking but also underlying cracks, provided that they are close enough to the surface to act as a thermal barrier, and it can also be used to detect delaminations of coated materials [She97].

Active thermography control consists in heating a material and monitoring its surface temperature. The energy source can be of several types: photonic, for instance a laser, induction heating, if the material is an electric conductor [Leh+92], or hot air jet heating [Leh+94; Har+94]. Surface temperature can then be measured by infrared radiometry.

In addition to the various source types, two excitation configurations can be distinguished: pulse thermography and lock-in thermography. In pulse thermography, heat diffusion is monitored near the excitation point in order to find potential disturbances in the heat flow: the 'flying-spot' method which is studied in this work is a particular case of this approach. Lock-in thermography, on the other hand, performs a spectral analysis of a sample submitted to a periodic thermal stimulation [Str+13b], in order to study the properties of the generated thermal "waves". An advantage of this approach is that the response can be averaged over several periods in order to improve signal/noise ratio. First results indicate that it can be used to estimate crack depth [Str+13a]. One of the main downsides of lock-in thermography, however, is that it may require the response to be averaged over many periods in order to obtain a sufficient signal/noise ratio, and the inspected zone cannot be larger than the infrared camera's receptive field. This is not the case for flying-spot thermography, since the camera and heat source move along the inspected specimen.

1.2 ATHENA Project

ATHENA — for Active THErmography for Nondestructive inspection Automation — is a collaborative project between academic (MINES ParisTech and Université de Bourgogne) and industrial

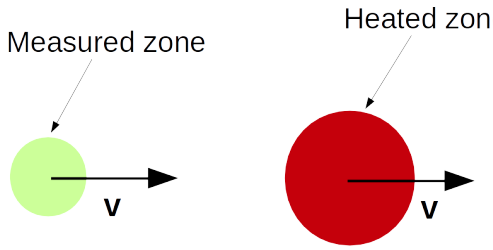


Figure 1.1: Illustration of a flying-spot inspection, in its simplest form: heat is deposited on the sample by an active source and infrared radiation is measured at another zone, both moving at the same speed **v** along the inspected sample. In this figure, the measurement is made behind the heated zone, but this is not necessarily the case.

actors (ARDPI, Intercontrôle, Ascot, Aubert & Duval, EDF), in relation with the competitive cluster "Nuclear Valley", in the form of a French "Fonds Unique Interministériel" grant.

The project aims at providing a complete solution to perform automated nondestructive inspection by active thermography; an active photothermal camera was developed, along with the tools to control a robotic arm upon which it can be mounted; a theoretical analysis of the system was the object of a PhD thesis at the Université de Bourgogne [Thi17], and an automated defect segmentation software was developed by Armines / MINES ParisTech, and has been the object of an international conference communication [Ala+17a].

This part of the manuscript is dedicated to this automated defect segmentation method. After a brief review of the 'flying-spot' thermal imaging technique, the requirements of our algorithm are stated. Chapter 2 details the algorithm itself and formalizes certain notions (defect symmetry, signal amplitude, noise level) that were so far ill-defined.

1.3 Flying-spot Camera

A particular active thermography technique consists in heating a localized mobile zone of the material and observing another localized mobile zone with an infrared sensor, the movements of the source and sensor being synchronized (see Fig.1.1). This idea was originally introduced in the late 1960's [Kub68] in order to detect cracks in aeronautic structures. In this precursor work, the heat source was a xenon arc lamp and the obtained resolution was quite poor; the authors at this point already suggested using a focalized laser beam as a heat source instead.

In the late 1980's, a first theoretical analysis of crack detection by flying-spot camera was performed in [Kau+87]. The authors mention that, from a radiative viewpoint, cracks act as black bodies. They also point out the problems caused by variations in surface absorptance and emissivity, which result in what they call "surface noise". Having observed that artifacts due to this surface noise can produce larger signals than those induced by cracks, they used two infrared detectors instead of one, measuring adjacent regions and used differentiation between the two in order to filter out this surface noise, making cracks easier to detect.

In a later work [Wan+90], it was proposed to use moving mirrors in order to control the motion of the source and detection beams; the laser and detector are motionless and aimed at a mirror whose rotation controls the motion of the source and detection zones upon the inspected sample. From a theoretical standpoint, in this same work, the authors claim that if there is no offset between the two zones (*i.e.* the source and detection beam are focused on the same zone), the resulting image only describes optic variations of the surface; for the image to contain 'real' thermal information, the detection zone must lag behind the heat source. Subsequent works [GB92; GLB93] further investigated this method, using larger mirrors, and found that the signature of cracks is bipolar, with a peak

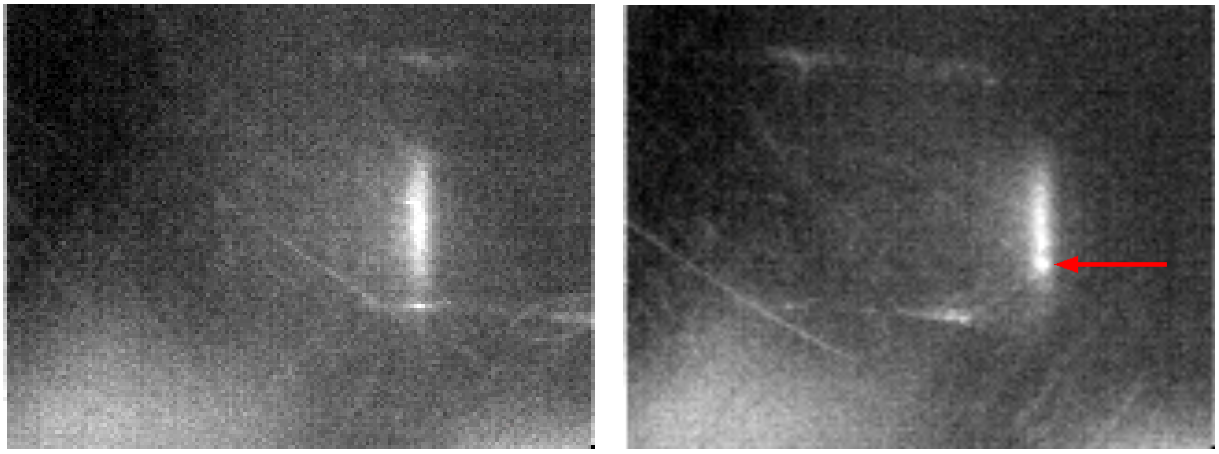
preceding a valley, the amplitude of this signal being maximal when the detector points slightly ahead of the heated zone.

Instead of punctual or quasi-punctual source and detection beams, it is possible to make the heated and inspected zone larger, which also has the advantage of making inspection faster. It was proposed in [Var+92] to focus the source laser on a line and to use a line scan infrared camera. The general procedure remains the same as that of Fig. 1.1, but the two quasi-punctual spots are replaced by lines orthogonal to the system's motion. With this system, instead of a one-dimensional signal, it is now possible to construct an image by concatenating the observations.

On this kind of image, cracks are visible as bright linear structures, with a more or less pronounced adjacent darker zone. However, cracks that are parallel to the inspection system's motion cannot be detected this way, as they offer little or no resistance to the lateral heat beam. To correctly detect those cracks as well, it was proposed in [Var+95] to perform a second scan, orthogonal to the first. By substracting the second image to the first, part of the surface noise can be removed, as details that are independent to scanning direction disappear. However, artifacts due to surface emissivity variation remain.

In order to eliminate these artifacts, it was proposed to perform two scans in opposite directions and subtract the back image to the forth image [Kra+98]. This way, the bipolar signature of defects can be almost doubled, while peaks induced by emissivity and absorptance variations are highly attenuated.

All the images studied in the present work were obtained by this back-and-forth procedure. The heating source used was a laser line, and the detector was an infrared camera whose field of view contains the laser line. Since the detector is not simply a line scan camera but measures a rectangular zone, from a given video inspection, several thermograms can be generated, by considering different detection zones, each detection zone consisting in a line parallel to the laser line.



(a) (b)

Figure 1.2: Sample images from a thermographic video: (a) the laser line is on a sane zone (b) it is upon a punctual crack, visible as a brighter dot at the bottom of the laser line, indicated by the red arrow.

1.4 Visible Defects And Their One-Dimensional Signatures

Throughout the rest of this study, we will consider images obtained by the back-and-forth scanning technique described in the previous section, and we will assume that the scan is made in the horizontal direction, the forward scan being made from left to right, and the back scan from right to left. The resulting image is the subtraction of the back scan to the forth scan.

In the following, we will consider two types of thermal blocking defects: cracks and delaminations.

1.4.1 Cracks

Cracks appear on the images as a light zone immediately followed by a dark one. An example is given in Fig.1.3: the crack can be seen on the left side of the image, between $x \approx 120$ at the bottom and $x \approx 160$ at the top. There are however lots of small structures that match the description "a light zone followed by a dark one". Based on this image alone, it is hard to tell whether they do correspond to small, quasi-punctual cracks or if they are due to variations in absorptance or emissivity, small scratches, or other possible artifacts. If we plot the gray values along a horizontal line, we can see that the crack is, as expected, characterized by a bipolar signal where a high local maximum is followed by a low local minimum. However, there are a lot of oscillations, some of which are of comparable amplitude.

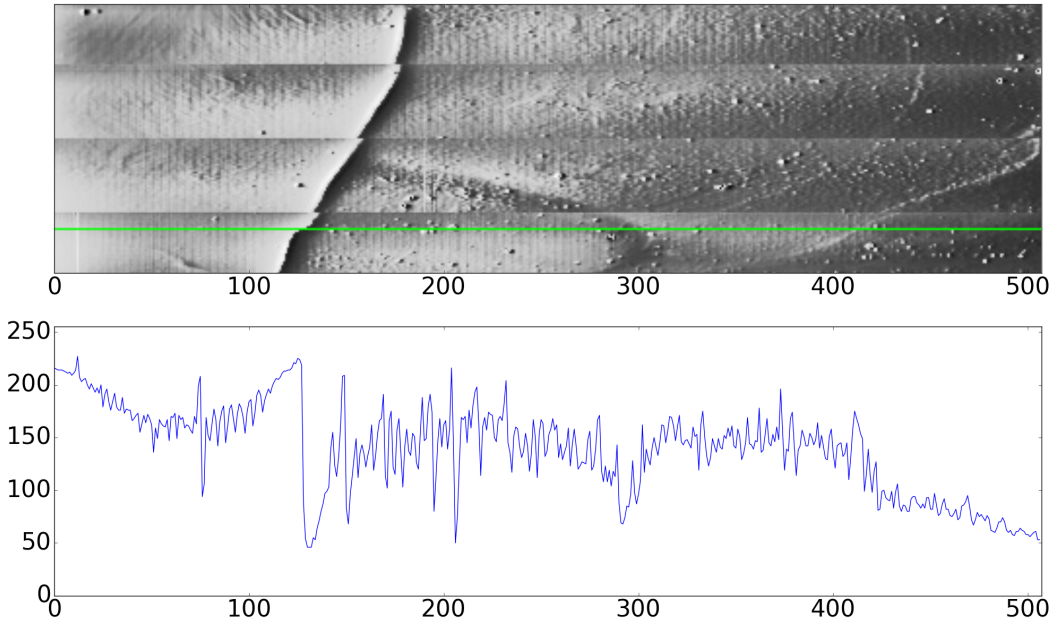


Figure 1.3: Sample with a visible crack and profile across the green line.

Another example can be seen in Fig. 1.4: in this image, there are multiple cracks on the right-most side of the image and corresponding oscillations in the gray values. The large valley visible on the signal plot around $x = 1200$ is not a defect but is due to the geometry of the inspected piece, in this case a Pelton wheel bucket. There is no corresponding large peak to this large valley, which is an information we can make use of when designing our automatic detection algorithm: cracks result in bipolar, roughly symmetrical signals, while this particular large valley does not have a corresponding large peak of similar amplitude.

1.4.2 Delaminations

Delaminations also have a bipolar signature, but in the opposite order: they appear as dark zones followed by light ones, possibly with a flat gray zone in between. The corresponding one-dimensional signal is bipolar, with the minimum preceding the maximum, and the corresponding valley and peak may be separated by an almost flat zone. An example is given in Fig. 1.5.

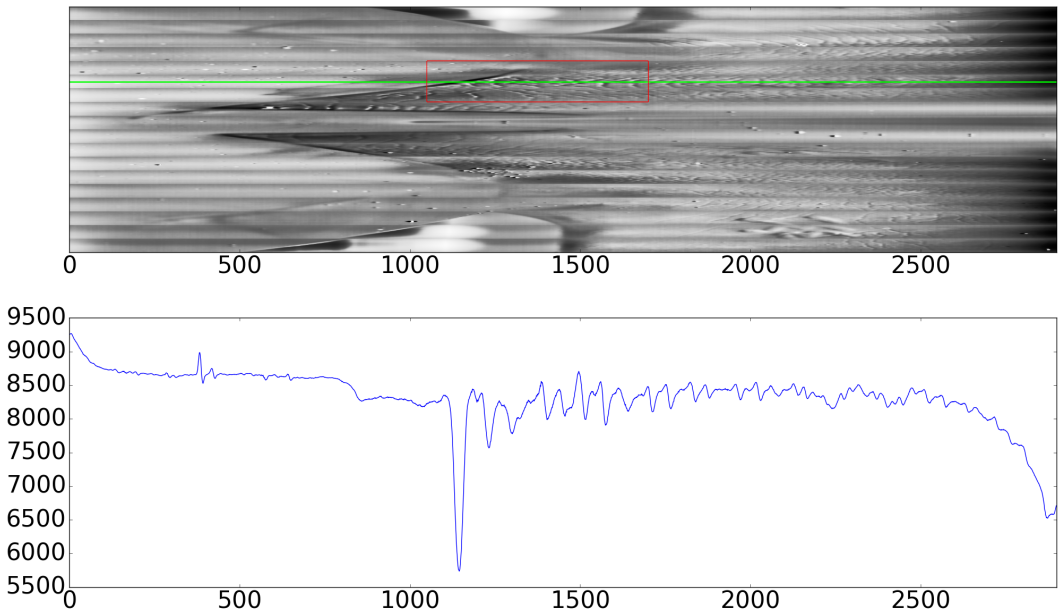


Figure 1.4: Pelton wheel bucket with multiple cracks and profile across the green line.

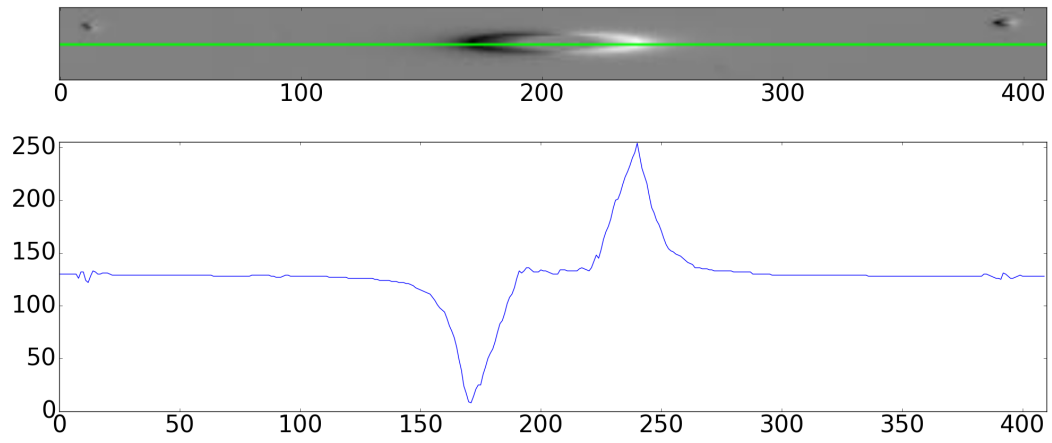


Figure 1.5: Sample with a delamination and profile across the green line. The two small structures at the left and right side of the image are likely smaller delaminations as well.

1.5 Acquisition Artifacts

1.5.1 Low-frequency trend

Because of thermal inertia, the sample may not cool down back to its equilibrium point immediately after the first scan, which can lead to an undesirable low-frequency component: the left part of images is brighter than the right part. If the scan is performed on a whole piece, from one end to the other, there are also accumulation effects on the sides: heat loss by convection at the edges of the sample is slower than heat diffusion throughout the material. This can be seen on Figures 1.3,1.4: the leftmost few pixels are very bright, while the rightmost are very dark. Fig 1.5 is actually a crop of a larger image, but the same phenomenon is visible on the full image. In Fig.1.3, which corresponds to a relatively small sample, the leftmost part of the image has an average gray value around 200, while in the rightmost part, it is around 50.

1.5.2 Acquisition Issues

In addition to physical phenomena, artifacts can occur because of camera-related issues. Figure 1.6 shows part of a thermogram that suffers from various problems: the first highlighted structure is a delamination, but either the subtraction of the backward scan from the forward scan was improperly done, or a problem occurred when saving the resulting image; either way, this example is an 8-bit image where negative values looped back up to 255; where a 'gray' value should be -10, on this image, it is 245, and so on.

On this same image, there are also fast oscillations at the bottom that cannot be explained by physical defects. As it turns out, the most likely explanation in this case is that the detection system was actually pointing outside of the examined sample and at the curtain behind it.

1.5.3 Image file format

As explained in Section 1.3, the thermograms we want to analyze in the framework of the ATHENA project are obtained by subtracting one image to another. The raw result of this operation is an 'image', or an array, containing both nonnegative and negative values. Sign provides an interesting information: a negative value means that the measured temperature was higher during the back scan than during the forth scan; a positive value means the contrary; a value close to zero means that the measured temperature was roughly the same for both scans.

Because of the parasitic low-frequency trend mentioned above, due to thermal inertia effects, this intuition may not perfectly accurate in some cases, notably on the edges of a sample; nevertheless, it could be an interesting baseline to work with. Unfortunately, all thermograms were saved as images, and therefore only contain nonnegative values. The 'neutral' value corresponding to pixels where the measured temperature was the same on both scans is not known, and while it is relatively easy to estimate on some images like Fig. 1.4, it is much harder to guess on images like Fig. 1.3.

The thermographic camera used in this project used a coding on 12-bits, which means that the subtraction of two images should be coded on 13 bits in order not to lose information. This was not taken into account in the former stages of the project: the first images were saved on 8 bits in the JPEG format, which in itself uses lossy compression, so a further layer of information may be lost because of it. Latter images were saved in a 16-bit TIFF format, but the neutral level was not provided.

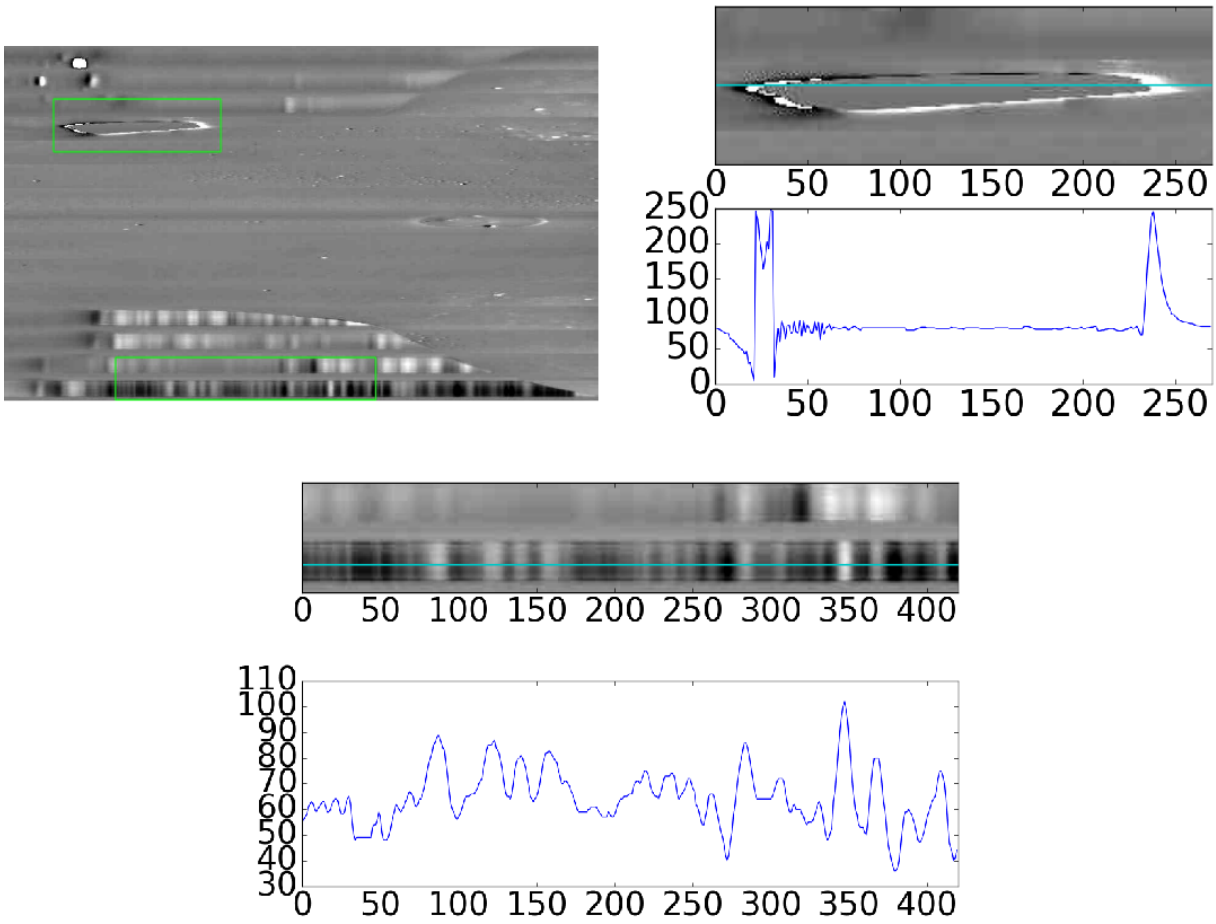


Figure 1.6: Thermogram exhibiting several artifacts: saturation issue on top, and unexplained oscillations at the bottom.

1.6 Manual Characterization of Cracks

When a thermogram is inspected by a human expert, the standard practice for characterizing the defect consists in drawing a line perpendicular to it and considering the profile along this line. The profile is then separated in three parts: the inferior part, the signal and the superior part. The inferior and superior parts are approximated by linear regression; the maximal amplitude of noise A_n is defined as the greatest absolute difference between these linear approximations and the profile on the inferior and superior parts.

The signal part is itself split in two; the linear regression on the inferior part is extended to the leftmost part of the signal, and the linear regression on the superior part is extended to the rightmost part of the signal. This way, two signal amplitude values A_1 and A_2 are defined by taking the greatest absolute difference between profile and linear interpolation on each half of the signal part. Only the greater of these two values is considered when computing the signal/noise ratio, which is defined as:

$$R = \frac{\max(A_1, A_2)}{A_n}.$$

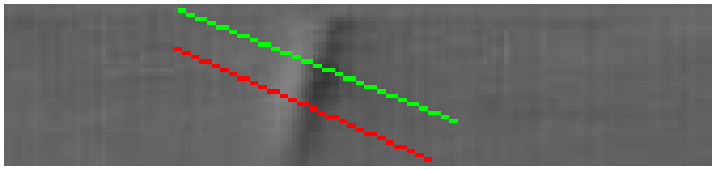
The ratio between A_1 and A_2 defines the symmetry value of the signal:

$$S = \frac{\min(A_1, A_2)}{\max(A_1, A_2)}.$$

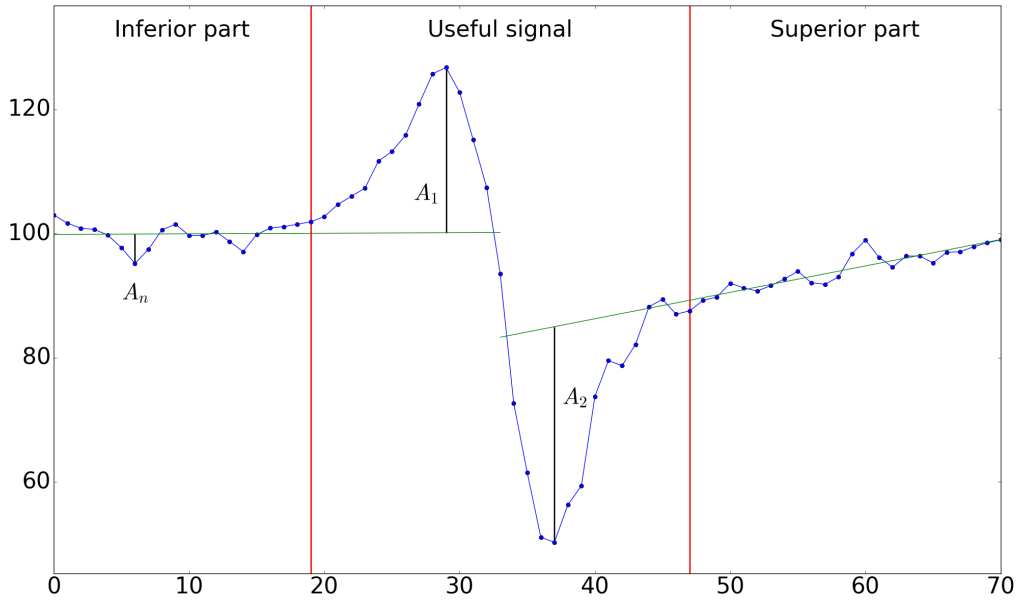
The process is illustrated in Fig. 1.7. The greatest weakness of this approach is that it is heavily dependent on several choices of the operator, each of which is somewhat arbitrary. The line along which the profile is plotted is hand-drawn; the segmentation in inferior part, useful signal and superior part is visually performed, but apart from the fact that the two extrema (the maximum and the following minimum, in the case of cracks) must be in the signal part, there is no hard rule for where exactly to put the borders.

In the example illustrated in Fig. 1.7, two different analyses of the same defect provide very different values of signal/noise ratio (7.3 and 3.2, respectively) and symmetry (0.77 and 0.96).

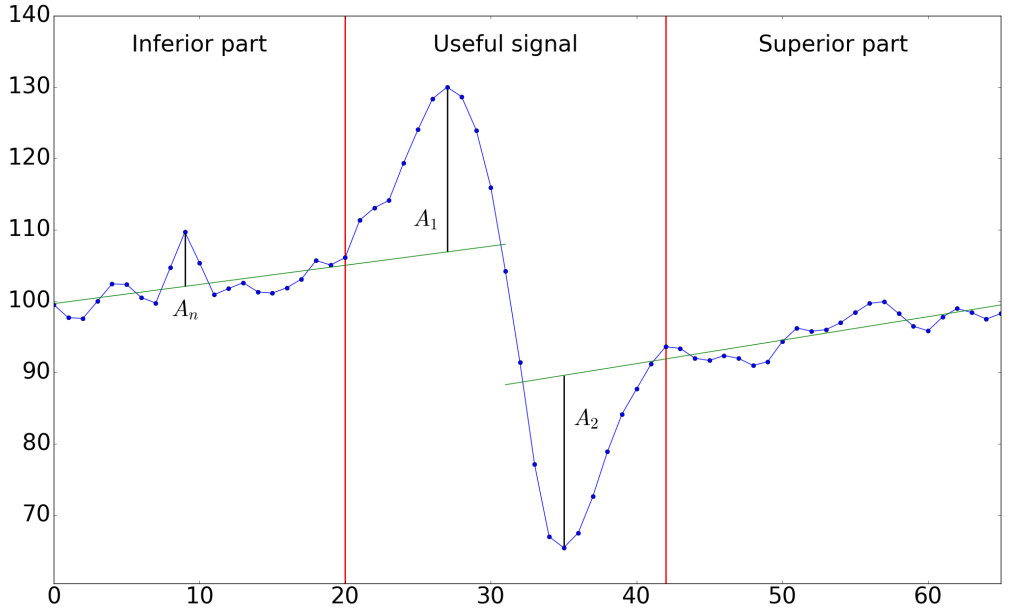
In the next chapter, we provide new definitions for these notions; although our algorithm has some input parameters, they mainly depend on the type of inspected material and on the scanning speed. Once these parameters are given, both detection and characterization are automated.



(a) Visible crack and two possible analysis lines.



(b) Profile analysis of the top (green) line. Here $A_n = 4.7$, $A_1 = 26.6$ and $A_2 = 34.75$. The estimated signal/noise ratio is $R = 7.3$ and the estimated symmetry is $S = 0.77$.



(c) Profile analysis of the bottom (red) line. Here $A_n = 7.6$, $A_1 = 23.1$ and $A_2 = 24.15$. The estimated signal/noise ratio is $R = 3.2$ and the estimated symmetry is $S = 0.96$.

Figure 1.7: Two different manual analyses of the same defect, with very different values in terms of noise level, signal amplitude, SNR and symmetry.

Chapter 2

Automatic Defect Segmentation

Ce chapitre détaille l'algorithme de segmentation utilisé pour détecter les défauts dans les thermogrammes du projet ATHENA.

Après avoir rappelé les contraintes auxquelles l'algorithme doit répondre, nous détaillons et illustrons les différentes étapes. Après un premier filtrage pour éliminer la composante basse fréquence de l'image, un premier masque de détection est obtenu en analysant séparément chaque ligne de l'image, puis ce masque initial est complété. Chaque défaut détecté dans le masque final est également caractérisé par différentes propriétés, fournies dans un rapport de détection. L'influence des quelques paramètres d'entrée de notre méthode est également étudiée.

La première phase de détection des défauts, ainsi que l'estimation du niveau de bruit, se base sur l'étude de certains extrema d'intérêt dans les signaux 1D obtenus après pré-traitement. Cette caractérisation des extrema reste élémentaire dans ce chapitre, mais elle a motivé une étude mathématique plus approfondie qui fait l'objet du chapitre suivant.

2.1 Requirements

As mentioned in the previous section, the method should be able, not only to accurately detect defects, but also to characterize them. In particular, for a given detected defect, it should provide measures of height, width (in pixels), area, perimeter, as well as measurements of signal/noise ratio and symmetry. These values are then provided in a detection report, so that it can be used by the operator in an interactive post-processing step, for instance in order to filter out small defects by thresholding on area or height, to discriminate between defects and artifacts by thresholding above a minimal symmetry, or to keep only significant enough defects by imposing a minimal signal/noise ratio. As mentioned in the previous chapter, these latter notions of signal/noise ratio and defect symmetry are ill-defined; in this chapter, we introduce a measure of global noise level on a thermogram, and a formal definition of symmetry.

2.2 General strategy

There are two types of defects to be detected: cracks and delaminations. The corresponding algorithms are very similar: in a first step, the image is filtered in order to remove the low-frequency artifact; a global noise level is then estimated on this filtered image. A first detection is then performed in one dimension, line by line, by looking for potential signatures of defects. In a postprocessing step, partial detections are combined, if needed, in order to produce the final detection mask, along with the detection report: to each connected component of the detection mask are associated various attributes, such as mean and median signal/noise ratio, mean and median symmetry, area or height (in pixels).

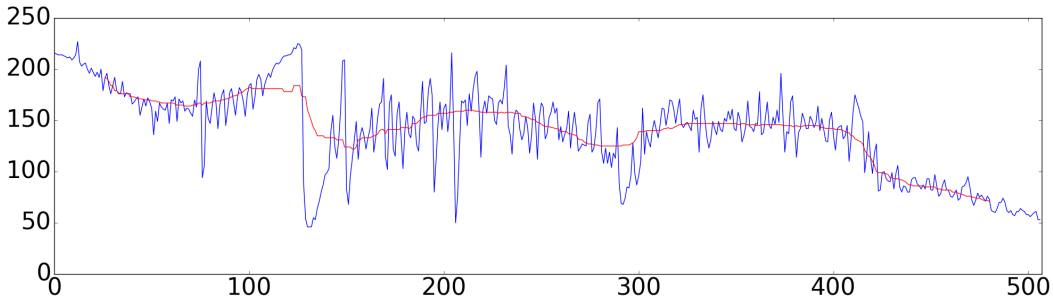


Figure 2.1: Profile along a thermogram line (in blue) and smoothed profile (red).

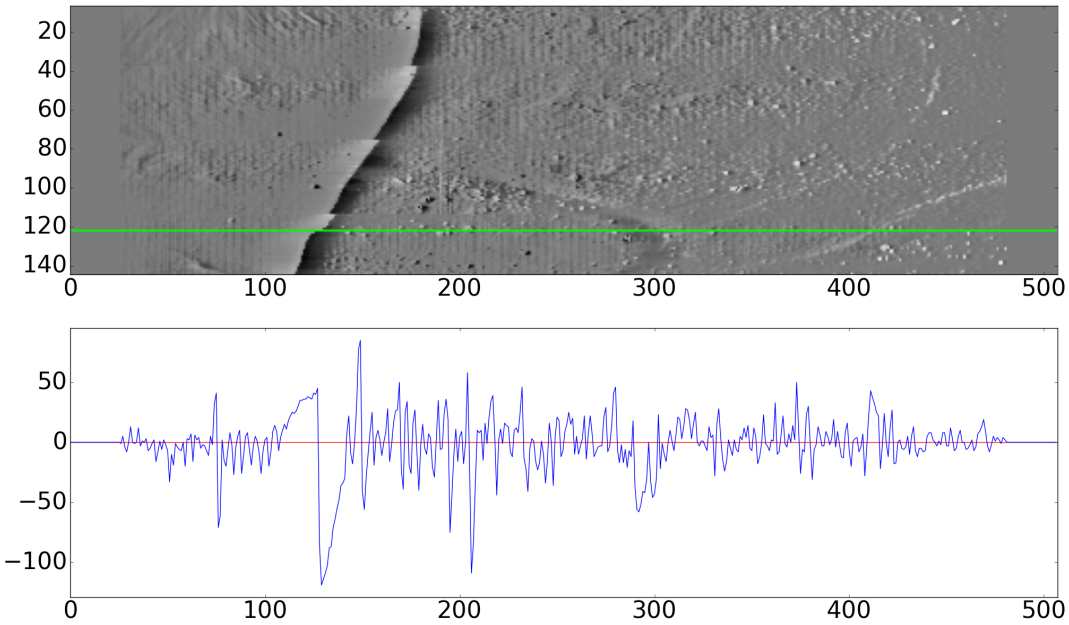


Figure 2.2: Filtered image and corresponding profile along the green line.

In an additional step, it is possible for the operator to interactively adjust the detection mask by specifying minimum or maximum values of these attributes.

2.3 Preprocessing

As explained in the previous chapter, both because of physical effects and because the thermograms were saved as images, it is not immediately obvious, for a given pixel, whether there was more energy deposited at this location during the forward or during the back scan. The aim of the preprocessing step is to recover this information by subtracting to each line of the image a smoothed version of the line profile. Figure 2.1 illustrates this on the same profile shown in Fig. 1.3, and Fig. 2.2 shows the filtered image and the line profile after filtering.

Given the scanning velocity and temporal resolution of a given acquisition, it is possible to estimate the expected peak-to-peak distance of cracks, in pixels. This value, denoted L in the following, is an input parameter of the method: we use a median filter of size $4L + 1$ as our estimation of the low-frequency component to be removed, and subtract it to the raw profile. The median filter is ill-defined on the first and last $2L$ points, which are discarded. Since the scans are usually performed on a whole piece, these regions of the image are unusable anyway because of thermal effects on the edges.

For delamination detection, we use a median filter as well, but its size is given by the maximum expected size of a delamination in the image; that is, the algorithm requires a parameter L_D, which is used as the size of the median filter. In both cases, the detrended profile now has both negative and nonnegative values, that we can interpret as regions where the sample was hotter during the back scan, and regions where it was hotter during the forth scan, respectively.

Subtracting the median filter might significantly alter the signal if the parameter L or L_D is underestimated, whereas overestimating it a little does not affect the outcome much; rather than a precise estimation of L or L_D, it is preferrable to provide a reasonable upper bound. A more thorough discussion on parameter influence is provided below in Section 2.10.

2.4 Relevant extrema

Our defect detection algorithm is based on the study of the extrema of the detrended profile. For crack detection, we will look for a large positive maximum, immediately followed by a large negative minimum. For delamination detection, we will look for a large negative minimum followed by a large positive maximum, possibly with some smaller extrema in between. In addition, our definition of the noise level of an image is also based on the study of the distribution of the values of local extrema.

Not all local extrema are useful for our task: typically, positive local minima or negative local maxima are of no interest. In addition, if there are several positive local maxima, or several negative local minima between two zero-crossings, we choose to keep only the greatest (in absolute value), as can be seen in Fig. 2.3

More formally, given a detrended profile f (obtained by substracting a median filter to the original profile), we consider all zero-crossings of this profile f. Between two zero-crossings, the sign of f does not change; if f is positive, we keep the position and value of the global maximum of f on this interval; if f is negative, we keep the position and value of the minimum. In the following, these will be referred to as the *relevant extrema* of f. In cases of equality between several maxima, we keep only the leftmost one; in cases of equality between several minima, we keep the rightmost one. Ties are broken this way because in the case of delamination detection, we will look for a minimum followed by a maximum with at most L_D pixels between the two; picking the rightmost minimum and leftmost maximum reduces the risk of false negatives.

2.5 Global Noise Estimation

Let X be the set of all values of relevant extrema on all (detrended) horizontal profiles of a given thermal image. X typically contains many small values corresponding to noise, and may contain some larger values corresponding to defects or to optical artifacts.

Our idea is to use as our global noise estimator the standard deviation of the noise's amplitude. If the set X contains too many or too large 'signal' (due to defects) or artifact values, the standard deviation of X is not a good approximation of the standard deviation of the noise alone. In order to obtain a sensible noise estimation, we first get rid of possible outliers in X, using iterative 3σ thresholding, as explained below.

The three sigma rule

If $X = \{x_1, \ldots, x_N\}$ is a finite set of reals, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ its mean and $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ its standard deviation, most of the values in X lie in the range $[\mu - 3\sigma; \mu + 3\sigma]$; this is often informally called the three-sigma rule of thumb and is used in various contexts in order to detect outliers of a (possibly unknown) distribution.

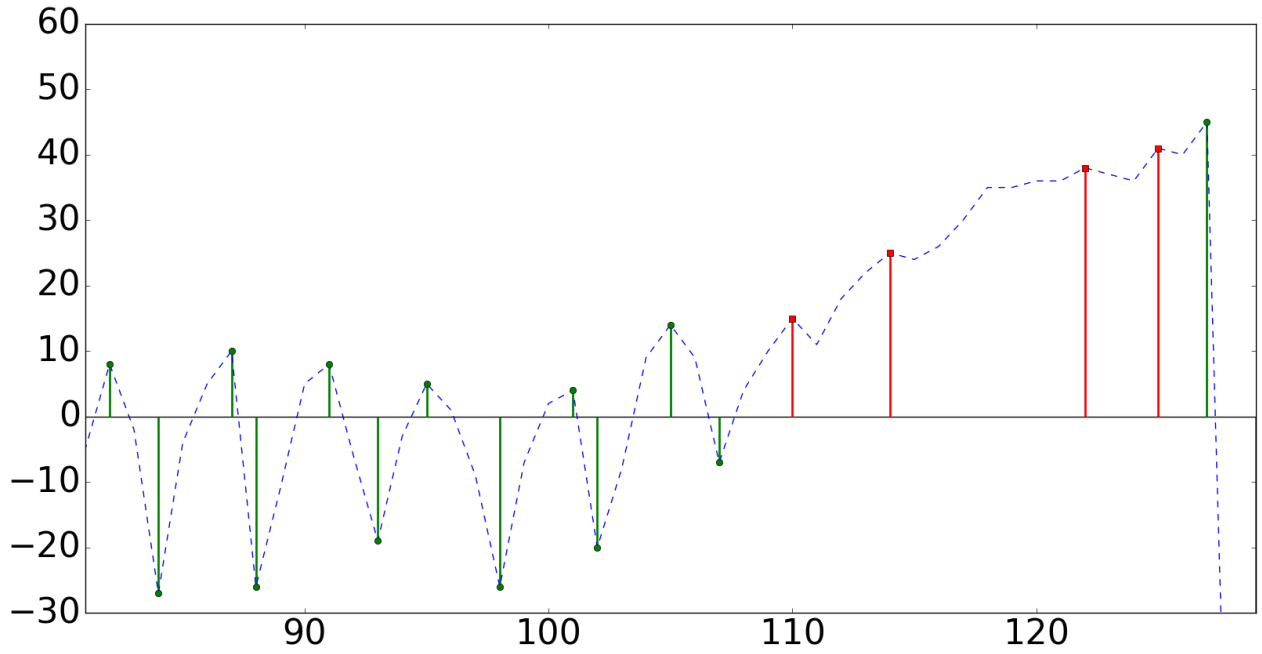


Figure 2.3: The relevant extrema of the detrended signal are marked with green dots and vertical green bars. The maxima indicated by the red squares and bars are discarded because there is a greater maximum in the same interval between two zero-crossings.

If the x_i's are drawn from a normal distribution with mean μ and standard deviation σ, the probability of falling in the interval $[\mu-3\sigma;\mu+3\sigma]$ is about 0.9973. In the more general case where the x_i's are drawn from a unimodal distribution with finite variance, the Vysochanskij–Petunin inequality [Pet80] states that this probability is over 0.95. Finally, for any random variable Y with finite mean μ and finite non-zero variance σ^2 are defined, Chebyshev's inequality [Che67] provides:

$$P(|Y-\mu| \geq \alpha\sigma) \leq \frac{1}{\alpha^2}$$

so in particular, for $\alpha = 3$, this inequality states that the probability of falling outside the range $[\mu - 3\sigma; \mu + 3\sigma]$ is less than one ninth.

The proof for Chebyshev's inequality is very simple and can be adapted in terms of number of outliers rather than probabilities: given $X = \{x_1, \ldots, x_N\}$ a finite set of reals, μ its mean and σ its standard deviation, no more than $\frac{1}{9}$ of the values are outside the range $[\mu - 3\sigma; \mu + 3\sigma]$.

We can easily prove this by contradiction: let us assume that strictly more than $\frac{1}{9}$ of the values in X are outside this range. We can write:

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \geq \frac{1}{N}\sum_{i;|x_i-\mu|>3\sigma}(x_i-\mu)^2 > \frac{1}{N}\sum_{i;|x_i-\mu|>3\sigma}(3\sigma)^2$$

and by hypothesis, strictly more than $\frac{1}{9}$ of the values x_i in X are such that $|x_i-\mu| > 3\sigma$, therefore:

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 > \frac{1}{N}\frac{N}{9}(3\sigma)^2 = \sigma^2$$

.

where by definition $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$: we reach a contradiction.

Iterative 3σ thresholding

Given the set X of relevant extrema, we compute its standard deviation σ and keep only the elements in X whose distance to the mean is less than 3σ, and iterate this process until convergence. Note that this procedure always converges to a non-empty set: the number of elements in the set decreases at each step but cannot reach zero, as this would contradict the property above.

After convergence is reached, σ is a reasonable estimation of the amplitude of positive noise, while $-\sigma$ is a reasonable estimation of the amplitude of negative noise: we define 2σ as our global noise estimation.

2.6 One-Dimensional Detection

2.6.1 Cracks

In a first step, we consider as a potential crack signature every 'large enough' couple consisting in a relevant maximum followed by a relevant minimum (with no other relevant extremum in between). In addition to the parameter L, our algorithm takes as input a parameter R, which is the minimum signal/noise ratio of cracks to be detected. An optional parameter L_{max}, which is the largest possible peak-to-peak distance for cracks, can be provided by the user; if not, it is set to $L_{max} = 2L$. For each detrended horizontal profile, if a relevant maximum at position x_1 with value $M > 0$ is immediately followed by a relevant minimum at position x_2 with value $m < 0$, with $M - m \geq 2\sigma R$ and $x_2 - x_1 \leq L_{max}$, we add the segment between x_1 and x_2 to our crack detection mask. It is also possible to specify a minimum symmetry value S_{min} (see below) as input, in which case the segment is added to the detection mask only if the couple (M, m) satisfies this symmetry condition.

2.6.2 Delaminations

Delaminations appear as a relevant minimum followed by a relevant maximum, but unlike cracks, it is possible that there are other relevant extrema in between (this would, for instance, be true of the profile shown in Fig. 1.5). For every line, for every couple $\{(x_1, m); (x_2, M)\}$ of relevant minimum/maximum with $M - m \geq 2\sigma R$ and x_2–$x_1 \leq L_D$, we add the segment between x_1 and x_2 to the delamination detection mask.

Unlike cracks, there might be other relevant extrema between x_1 and x_2. In particular, theoretically, there could be a relevant minimum x_1' between x_1 and x_2 and a relevant maximum at $x_2' > x_2$; if, in addition, this couple $(x_1'; x_2')$ meets the signal/noise ratio criterion, the whole segment between x_1 and x_2' ends up being added to the detection mask, although it is possible that $x_2' - x_1 \geq L_D$. However, even though this is not forbidden from a mathematical point of view, this is unlikely to happen if the parameter R is not too close to 1. In order to prevent this kind of phenomenon, it is also possible to specify a minimal symmetry value.

2.7 Defect Symmetry

For cracks as well as for delaminations, a defect is characterized by a minimum $m < 0$ and a maximum $M > 0$. In both cases, we define the symmetry of the defect as the ratio between the smaller and the greater (in absolute value); formally:

$$S = \frac{\min(-m, M)}{\max(-m, M)}$$

In order to better discriminate real defects from potential artifacts, a minimal symmetry criterion (between 0 and 1) can be specified as input to either detection algorithm. When considering a

candidate couple, this criterion is tested in addition to the signal/noise ratio requirement, and the candidate kept only if it meets both criteria. It is also possible to use mean or median symmetry of suspected defects as a postprocessing refining tool.

2.8 Detection Mask Completion

On noisy images, 1D analysis is not always sufficient to recover a crack in its entirety, as some profiles may exhibit an insufficient SNR, although the defect as a whole is clearly visible, as can be seen on Fig. 2.4.

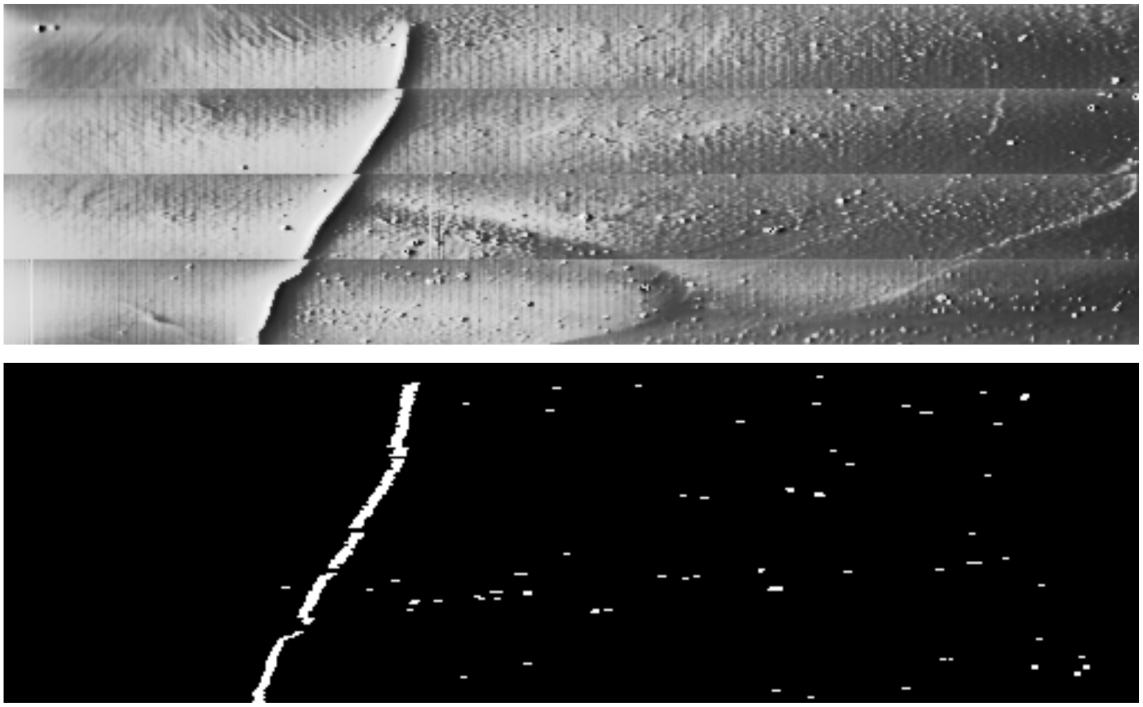


Figure 2.4: First detection mask with $R = 3.4$; the crack is only partially recovered and has several connected components.

To overcome this problem, we relax the condition $M-m \geq 2\sigma R$ for candidate maximum/minimum couples next to annotated cracks if their amplitude is similar enough to neighboring annotated couples: more precisely, if a crack of amplitude $M - m = A$ is annotated in line i between positions x_1 and x_2, we consider the above and below lines $i - 1$ and $i + 1$ between columns $x_1 - a$ and $x_2 + a$, where a is a small parameter allowing for a horizontal shift in the signal we are looking for (typically $a = 1$ or 2 pixels). If line $i - 1$ (respectively line $i + 1$) contains a maximum M' at position x'_1 immediately followed by a minimum m' at position x'_2, and M'–$m' \geq \lambda A$, we add line $i-1$ (resp. $i+1$) between x'_1 and x'_2 to our crack detection mask. The parameter λ is a tolerance parameter controlling how similar candidate defects should be to their neighbors in order to be added to the detection mask. In all our tests, we used $\lambda = 0.9$, with satisfying results. This process is iterated until convergence, which can lead to some candidate couples being annotated even though their amplitude is below the threshold $2\sigma R$. To prevent false detections and erroneous reconstructions between close but separated defects, we add the constraint that for each connected component in the obtained mask, the median SNR (the median value of all line SNRs) must be at least R, which is the minimum SNR demanded by the user. If a minimum symmetry is specified, in the same way, we can ensure that the median symmetry is at least the one specified by the user.

The final crack detection mask obtained, after completion, with the same parameters, can be seen in Fig. 2.5; the crack is now correctly detected as one connected component. The same strategy could be used for delaminations, but in practical cases, it has been found unnecessary.

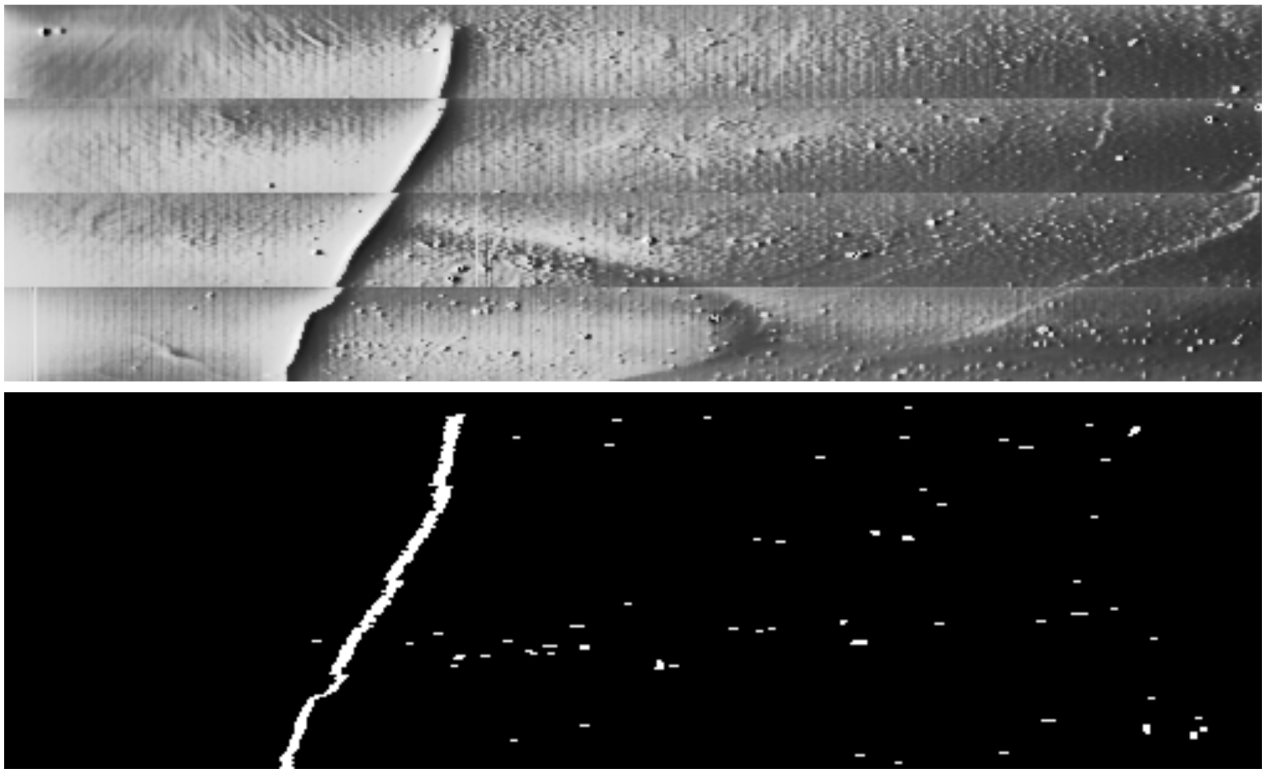


Figure 2.5: Final crack detection mask

An example of delamination detection can be seen in Fig. 2.6.

Figure 2.6: Delamination detection mask

2.9 Post-processing

In addition to the detection mask, to each defect are assigned various features, such as mean/median SNR, mean/median symmetry, perimeter, area, height, width (in pixels). Each of these features can be used in order to further discriminate between defects and artifacts, or simply between defects of interest to the user and defects small enough to be considered harmless. In more realistic cases, symmetry has been found to be an interesting criterion in order to discriminate between cracks and artifacts due to the sample's geometry. In Fig. 2.5, it is obvious that by retaining only large enough defects (in terms of area), only the crack remains. This kind of user-driven post-processing is all the more interesting as it can be done in real time: the algorithms described above in order to get a crack or delamination detection mask can take up to 2 or 3 minutes for 1000x1000 pixel images, but the post-processing can be done interactively. Screenshots of a simple demonstrator for the thermogram shown in Fig. 2.7 can be seen in Fig. 2.8.

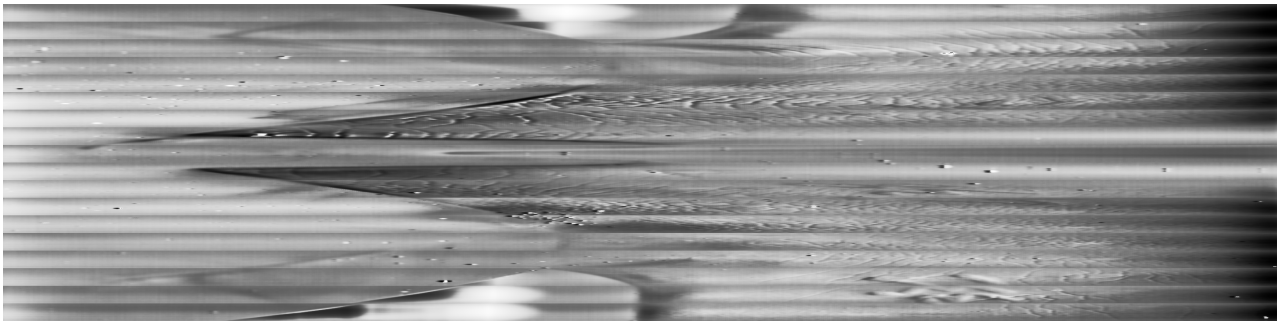


Figure 2.7: Example thermogram with many cracks as well as optic artifacts on top and bottom.

2.10 Parameter Influence

Our crack detection algorithm takes as input two mandatory parameters, L and R, which are the expected size (peak-to-peak distance) of cracks, and the minimal signal/noise ratio, as well as two optional parameters L_{max} and S_{min}, which are the maximum peak-to-peak distance and the minimal symmetry for 1D detection.

Parameter L is used in order to estimate the size of the median filter used in pre-processing; as such, underestimating this parameter can degrade the signal, by making the low-frequency estimation too close to the original profile. Overestimating it, on the other hand, makes the low-frequency estimation somewhat 'smoother', which could in principle cause problems as well, especially if there are geometry issues to be taken care of (for instance around column 300 in Fig. 2.2). In practice, however, it is preferrable to slightly overestimate L than to underestimate it, and the precision of this parameter is not crucial.

Parameter L_{max}, which is set by default to $L_{max} = 2L$, is mainly there to avoid false crack detections, by preventing the pairing of a very distant maximum/minimum couple, for instance if there are two consecutive delaminations on a very smooth image, like in Fig. 2.6. It was made available as an input parameter at the request of the end users, in the eventuality of a specific case where $2L$ would be too loose an upper bound, but there were no examples in our dataset where setting it to $2L$ caused problems.

Parameter R is the minimal signa/noise ratio of cracks to be detected, so it is basically a sensitivity parameter: the smaller this value, the higher the number of detections. It is possible to raise this value *a posteriori* in the postprocessing step, as illustrated in Fig. 2.8; however, because of the reconstruction step described in Section 2.8, the initial value should not be set too low: if R is too small, this reconstruction step may incorrectly connect different zones with a high signal/noise ratio into a single zone, the SNR of which will be much smaller. Thus, applying the algorithm with an initial value R_1 then thresholding the median signal/noise ratio on a value $R_2 > R_1$ is not equivalent to applying the algorithm with initial value R_2.

Parameter S_{min} can be given as input to specify a minimal symmetry value; it can be useful in order to discriminate between defects and artifacts, but even real defects can have quite low symmetries; for instance, on the profile shown in Fig. 2.2, the symmetry for the crack is only $\frac{45}{119} \approx 0.38$.

The delamination detection algorithm takes as input the parameters L_D, R and the optional S_{min}; except for the fact that L_D plays a role analogous to both L and L_{max}, their influence is very similar to the crack detection case; however, whereas S_{min} should be used with caution for cracks, it can arguably be used more freely here: empirically, the symmetry seems to be quite high for delaminations, but it is unclear if this due to a physical property of laminated materials, or if it is specific to our dataset.

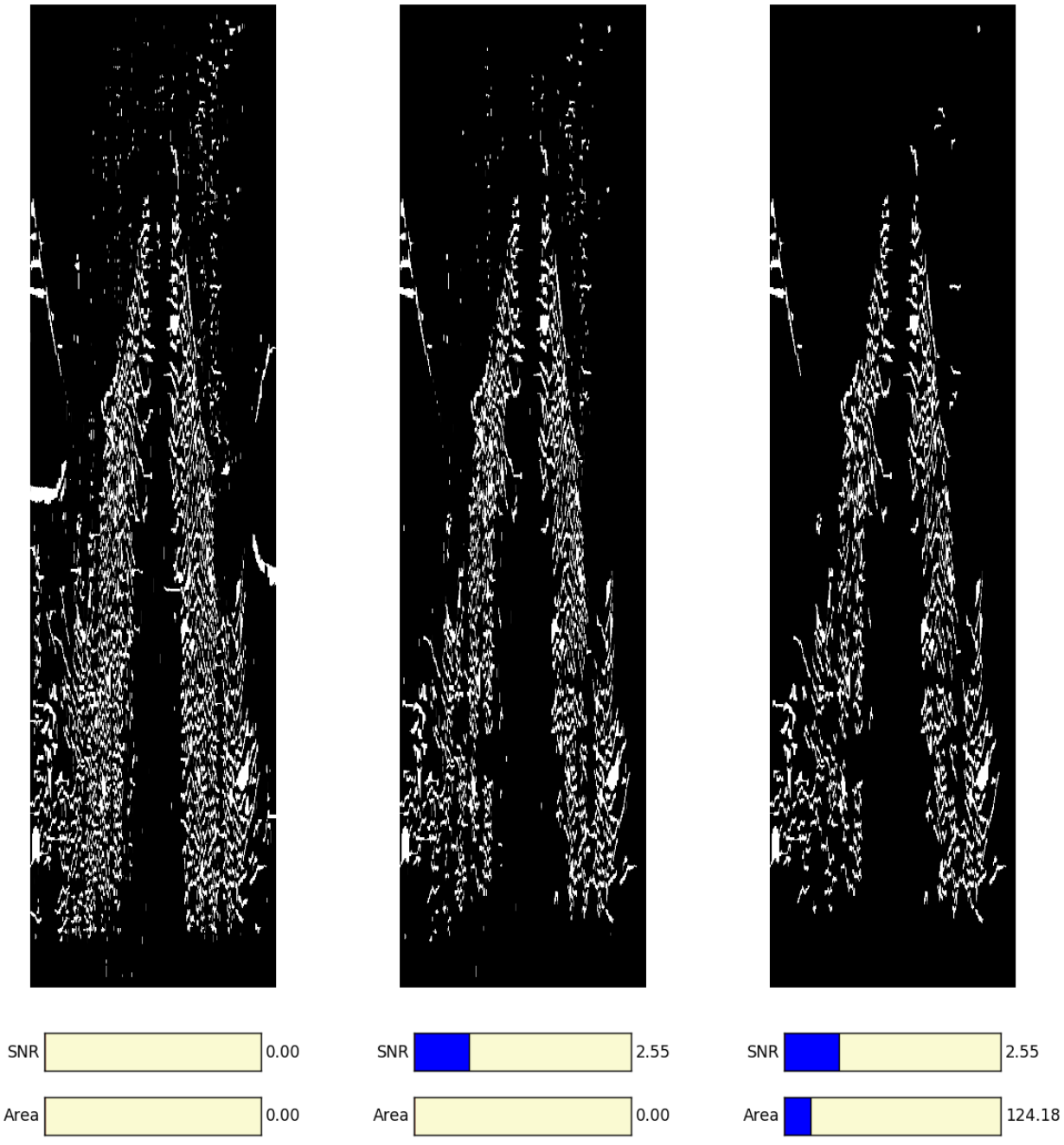


Figure 2.8: Three screenshots of a simplified interactive postprocessing software, illustrating a possible postprocessing on the crack detection mask of the thermogram shown in Fig. 2.7. In this particular case, the images are shown with the forward scan from top to bottom instead of from left to right.

2.11 Conclusion and Perspectives

In this first part, we presented an automatic algorithm for detecting cracks and delaminations on thermal images. Our algorithm needed to be as simple, and with as few input parameters as possible, for it to be used by operators with possibly little background in image processing. The input parameters are for the most part easily deducible of the experimental settings (source velocity, camera resolution). The optional parameters L_{max} and S_{min} may be used as fine-tuning parameters if the first detection mask looks unsatisfactory. We also provided definitions for signal symmetry and signal/noise ratio (by defining the noise level of an image). These properties were so far ill-defined but widely used in this domain. The interactive post-processing we propose is therefore quite intuitive for operators. From a mathematical standpoint, we introduced the concept of *relevant extrema*: for the thermal signals of this part, we simply defined those as the highest (positive) maximum or lowest (negative) minimum on each connected component of the support of a function. However, trying to quantify the importance of extrema led us to explore in more detail the various possibilities for assigning them some kind of importance metrics. A classic approach is to apply more and more severe operators until an extremum disappears, and assigning an *extinction value* related to a family of operators [VM95; Vac95]. By pursuing this idea, we discovered novel morphological decompositions as well as new kinds of extinction values. This is the topic of the next chapter.

Chapter 3

Generalized Extinction Values

Ce chapitre purement théorique peut être lu indépendamment du reste du manuscrit.

Les valeurs d'extinction sont des propriétés quantitatives des maxima ou des minima, utilisées par exemple pour sélectionner les minima les plus importants comme marqueurs d'un algorithme de partage des eaux. Dans ce chapitre, nous montrons comment une valeur d'extinction peut être utilisée pour définir la décomposition d'une fonction comme somme de composantes élementaires, ou pics*, associées à ses maxima. Ces décompositions et leurs propriétés sont illustrées ; en particulier, de nouvelles valeurs d'extinction sont définies.*

Nous définissons également de nouveaux opérateurs morphologiques, en conservant uniquement les pics les plus "grands" (selon un certain critère) d'une décomposition donnée, chaque choix du critère et de la décomposition produisant une famille d'opérateurs différente.

Les opérateurs définis ici étant connectés — il ne créent pas de nouveaux contours — ils sont particulièrement adaptés aux algorithmes de segmentation.

3.1 Mountaineering Analogy

The International Mountaineering and Climbing Federation recognizes fourteen *eight-thousanders*: mountains that are more than 8,000 metres in height above sea level. Mount Everest is the highest, with an elevation of 8,848 metres, and K2 is generally considered the second-highest, with an elevation of 8,611 metres. However, if we consider *summits* — a place or plateau from which one can only go down — the second-highest summit on Earth is not K2, but the South Summit of Mount Everest (Fig. 3.1), with an elevation of 8,749 metres.

The South Summit is considered as a sub-peak of Mount Everest, failing at earning its title of proper mountain, whereas Lhotse (Fig. 3.2), which is also part of the Everest massif, despite a lower elevation of only 8,516 metres, is considered a different mountain — the fourth-highest one on Earth.

There is no qualitative difference between the South Summit and Lhotse — both are local maxima — but where the South Summit is only 11 metres higher than the pass between it and the main summit, Lhotse rises 610 metres above the South Col. This notion is known in mountaineering as the *topographic prominence* of a peak or summit: it is the minimum height one must descend before reaching higher terrain. In the field of mathematical morphology, the same notion is known as the *dynamics* of a regional maximum [Gri92].

Dynamics (or topographic prominence) is one of several attributes assigned to extrema, known as *extinction values*, which were originally introduced in [VM95], and later generalized in [Vac95]. The extinction value of a regional maximum, with respect to a decreasing family of anti-extensive connected operators $(\psi_\lambda)_{\lambda\geq 0}$, is the smallest value of λ for which this maximum disappears. By duality, it can be defined for regional minima as well.

Figure 3.1: Mount Everest as seen from Kala Patthar. The South Summit is the small peak on the immediate right of the main summit. (Wikimedia Commons, public domain)

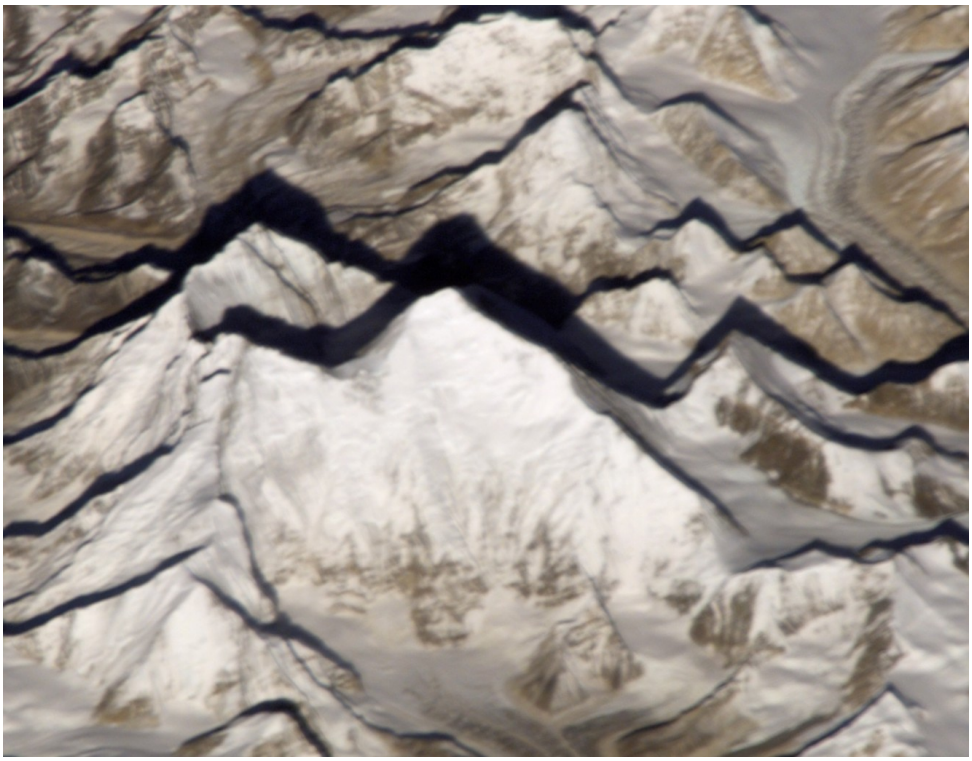


Figure 3.2: Lhotse (on the left) is connected to Mount Everest (center) by the South Col. (NASA, public domain)

Although many kinds of extinction values can be defined, the most commonly used are the *area* extinction value, related to the granulometry of *area openings* [Vin93], the *dynamics* [Gri92], related to h- reconstructions, and the *volume* [Vac95], related to volumic razings [Vac95].

These were introduced as a way to quantify the importance of regional extrema of an image, according to their size (area), contrast (dynamics) or a tradeoff between both (volume). The corresponding families of operators filter out 'small' extrema. The extinction values themselves are a common criterion to select relevant markers prior to watershed segmentation [Mac+15; Vac95]; histograms of extrema characteristics can also be used, in addition or replacement of granulometric functions [AS03].

Instead of assigning importance metrics to extrema, it is possible to use extinction values or their corresponding families of operators to obtain a decomposition of a function as a sum of smaller components, or *peaks*. This is illustrated in [Ala+17b], in the case of the dynamics; in section 3.2, we show how this idea can be generalized to other extinction values. Having obtained a decomposition as a sum of peaks — according to some criterion — new morphological operators can be defined by removing the 'smallest' peaks, not necessarily with the same definition of 'small'. For instance, it is possible to compute the area decomposition of a function, then keep the highest peaks, thus defining an operator preserving maxima that are either sufficiently large (in terms of their area extinction value) or sufficiently contrasted, contrary to classical operators like h-reconstructions or area openings, which erase all maxima with the same dynamics, or with the same area, respectively. For instance, an h-reconstruction with $h = 2$ would remove maxima A and C in Fig. 3.3 and alter maximum B; similarly, an area opening of size 2 would remove maxima A and B. As previously mentioned, the operators introduced in this work are able to encompass two properties of maxima, like area and dynamics, or dynamics and volume, and to filter out only maxima that are too 'small' in both senses. Likewise, the associated extinction values are able to quantify notions such as 'high enough *or* large enough'.

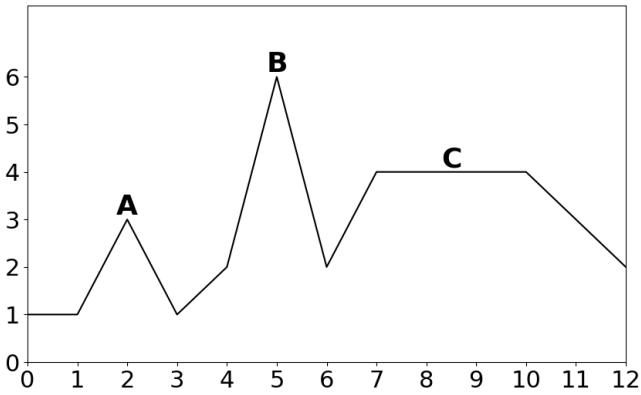


Figure 3.3: Maxima A and B have the same area extinction value (2) but different dynamics, while maxima A and C have the same dynamics but different area extinction values.

3.2 Extinction Decompositions: General Idea

This section intends to first present the main ideas behind extinction decompositions in the clearest possible way, before diving into the formal construction in the next section. We introduce here some required definitions, discuss our design choices, and illustrate the dynamics and area decompositions in a simple case.

3.2.1 Definitions

Although the concepts of regional maxima and extinction values are well-known, their definitions may suffer slight differences according to the problem at hand. In particular, we must specify how we handle the ambiguous case of constant functions, how we break ties between maxima in cases of equality, or which arbitrary value we assign to the dynamics of a global maximum. Another reason for redefining these concepts here is that the decomposition we introduce can be defined in a more general case than the usual framework for image processing: we consider a function $f : X \to Y$, but where X would typically be a pixel or voxel grid, the only requirement here is that X has a finite simple graph structure. Additionally, where Y is often required to be a set of integers, we only require that $Y \subset \mathbb{R}^+$ (but not $\mathbb{R}$, since, as we shall see, 0 plays a particular role). Note: in this work, we denote by $\mathbb{R}^+$ the set of all nonnegative real numbers, including zero, and by $\mathbb{N}$ the set of all nonnegative integers, including zero.

Definition 1 (Regional Maximum)**.** *A regional maximum of a function $f : X \to \mathbb{R}^+$ at level $y > 0$ is a non-empty connected set $M \subset X$ such that:*
$\forall x \in M,\ f(x) = y$ *and* $\mathcal{C}^M(X_y^+(f)) = M$

where $X_y^+(f) = \{x \in X | f(x) \geq y\}$ denotes the upper level set of f at level y, and $\mathcal{C}^M(S)$ (respectively $\mathcal{C}^x(S)$) denotes the connected component of set S containing set M (respectively point x).

In other words, a regional maximum M is commonly described as a *plateau* of f at level $y > 0$ with no higher neighbors [Vac95]. This definition does not require that f takes lower values elsewhere, so if f is equal to a constant $\lambda > 0$ on X, we consider that X is a regional (and a global) maximum of f at level λ. If, however, $f = 0$, we consider that f has no regional maximum. Following the moutaineering analogy, 0 can be thought as the sea level.

Definition 2 (Dynamics)**.** *The dynamics of a regional maximum M at level y is given by:*
$\mathrm{Dyn}(M) = y - \sup\{z | \exists x \in \mathcal{C}^M(X_z^+(f)), f(x) > y\}$
or
$\mathrm{Dyn}(M) = y$ *if* $\max(f) = y$

With this definition, the value we assign to global maxima might seem arbitrary. The definition of extinction values given in [Vac95] is actually based on families of operators. It is reminded here.

Definition 3 (Extinction values)**.** *Given $\Psi = (\psi_\theta)_\theta$ a decreasing family of connected, anti-extensive operators, the extinction value of a regional maximum M w.r.t. Ψ is given by:*
$\mathcal{E}_\Psi(M) = \sup\{\theta | \forall \theta' \leq \theta, M$ *is a subset of a regional maximum of* $\psi_{\theta'}(f)\}$

Dynamics can be defined this way, with the family of h-reconstructions: the h-reconstruction of a function f is the geodesic reconstruction [Beu01] of $f - h$ under f. For $h = \max(f) - \min(f)$, the h-reconstruction of f is a constant function, equal to $\min(f)$, which is why some authors would define the dynamics of global maxima to be equal to $\max(f) - \min(f)$. However, in order to be consistent with definition 1, we have to assign the value $\max(f)$ to the dynamics of global maxima, hence definition 2.

The *area* extinction value of a regional maximum M can be defined the same way, with the family of *area openings*[Vin93].

Definition 4 (Area opening)**.** *The area opening of size N of $f : X \to \mathbb{R}^+$ is defined for $x \in X$ by:*

$$\gamma_N^a(f)(x) = \sup\left(\{y | \#\{\mathcal{C}^x(X_y^+(f))\} \geq N\} \cup \{0\}\right)$$

where $\#S$ denotes the cardinal of set S. In other words, an area opening of size N razes maxima until their area (number of pixels) is at least N. In the - arguably degenerate - case where $N > \#X$, the set $\{y|\#\{\mathcal{C}^x(X_y^+(f))\} \geq N\}$ is empty, and we assign the value 0.

The volume extinction value corresponds to the family of *volumic razings* [Vac95].

Definition 5 (Volumic razing)**.** *The volumic razing of size $\varepsilon \geqslant 0$ of $f : X \to \mathbb{R}^+$ is defined for $x \in X$ by:*

$$r_\varepsilon^v(f)(x) = \sup\left(\{y| \sum_{z \in \mathcal{C}^x(X_y^+(f))} (f(z) - y) \geq \varepsilon\} \cup \{0\}\right)$$

Like area openings, volumic razings replace peaks with flat zones, but unlike area openings, they are not idempotent. It must be noted that the set

$$\{y| \sum_{z \in \mathcal{C}^x(X_y^+(f))} (f(z) - y) \geq \varepsilon\}$$

is never empty if we allow y to take negative values. Here, we impose that $r_\varepsilon^v(f)$ is still a function from X to $\mathbb{R}^+$, even for large values of ε.

In this article, we introduce a new kind of extinction value, namely the L^1 extinction value, which is similar to the volume in that it constitutes a tradeoff between area and contrast. It is hard to give a formal definition here, without first introducing the corresponding decomposition and family of operators, but the notion and the way it differs from the volumic extinction value should appear clear in the following.

3.2.2 Label Propagation

The algorithm for computing the decomposition is very similar to the algorithm for computing the extinction values of the maxima of f, which itself bears much resemblance to the watershed computation algorithm [VS91]. We start by labeling all regional maxima; we then consider the image's upper levels in decreasing order of height and propagate the labels. Figure 3.4 illustrates the first steps of this process, which are common to all these algorithms (except for the fact that the watershed algorithm usually operates on minima rather than on maxima): if a connected set of an upper level contains only one label, it is assigned this label.

When two labels meet at some point M, as in Fig. 3.4c, we can compute the extinction value of the 'smallest' maximum and propagate the label of the 'greatest' one, the meaning of 'smallest' and 'greatest' depending on the criterion chosen for the decomposition. In the case of dynamics, maximum A is lower than maximum B, so M would be assigned the label B, and at this point we would learn that the dynamics of A is equal to $f(A) - f(M) = 4$. In the case of the area extinction value, here, the surface labeled A so far is greater than the surface labeled B, so M would be labeled with A, and we would compute the area extinction value of B, which is simply the number of pixels labeled B at this step of the algorithm. The same idea applies for the volume or other extinction values, as long as they can be calculated at this point. In the subsequent steps of the algorithm, the label of the smallest maximum is replaced by the label of the greatest, and the process iterates until convergence.

3.2.3 Sketch of the Decomposition Algorithm

When labels meet, instead of computing the extinction value of the smallest maximum, we can assign a function to it: the part of the graph that has been colored with this label so far. In Fig. 3.4c, if we compute a decomposition based on dynamics, A is the smallest maximum and we can assign to it the blue function; if the decomposition is based on area, B is the smallest maximum and we can assign to

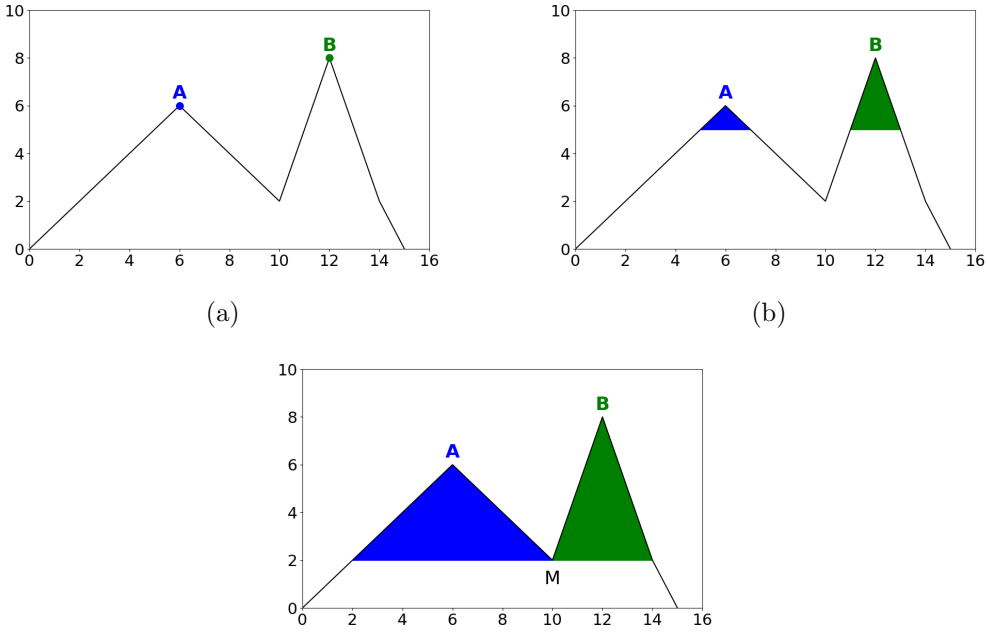
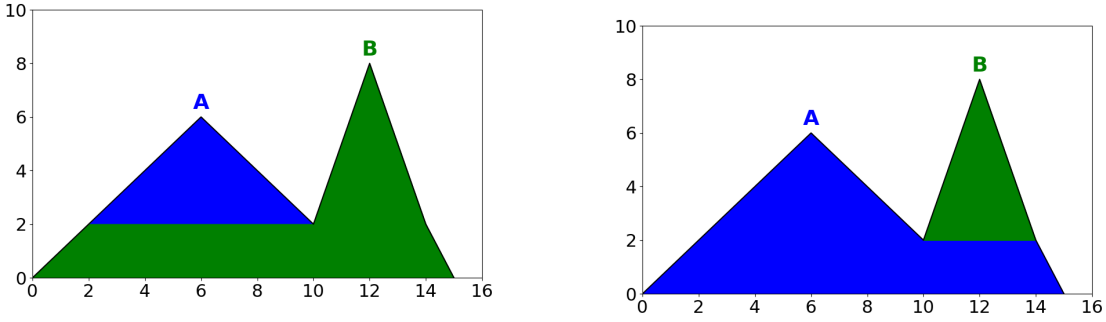


Figure 3.4: First steps of all label-propagating algorithms: maxima A and B are assigned different labels (a), labels a propagated downwards (b), labels meet at point M (c)

it the green function. We could also compare the L^1 norms (integrals) of the green and blue function, which is what we will do in order to define the L^1 decomposition in the next section.

More formally, when two or more labels $l_1, l_2, \ldots$ meet at level y, for the labels that go extinct at this step, we define the *peak* $f_{l_i} = f - y$ restricted to the part of the domain that has been labeled with l_i. In the case of dynamics decomposition, in Fig. 3.4c, we assign to A the function $f_A = f - 2$ on the interval $[2; 10]$. In the case of area decomposition, we would assign to B the function $f_B = f - 2$ on the interval $[10; 14]$. Before proceeding to the next step of the algorithm, we retrieve the component f_{l_i} from f, and replace label l_i with the label of the highest neighboring maximum. Figure 3.5 illustrates this: labels A and B meet at level 2; the smaller maximum goes extinct at this step, while the greater maximum's label continues propagating below.



(a) Dynamics decomposition: A is the first maximum to go extinct, while label B continues propagating.

(b) Area decomposition: B goes extinct at level 2, label A propagates at levels 1 and 0.

Figure 3.5: Dynamics and area decompositions.

Equality Cases

It is possible that when labels meet, two or more corresponding maxima are the 'greatest' in the sense of the chosen criterion. It should be emphasized that this is not a degenerate case, especially if the chosen criterion is the dynamics, since even a moderate-sized image can have many more maxima than grayscale levels. A solution would be to break ties arbitrarily; however, this leads to similar maxima being associated very dissimilar peaks, yet the operators to be introduced in Sec. 3.6 rely on properties of the peaks, such as area or volume; arbitrarily breaking ties would eventually lead to some part of the image being arbitrarily filtered out rather than another for no meaningful reason. In order to avoid this, in case of equality between two or more 'greatest' maxima, we merge the corresponding labels.

Maximum Fusion Tree

The idea of a minimum fusion tree was introduced in [Vac95]. In the present work, we focus on maxima rather than minima, but the idea is the same: the maximum fusion tree describes which label corresponded to the greatest maximum at the step when a label went extinct: in Fig. 3.5a, A is extinguished by B, which we denote by making B the father of A in that scenario, while A is the father of B in Fig. 3.5b. The maximum fusion tree depends on the chosen extinction value; actually, because of the way equality cases are handled, even the number of nodes may differ.

Memory Representation

Aside from the equality cases mentioned above, the decomposition we are presenting associates one component, or peak, to each regional maximum of a function. In both illustrations of Fig. 3.5, the decomposition is simply $f = f_A + f_B$.

Storing each peak as a separate function would be very inefficient in terms of memory, and infeasible in practice even for moderately large images. Instead, the decomposition can be stored as a label image, a maximum fusion tree, and a list of extinction heights. Their computation is detailed in section 3.3, but the idea can be illustrated here. The label image contains the first label to be assigned to each pixel (even if in subsequent steps of the algorithm, the pixel is assigned the label of a greater maximum). If peaks are visualized as stacked on top of each other as in Fig. 3.5, the image label contains the label of the top-most peak; in Fig. 3.5a, pixels x between 2 and 10 are labeled A (blue), and other pixels are labeled B (green).

Given the image label, the maximum fusion tree and the list of extinction heights, we can easily retrieve the peak associated to a label l: its support is the set of points labeled with l or one of its descendants in the maximum fusion tree. If we denote EH_j the extinction height of label j, for points labeled l, f_l is equal to $f - EH_l$, and for points labeled with a descendant d of l, f_l is equal to $EH_m - EH_l$, where m is the son of l on the branch leading to d.

3.3 Extinction Decompositions

3.3.1 Labeling Algorithm

We consider a function $f : X \to \mathbb{R}^+$, where set X has a finite simple graph structure. In particular, set X is finite, so $Y = f(X) = \cup_{x \in X}\{f(x)\}$ is finite. We denote $y_1, y_2, \ldots y_n$ the *non-zero* values of Y, in ascending order: $0 < y_1 < y_2 < \ldots < y_n$.

As mentioned in section 3.2.3, the decomposition is stored as a label 'image' or function $L : X \to \mathbb{N}$, a maximum fusion tree, and a list of extinction heights. In order to keep track of which labels are still 'active', we will also use a temporary label image $T : X \to \mathbb{N}$; labels in T will be called active, while

the others will be called inactive. The first step consists in labeling all regional maxima $M_1, M_2, \ldots$, in both the final and temporary label images:

$$L(x) = T(x) = \begin{cases} l & \text{if } x \in M_l \\ 0 & \text{elsewhere.} \end{cases}$$

For each level y in descending order (from y_n to y_1), we consider the connected components $\mathcal{C}^1, \mathcal{C}^2, \ldots$ of X_y^+. For each such connected component $\mathcal{C}$:

- find the label m of the greatest maximum included in $\mathcal{C}$, in the sense of the chosen criterion
- in case of equality between several greatest maxima, merge their labels first (in both L and T)
- label unlabeled pixels of $\mathcal{C}$ with m:

$$\forall x \in \mathcal{C} | L(x) = T(x) = 0, \begin{cases} L(x) & \leftarrow & m \\ T(x) & \leftarrow & m \end{cases}$$

- for each active label $l \neq m$ in $\mathcal{C}$:
  - make m the parent of l in the maximum fusion tree
  - store the extinction height $EH_l = y$
  - deactivate label l:

$$\forall x \in \mathcal{C} | T(x) = l, \ T(x) \leftarrow m$$

The algorithm stops just before reaching level zero, meaning that pixels x such that $f(x) = 0$ remain labeled with $L(x) = 0$. If the support of f is not connected, several labels are still active after level y_1, and we have not defined a maximum fusion tree, but a maximum fusion forest, with exactly one tree per connected component of the support. In the present work, it does not matter whether we obtain a tree or a forest, and if the support of f is disconnected, we can process each connected component separately. Following the mountaineering analogy of the introduction, with 0 the sea level, each maximum fusion tree corresponds to one island.

A detailed example of the four decompositions condidered in this work is given in section 3.4.

3.3.2 Attribute Updates

Let us denote by N_L the number of non-zero values in the label image L (which might be strictly inferior to the number of regional maxima because of label merging), and if necessary, modify the names of the labels so they range from 1 to N_L. We have obtained a decomposition $f = \sum_{l=1}^{N_L} f_l$; in the next section, we define operators based on attributes of these peaks, such as

maximum value:	$\delta(l)$	$=$	$\max(f_l)$
area:	$\alpha(l)$	$=$	$\#\{x \in X \mid f_l(x) \neq 0\}$
L^1 norm:	$\lambda(l)$	$=$	$\|f_l\|_1 = \sum_{x \in X} f_l(x)$
volume:	$\mu(l)$	$=$	$\|f_l\|_1 + \sum_{d \in D(l)} \|f_d\|_1$

where $D(l)$ denotes the descendants of l in the maximum fusion tree.

Notice the difference between the definitions of the λ and μ attributes: the first is simply the L^1 norm of the peak and does not depend on anything else; the second may seem less natural, as it depends not only on the considered peak, but also on neighboring smaller peaks. As we will see in Sec. 3.5, the μ attribute is defined this way so that it is equal to the volumic extinction value if the volume is chosen as the discriminating criterion. A detailed example is provided in the next section.

Instead of computing these four attributes *a posteriori*, we can compute them along with the decomposition. In the initialization step, we assign to the regional maximum M_l the values:

$$\begin{array}{rcl} \delta(l) & = & f(M_l) \\ \alpha(l) & = & \#M_l \\ \lambda(l) & = & \sum_{x\in M_l} f(x) = \#M_l \times f(l) \\ \mu(l) & = & \sum_{x\in M_l} f(x) = \#M_l \times f(l) \end{array}$$

The update rules at level y for each component $\mathcal{C}$ of $X_y^+(f)$ become:

- find the label m of the greatest maximum included in $\mathcal{C}$, in the sense of the chosen criterion
- in case of equality between k greatest maxima m_1, ..., m_k, merge their labels into $m = m_1$:

$$\forall x \in \mathcal{C} | L(x) \in \{m_2, \dots m_k\}, \left\{ \begin{array}{rcl} L(x) & \leftarrow & m_1 \\ T(x) & \leftarrow & m_1 \end{array} \right.$$

- and merge their attributes:

$$\begin{array}{rcl} \delta(m_1) & \longleftarrow & \max\limits_{1\le i\le k} \delta(m_i) \\ \alpha(m_1) & \longleftarrow & \sum\limits_{i=1}^{k} \alpha(m_i) \\ \lambda(m_1) & \longleftarrow & \sum\limits_{i=1}^{k} \lambda(m_i) \\ \mu(m_1) & \longleftarrow & \sum\limits_{i=1}^{k} \mu(m_i) \end{array}$$

- update m's area, L^1 norm and volume attributes:

$$\begin{array}{rcl} \alpha(m) & += & \#\{x \in \mathcal{C} | L(x) = 0\} \\ \lambda(m) & += & y \times \#\{x \in \mathcal{C} | L(x) = 0\} \\ \mu(m) & += & y \times \#\{x \in \mathcal{C} | L(x) = 0\} \end{array}$$

- label unlabeled pixels of $\mathcal{C}$ with m:

$$\forall x \in \mathcal{C} | L(x) = T(x) = 0, \left\{ \begin{array}{rcl} L(x) & \leftarrow & m \\ T(x) & \leftarrow & m \end{array} \right.$$

- for each *active* label $l \neq m$ in $\mathcal{C}$:
  - make m the parent of l in the maximum fusion tree
  - increment m's area, L^1 norm and volume:

$$\begin{array}{rcl} \alpha(m) & += & \alpha(l) \\ \lambda(m) & += & y\alpha(l) \\ \mu(m) & += & \mu(l) \end{array}$$

  - store the extinction height $EH_l = y$
  - compute label l's final attributes:

$$\begin{array}{rcl} \delta(l) & \leftarrow & \delta(l) - y \\ \lambda(l) & \leftarrow & \lambda(l) - y\alpha(l) \\ \mu(l) & \leftarrow & \mu(l) - y\alpha(l) \end{array}$$

– deactivate label l:

$$\forall x \in \mathcal{C} | T(x) = l,\ T(x) \leftarrow m$$

No specific post-processing is required for labels still active after the last level $y_1 > 0$; their attributes are already correct.

3.3.3 Finding the Greatest Maximum

When computing the dynamics decomposition, the greatest maximum is the one with the highest value, so we can directly compare the values of $\delta(l)$; similarly, when computing the area decomposition, we can directly compare the values of $\alpha(l)$. In the case of the volume decomposition, in compliance with definition 5, we need to consider the volume above level y, which is equal to $\mu(l) - y\alpha(l)$.

3.3.4 L^1 decomposition

In addition to the three former decompositions (dynamics, area, volume), it is also possible to choose the greatest maximum as the one associated to the component with the greatest L^1 norm. Formally, at height y, if we consider that the 'greatest' maximum of a connected component $\mathcal{C}$ is the one with the highest value of $\lambda(l) - y\alpha(l)$, we obtain another decomposition.

3.4 Detailed Example

In this section, we will illustrate the decompositions on a synthetic example, shown in Fig. 3.6. In the following, we will use the subscripts d, a, L^1 and v to disambiguate between the dynamics, area, L^1 and volume decompositions, respectively.

The heights and areas of the four regional maxima and three flat zones F_1, F_2, F_3 are summarized in Table 3.1. We will think of the lowest non-zero flat zone as the 'background' B, although there is a lower flat zone Z at level zero. The background B will always be labeled last; it will be attributed the label of the 'greatest' of the four maxima - in the sense of the chosen criterion - and Z will not be labeled, so they are not included in the table.

Zone	Height	Area
M_1	10	9
M_2	8	40
M_3	5	66
M_4	6	35
F_1	2	6
F_2	3	9
F_3	4	6

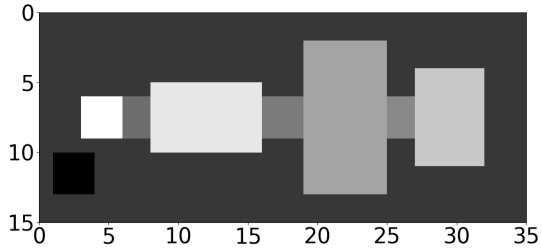
Table 3.1: Properties of the main flat zones of the image.

The first step of labeling algorithms consists in labeling the regional maxima, in this case M_1, M_2, M_3 and M_4, as illustrated in Fig. 3.6b.

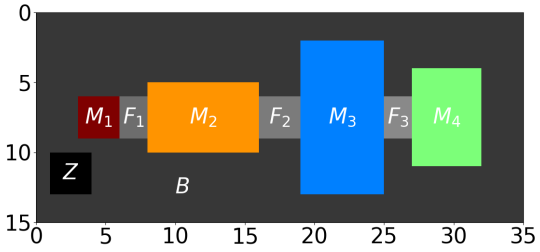
3.4.1 First labeling step, level $y = 4$

Our example image has 9 different values: in decreasing order, 10, 8, 6, 5, 4, 3, 2, 1 and 0. The first four values correspond to regional maxima, which are already labeled.

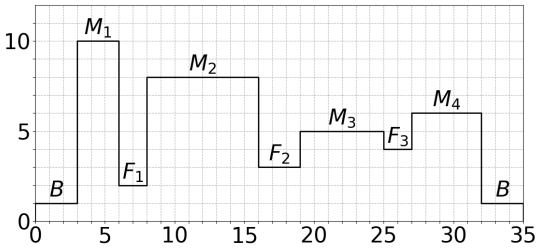
The first level to consider is thus $y = 4$. The upper level set at this value, $X_4^+(f)$, has three connected components: M_1, M_2 and $M_3 \cup F_3 \cup M_4$; M_1 and M_2 contain no unlabeled pixels, so the



(a) Original image



(b) First step of all decomposition algorithms. The regional maxima M_1 to M_4 are labeled (here with different colors) and the non-maximal flat zones annotated.



(c) Image profile along the middle lines.

Figure 3.6: Example image with four regional maxima and five non-maximal flat zones. The flat zone denoted by Z is at level 0 and thus will not be labeled. The flat zone denoted by B is the background, at level 1, which will be labeled last. The three flat zones F_1, F_2, F_3 will be labeled in decreasing order of height, according to the chosen criterion for a given decomposition.

only connected component to consider here is the last one, which contains an unlabeled flat zone, F_3, and two regional maxima: M_3 and M_4. Depending on the chosen criterion, either M_3 or M_4 will be considered the greater: F_3 will be labeled accordingly, and the smaller maximum will have its label deactivated.

Since we will illustrate the four decompositions in parallel, we will use the subscripts d, a, L^1 and v in order to specify the considered decomposition.

Dynamics decomposition

In the case of the dynamics decomposition, we can simply compare the heights of the maxima: M_4 is higher than M_3, so F_3 gets labeled with M_4's label, and M_3 gets deactivated. At this point we can finalize M_3's attributes. In particular we have $\delta_d(M_3) = 1$, meaning that the dynamics of M_3 is equal to 1. We do not know yet the dynamics of M_4; the temporary value remains $\delta_d(M_4) = 6$.

Area decomposition

Here M_3 is the greater maximum: its area is 66, whereas the area of M_4 is only 35. F_3 gets labeled with M_3's label, and it is M_4 that gets deactivated. Here we obtain the final value $\delta_a(M_4) = 2$, which is *not* the dynamics of M_4, but a novel kind of importance measure, encompassing both the relative height and spatial range of a maximum compared to its surroundings.

L^1 decomposition

We have to evaluate the quantities $\lambda_{L^1}(l) - y\alpha_{L^1}(l)$ for M_3 and M_4. At initialization we had:

$$\begin{array}{rclcl} \lambda_{L^1}(M_3) & = & 66 \times 5 & = & 330 \\ \lambda_{L^1}(M_4) & = & 35 \times 6 & = & 210 \\ \alpha_{L^1}(M_3) & = & 66 & & \\ \alpha_{L^1}(M_4) & = & 35 & & \end{array}$$

The calculations yield $\lambda_{L^1}(M_3) - y\alpha_{L^1}(M_3) = 66$ and $\lambda_{L^1}(M_4) - y\alpha_{L^1}(M_4) = 70$. M_4 is greater than M_3 in the sense of the L^1 criterion, so F_3 gets labeled with M_4's label, and M_3 gets deactivated.

We can compute the final attribute values $\lambda_{L^1}(M_3) = \mu_{L^1}(M_3) = 66$, and we can update the (temporary) values for M_4, yielding $\lambda_{L^1}(M_4) = 498$ and $\mu_{L^1}(M_4) = 564$.

Volume decomposition

At this step, since the λ and μ attributes are initialized in the same way, the exact same reasoning as in the previous paragraph is applicable.

3.4.2 Step at level $y = 3$

The upper level set $X_3^+(f)$ has two connected components: M_1 and $M_2 \cup F_2 \cup M_3 \cup F_3 \cup M_4$. Again, nothing needs to be done about M_1, and only the second connected component has to be considered. It contains three regional maxima, but only two active labels: M_2 and whichever of M_3 and M_4 was considered the greater in the previous step.

Dynamics decomposition

Since M_3 was deactivated in the previous step, we simply have to compare the heights of M_2 and M_4; M_2 is higher, so at this step F_2 gets labeled with M_2's label and M_4 gets deactivated. We can compute its final attributes; in particular, we learn that its dynamics is 2.

Area decomposition

In the previous step, the α attribute of M_3 was updated to $\alpha_a(M_3) = \#M_3 + \#F_3 + \#M_4 = 107$, which is greater than $\alpha_a(M_2) = \#M_2 = 40$. Therefore, F_2 gets M_3's label and M_2 is deactivated.

L^1 decomposition

In order to determine which of M_2 or M_4 is the greater maximum, we have to compare the values $\lambda_{L^1}(M_2) - y\alpha_{L^1}(M_2)$ and $\lambda_{L^1}(M_4) - y\alpha_{L^1}(M_4)$. At this point, $\lambda_{L^1}(M_2)$ and $\alpha_{L^1}(M_2)$ have not been modified since their initializations to $\lambda_{L^1}(M_2) = 8 \times 40 = 320$ and $\alpha_{L^1}(M_2) = \#M_2 = 40$. On the contrary, in the previous step at $y = 4$, the values for M_4 have been updated to $\lambda_{L^1}(M_4) = 498$ and $\alpha_{L^1}(M_4) = \#M_3 + \#F_3 + \#M_4 = 66 + 6 + 35 = 107$.

The calculations yield $\lambda_{L^1}(M_2) - y\alpha_{L^1}(M_2) = 320 - 3 \times 40 = 200$ and $\lambda_{L^1}(M_4) - y \times \alpha_{L^1}(M_4) = 498 - 3 \times 107 = 177$. Therefore, M_4 gets deactivated and F_2 gets the label of M_2.

Volume decomposition

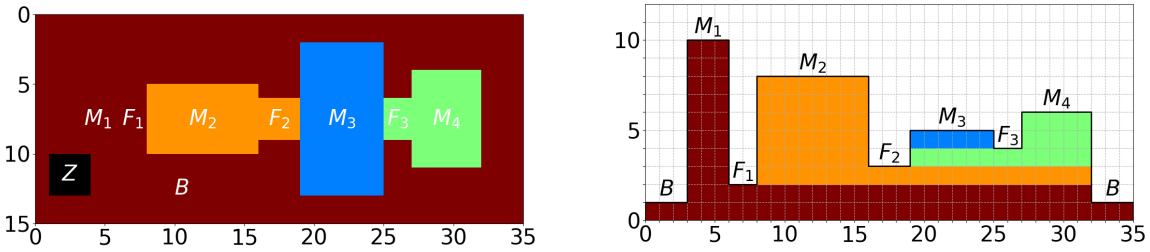
Most of the reasoning is the same as above, but we have to compare the quantities $\mu_v(M_2) - y\alpha_v(M_2) = 200$ and $\mu_v(M_4) - y\alpha_v(M_4) = 564 - 3 \times 107 = 243$. This time, M_4 is considered the greater maximum; M_2 is deactivated and F_2 gets labeled with M_4.

3.4.3 Last steps: $y = 2$ and background

The step at level $y = 2$ is straightforward: there is only one connected component to consider, containing all four maxima, and the only case where M_1 is considered the greatest is the dynamics decomposition.

The final step at level $y = 1$ is even simpler: the upper level set $X_1^+(f)$ consists of all the image with the exception of Z and contains only one active label at this point.

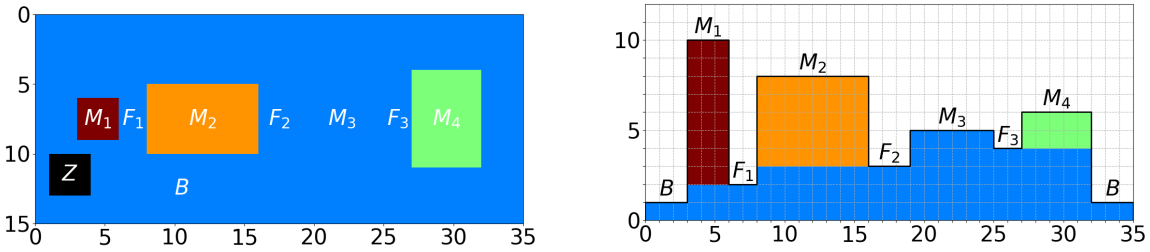
The final label images, as well as a stacked representation of the four components, are shown in Figs. 3.7, 3.8, 3.9, and 3.10.



(a) Final label image for the dynamics decomposition

(b) Stacked view of the middle lines

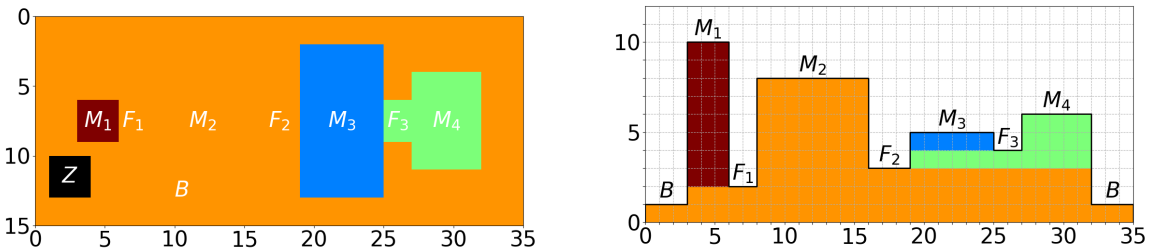
Figure 3.7: Completed dynamics decomposition



(a) Final label image for the area decomposition

(b) Stacked view of the middle lines

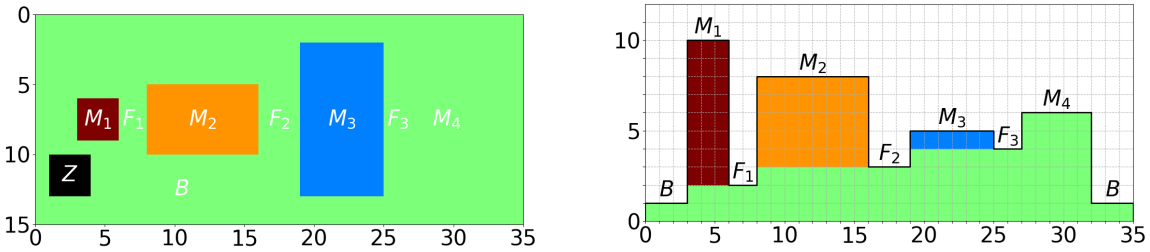
Figure 3.8: Completed area decomposition



(a) Final label image for the L^1 decomposition

(b) Stacked view of the middle lines

Figure 3.9: Completed L^1 decomposition



(a) Final label image for the volume decomposition

(b) Stacked view of the middle lines

Figure 3.10: Completed volume decomposition

3.5 Decomposition Properties

3.5.1 Discriminating Criterion Conservation

By construction, the dynamics of a regional maximum M is equal to its δ attribute in the dynamics decomposition, $\delta_d(M)$.

Similarly, its area extinction value is equal to its α attribute in the area decomposition $\alpha_a(M)$, and its volumic extinction value is equal to its μ attribute in the volume decomposition $\mu_v(M)$.

By analogy, we can define the L^1 extinction value of a maximum as its λ attribute in the L^1 decomposition, $\lambda_{L^1}(M)$. It is indeed an extinction value as per definition 3; the corresponding family of operators is introduced in the next section.

3.5.2 Support Inclusion

For any of the four decompositions $f = \sum f_l$, if s is a descendant of p in the corresponding maximum fusion tree (or forest), then the support of f_s is strictly included in the support of f_p. It immediately follows that $\alpha(f_s) < \alpha(f_p)$: the α attribute is increasing when going from leaf to root.

3.5.3 Root-to-leaf Decreasingness of the μ Attribute

For all decompositions, the μ attribute is clearly increasing when going from leaf to root. In the case of the volume decomposition, $\mu_v(l)$ is the volumic extinction value of maximum l; for the three other decompositions, it is not a classic measurement. It can still be thought of as a 'volume', but it is the volume above EH_l on the support of f_l; in more precise terms, it is the sum of the L^1 norms of the component f_l and of all its descendants.

3.5.4 Root-to-leaf Decreasingness of the Discriminating Criterion

In addition to the former properties, which hold true for any of our four decompositions, by construction, the criterion (dynamics, area, volume, L^1 norm) chosen to perform the decomposition is also increasing from leaf to root of the corresponding fusion tree. Namely, this property states that δ_d, α_a, λ_{L^1} and μ_v are decreasing from root to leaf.

3.5.5 Other Attributes

The α and μ attributes are increasing from leaf to root in all decompositions; this is also true of the δ attribute for the dynamics decomposition and the λ attribute in the L^1 decomposition.

There are no monotony properties for δ and λ in the other cases: in the example from Section 3.4, we have for instance $\delta_a(M_1) = 8$ and $\delta_a(M_3) = 5$, although M_3 is the parent of M_1 in the area decomposition.

Although in all decompositions of Section 3.4, the λ attribute is decreasing from root to leaf, it is easy to construct examples, even with only two maxima, where this is not true for λ_d and λ_a (consider a narrow peak and a large but lower one). A counter-example for λ_v with three maxima is given in Fig. 3.11: B is the root of the fusion tree but $\lambda_v(A) > \lambda_v(B)$.

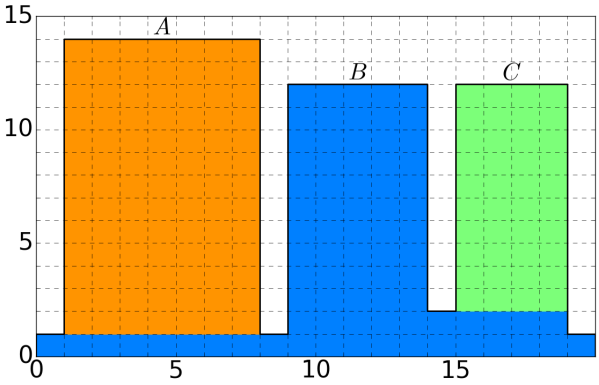


Figure 3.11: Volumic decomposition example in 1 dimension. The L^1 norms (areas) are $\lambda_v(A) = 91$, $\lambda_v(B) = 80$ and $\lambda_v(C) = 40$. The area colored in orange (91) is less than the sum of the blue and green areas above level 1 (100), so B is considered greater than A.

3.6 Associated Operators

3.6.1 Leaf Removal

Before considering more general operations, let us consider the simple case where one leaf component is removed from a decomposition. Let $f = \sum_{l=1}^{N} f_l^c$ be one of the decompositions defined in the previous section, with respect to criterion c. Let us assume for simplicity that no label merging occurs, meaning that f has exactly N regional maxima, and that the leaf we remove is labeled N (meaning N is a leaf in the maximum fusion tree for this particular criterion c). Let us now consider the function $g = f - f_N$. What happens if we compute the decomposition of g with respect to the same criterion c?

By construction, we have $g = f$ everywhere except on the support of f_N, where g is constant and equal to EH_N (note that for g, this is a non-maximal plateau). Considering the algorithm from section 3.3, it is clear that the first steps, for $y > EH_N$, are exactly the same for f and g, except for the propagation of label N.

Let us consider the point where the algorithm has just reached level $y = EH_N$, and let us consider the connected component $\mathcal{C}$ of $X_y^+(f)$ containing the support of f_N. Since N goes extinct at this height, the greatest maximum of f included in $\mathcal{C}$ has a label $m \neq N$. Since $g = f$ everywhere except on the support of f_N where $g = EH_N = y$, the decomposition algorithm for g will at this point treat the same component $\mathcal{C}$ and find the same greatest maximum m.

At this point, just before they are updated, the area, L^1 norm and volume attributes of m are the same for f and g, and the label images L_f and L_g are equal everywhere except on the support of f_N, where $L_f = N$ and $L_g = 0$. After the updates:

$$\begin{aligned}\alpha(m) &+= \#\{x \in \mathcal{C} | L(x) = 0\} \\ \lambda(m) &+= y \times \#\{x \in \mathcal{C} | L(x) = 0\} \\ \mu(m) &+= y \times \#\{x \in \mathcal{C} | L(x) = 0\}\end{aligned}$$

we obtain:

$$\begin{aligned}\alpha_g(m) &= \alpha_f(m) + \alpha_f(N) \\ \lambda_g(m) &= \lambda_f(m) + y\alpha_f(N) \\ \mu_g(m) &= \mu_f(m) + y\alpha_f(N).\end{aligned}$$

The next step is labeling the unlabeled pixels of $\mathcal{C}$: after this, we have $L_f = L_g$ and $T_f = T_g$ everywhere except on the support of f_N, where at this point $L_f = T_f = N$ and $L_g = T_g = m$.

Then, we consider the active labels $l \neq m$ in $\mathcal{C}$: these are exactly the same for f and g except for label N, so the increments in m's area, L^1 norm and volume are the same except for the extra increments:

$$\begin{aligned}\alpha_f(m) &+= \alpha_f(N) \\ \lambda_f(m) &+= y\alpha_f(N) \\ \mu_f(m) &+= \mu_f(N)\end{aligned}$$

These exactly compensate the extra increments we had for $\alpha_g(m)$ and $\lambda_g(m)$, so we now have $\alpha_f(m) = \alpha_g(m)$ and $\lambda_f(m) = \lambda_g(m)$, but we have:

$$\mu_f(m) - \mu_g(m) = \mu_f(N) - y\alpha_f(N) = \|f_N\|_1.$$

The rest of this step - storing extinction heights, computing final attributes and deactivating labels that are not m - is exactly the same for f and g.

If the criterion c is area, dynamics or L^1-norm, at subsequent steps $y < EH_N$, things remain the same for f and for g until the algorithm ends. This means that if we write the c-decompositions, respectively:
$f = \sum_{l=1}^{N} f_l^c$ and $g = \sum_{l=1}^{N-1} g_l^c$, we have for all l in $\{1, \ldots, N-1\}$, $f_l^c = g_l^c$, as well as for all $l \in 1, \ldots N-1$:

$$\begin{aligned}\alpha_f(l) &= \alpha_g(l) \\ \delta_f(l) &= \delta_g(l) \\ \lambda_g(l) &= \lambda_g(l)\end{aligned}$$

The former reasoning applies even if label mergings occur: the area, dynamics and L^1-norm decompositions are stable by leaf removal.

This is not true for the volume decomposition; for instance, let us consider the very simple example given in Fig. 3.11. C is a leaf in the maximum fusion tree, but removing f_C changes the volume decomposition: A would become the root instead of B, and components f_A and f_B would be different.

3.6.2 Thresholding Operators

In the former sections, we have described four decompositions - actually, any binary granulometry defines such a decomposition, but in the present work, we will restrict our attention to these four. Considering a decomposition associated to one of these criterions, c, then thresholding on an attribute θ over a value $\varepsilon \geq 0$ (θ being one the four attributes α, δ, λ or μ), we can define the operator

$$\begin{aligned}T_\varepsilon^{c,\theta}(f) &= \sum_{l|\theta(l) \geqslant \varepsilon} f_l^c \\ &= f - \sum_{l|\theta(l) < \varepsilon} f_l^c\end{aligned}$$

This operator leaves intact the peaks that are important enough while filtering out those who are not - the meaning of 'important' here being defined by the chosen decomposition and criterion. In particular, thresholding operators are antiextensive :

Property 1. *Antiextensivity*
$\forall f : X \to \mathbb{R}^+, T_\varepsilon^{c,\theta}(f) \leqslant f.$

In the particular cases of $T_\varepsilon^{d,\delta}$, $T_\varepsilon^{a,\alpha}$ and $T_\varepsilon^{v,\mu}$, they are less active than their 'usual' counterparts: h-reconstruction, area opening and volumic razing, respectively.

Property 2. *Activity*
$\forall f : X \to \mathbb{R}^+, \forall \varepsilon \geqslant 0,$

$$\begin{array}{ccccc} G_\varepsilon(f) & \leqslant & T_\varepsilon^{d,\delta}(f) & \leqslant & f \\ \gamma_\varepsilon^a(f) & \leqslant & T_\varepsilon^{a,\alpha}(f) & \leqslant & f \\ r_\varepsilon^v(f) & \leqslant & T_\varepsilon^{v,\mu}(f) & \leqslant & f \end{array}$$

where $G_\varepsilon(f)$ denotes the geodesic reconstruction of $f - \varepsilon$ under f, γ_ε^a the area opening of size ε and r_ε^v the volumic razing of size ε.

This is illustrated on the toy image shown in Fig. 3.12. Figure 3.13 shows the difference between an h-reconstruction and the operator $T_\varepsilon^{d,\delta}$ with the same value. Figure 3.14 compares area opening and $T_\varepsilon^{a,\alpha}$. Finally, Fig. 3.15 compares volumic razing and $T_\varepsilon^{v,\mu}$.

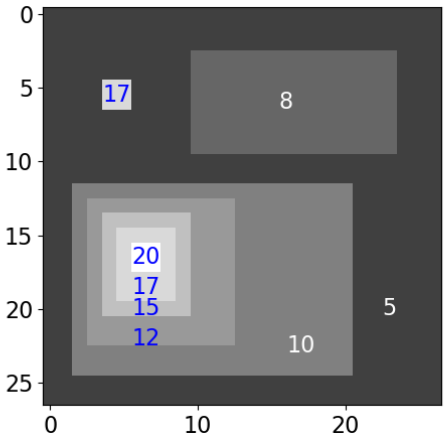


Figure 3.12: Toy image used to illustrate the activities of operators $T_\varepsilon^{d,\delta}, T_\varepsilon^{a,\alpha}$ and $T_\varepsilon^{v,\mu}$ compared to their usual counterparts. The numbers indicate the gray level of each flat zone.

Another property of the thresholding operators $T_\varepsilon^{c,\theta}$ is the fact that they do not create new contours. More formally, they are *connected* operators [SOG98], of which we remind the definition here:

Definition 6. *Connected operator*
An operator T is connected if for all function f, the partition of the flat zones of f is finer than the partition of the flat zones of $T(f)$.

Property 3. *Connectedness*
For all c, θ and ε, $T_\varepsilon^{c,\theta}$ is a connected operator.

Proof. In order to prove this property, let us consider the removal of one component f_l. By construction, any flat zone of f is either entirely included in the support of f_l, or entirely outside of it (the support of f_l is a connected component of the strict upper level set $\{x \in X | f(x) > EH_l\}$). Similarly,

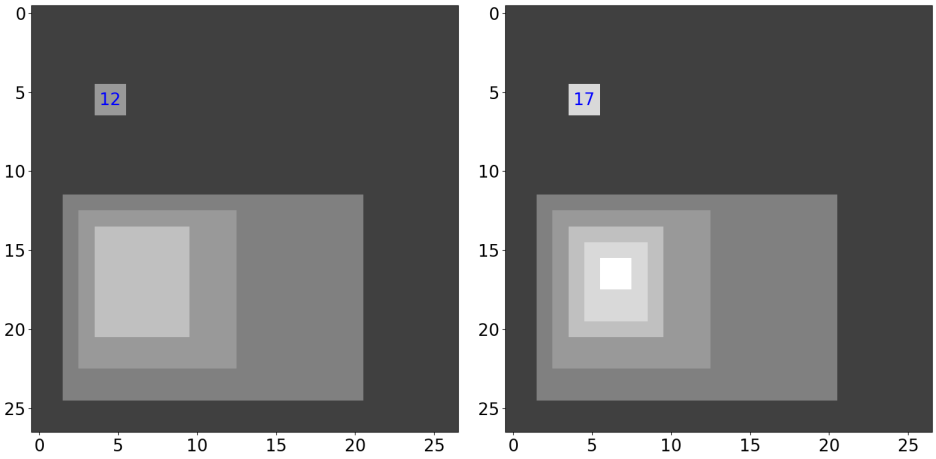


Figure 3.13: Comparison between an h-reconstruction with $h = 5$ on the left, and $T_5^{d,\delta}$ on the right. In both cases, the maximum on the top right of the original image disappears, but the maxima that are kept are unaltered by $T_5^{d,\delta}$, whereas they are modified by the h-reconstruction.

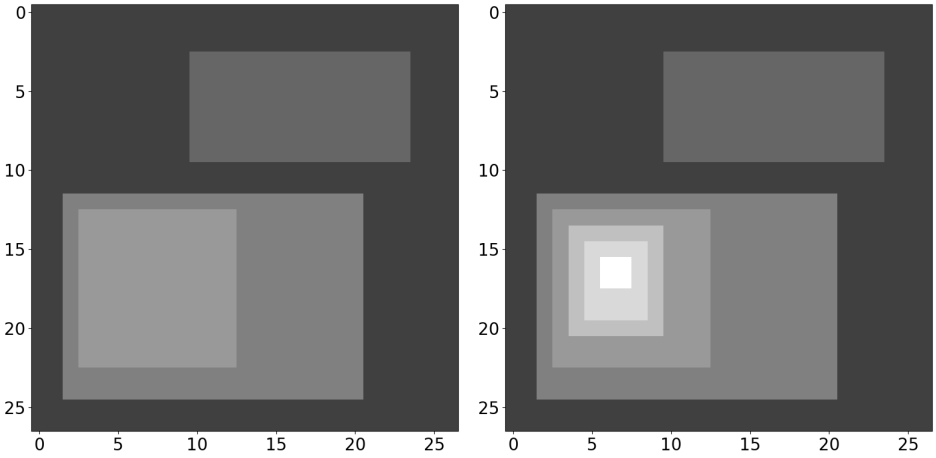


Figure 3.14: Comparison between an area opening of size 50 on the left, and $T_{50}^{a,\alpha}$ on the right. In both cases, the maximum on the top left disappears and the maximum on the top right is kept intact. The maximum at the bottom is unaltered by $T_{50}^{a,\alpha}$, whereas it is partly erased by the area opening.

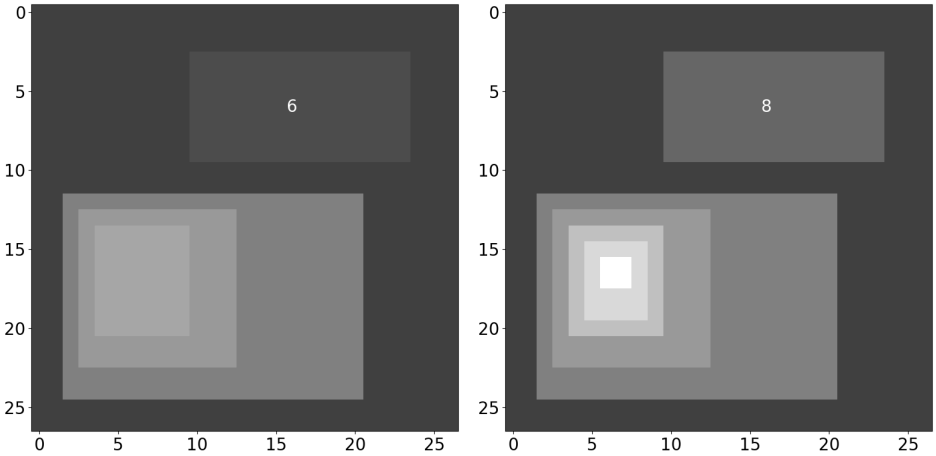


Figure 3.15: Comparison between a volumic razing of volume 100 on the left and $T^{v,\mu}_{100}$ on the right. The maximum on the top left is erased by both; the other two maxima are partly erased by the volumic razing but kept intact by $T^{v,\mu}_{100}$. Note: in this example, we use a slightly different definition for the volumic razing so that the resulting image is integer-valued.

any flat zone in the support of f_l is either entirely within the support of f_s with s a son of l in the fusion tree, or entirely within the support of f_l but outside the support of any descendant of l in the fusion tree. In the first case, since $f - f_l$ is equal to $f - (EH_s - EH_l)$ on the support of f_s, the flat zones are unchanged; in the second case, $f - f_l$ is equal to EH_l, so that each flat zone of f is still included in a — possibly larger — flat zone of $f - f_l$. Applying an operator $T^{c,\theta}_\varepsilon$ is equivalent to successive removals, and the composition of connected operators is a connected operator, which proves the proposition. □

Property 4. *Decreasingness*
For all c, θ, $(T^{c,\theta}_\varepsilon)_\varepsilon$ is a decreasing family of antiextensive operators.

Since for any choice of decomposition criterion c and attribute θ, we obtain a decreasing family of connected, antiextensive operators, any couple (c, θ) defines an extinction value by definition 3. Some are already well-known: the couple (d, δ) defines the dynamics, the couple (a, α) the area extinction value and the couple (v, μ) the volumic extinction value. It must be noted that although the extinction values are the same, the families of operators are not.

The other couples define new kinds of extinction values: we already mentioned the L^1 extinction value, naturally defined by the couple (L^1, λ); the others can be thought of as second-order extinction values. All operators and their respective effects are illustrated in Sec. 3.7.

3.6.3 Stable Thresholding Operators

As we have seen, the dynamics, area and L^1 decompositions are stable by leaf removal; if c is one of these three criteria and θ is decreasing from leaf to root for the corresponding decomposition, applying $T^{c,\theta}_\varepsilon$ is equivalent to successive leaf removals and the decomposition of $T^{c,\theta}_\varepsilon(f)$ is simply $T^{c,\theta}_\varepsilon(f) = \sum_{l|\theta(l)\geqslant\varepsilon} f^c_l$.

The five thresholding operator families which satisfy both conditions are $T^{d,\delta}$, $T^{d,\alpha}$, $T^{a,\alpha}$, $T^{L^1,\lambda}$ and $T^{L^1,\alpha}$. For these five families, the following properties are immediate consequences of the decomposition's stability by these operators:

Property 5. *Idempotence*

$\forall \varepsilon \geq 0,\ T^{c,\theta}_{\varepsilon} \circ T^{c,\theta}_{\varepsilon} = T^{c,\theta}_{\varepsilon}$

Property 6. *Absorption property*
$\forall a$ *and* $b \geq 0,\ T^{c,\theta}_{a} \circ T^{c,\theta}_{b} = T^{c,\theta}_{a \vee b}$

This last property is remindful of granulometries: applying two or more operators is equivalent to applying the most active one. However, even the stable thresholding operators of this section are not morphological openings: they are anti-extensive and idempotent, but they are not increasing: for each operator $T^{c,\theta}_{\varepsilon}$, it is possible to find f and g with $f \leq g$ such that the inequality $T^{c,\theta}_{\varepsilon}(f) \leq T^{c,\theta}_{\varepsilon}(g)$ does not hold.

Operators belonging to these five families are, however, levelings [Mey04]. Levelings are a particular case of connected operators, introduced in [Mey98].

Definition 7. *Leveling*
A function g is a leveling of f if for any pair (p, q) of neighboring pixels:

$$g(p) < g(q) \Rightarrow f(p) \leq g(p) \text{ and } g(q) \leq f(q)$$

Property 7. *Stable thresholding operators are levelings*
For the five stable thresholding families of operators $T^{c,\theta}$, for all $\varepsilon \geq 0$, for all $f : X \to \mathbb{R}^+$, $T^{c,\theta}_{\varepsilon}(f)$ is a leveling of f.

Proof. Applying $T^{c,\theta}_{\varepsilon}$ is equivalent to successive leaf removals: if a component is removed by $T^{c,\theta}_{\varepsilon}$, so are all its descendants in the maximum fusion tree. Let f be a function from X to $\mathbb{R}^+$ and $f = \sum_l f^c_l$ its decomposition with respect to criterion c. Let us consider the list of labels $m_1, \ldots, m_n$ which are erased by $T^{c,\theta}_{\varepsilon}$ but whose father component in the maximum fusion tree is not (or which have no father in the maximum fusion tree). If we write $g = T^{c,\theta}_{\varepsilon}(f)$, we have $f = g$ everywhere except on the supports of the f_{m_i}, which are the connected components of the set $\{x | f(x) \neq g(x)\}$. For each connected component of this set, g is constant and equal to EH_{m_i}. Thus, for neighboring pixels p and q, $g(p) \neq g(q)$ implies that $g(p) = f(p)$ and $g(q) = f(q)$, which in particular proves that g is a leveling of f. □

3.7 Example Illustrations

In this section, we illustrate the concepts defined above on a natural image $f : X \to \{0, \ldots, 255\}$ shown in Fig. 3.16. Here X is the set of pixels with 8-connectivity.

There are 14420 regional maxima on this image; after label merging, there are 14156 peaks in the dynamics decomposition, 13785 peaks in the area decomposition, and 14306 peaks in both the L^1 and volume decompositions.

The root component of the dynamics decomposition is shown in Fig. 3.17. There are several global maxima in the image, all of them located in the patches of snow at the top of the image.

The root of the area, L^1 and volume decompositions (Fig. 3.18) is in this case the same, since the 'greatest' maximum is the same in all three senses: a point on the left shoulder of the shirt at the bottom left.

In both cases, the root component is the geodesic reconstruction of the greatest maximum or maxima under the image.

3.7.1 Thresholdings on the δ attribute

In this subsection, we focus on the operators $T^{c,\delta}_{\varepsilon}$: in other words, for each decomposition, a peak f_l is kept only if $\max(f_l)$ is above the threshold ε. As mentioned above, for the dynamics decomposition,

Figure 3.16: Example image used to illustrate the decompositions and operators defined in this work. The image is 640 by 480 pixels, 8-bit grayscale (256 levels from 0 to 255).

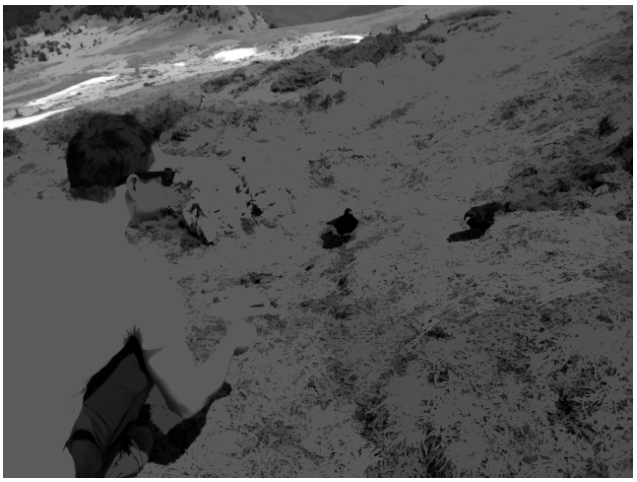


Figure 3.17: Root component of the dynamics decomposition

Figure 3.18: Root component of the area, L^1 and volume decompositions

(a) Illustration of $T_{100}^{d,\delta}$.

(b) Illustration of $T_{100}^{a,\delta}$.

(c) Illustration of $T_{100}^{L^1,\delta}$.

(d) Illustration of $T_{100}^{v,\delta}$.

Figure 3.19: Comparison of the four operators $T_{\varepsilon}^{c,\delta}$ applied to the original image with $\varepsilon = 100$.

$T_\varepsilon^{d,\delta}$ erases peaks with low prominence — dynamics — and keeps the most prominent ones. Contrary to h-reconstructions, which perform a similar task, peaks that are kept are unaltered.

In the other decompositions, peaks that are very high above the nearest 'col' will be kept, too, no matter their other attributes: more formally, if M is the only maximum in its connected component of the upper level set $X^+_{f(M)-\varepsilon}(f)$, then by construction, since M does not meet any other label until a lower level, $\max(f_M) > \varepsilon$, so f_M is preserved by the operator $T_\varepsilon^{c,\delta}$. It is not a necessary condition: a maximum M is kept as long as it is the 'greatest' — according to the decomposition criterion — of its connected component of $X^+_{f(M)-\varepsilon}(f)$ (which does, however, require that $f(M) \geq \varepsilon$). For the area decomposition, for instance, a lower peak surrounded by higher ones may be kept if it is broad enough. In the case of the L^1 or volume decomposition, lower peaks may be kept if they are 'massive' enough; the difference between the two criteria being that in the case of the volume, the 'mass' (L^1 norm) of the descendants in the fusion tree is taken into account, whereas it is not in the L^1 decomposition. In practice, unsurprisingly, these two decompositions and their associated operators are often very similar; nevertheless, as detailed in the previous section, they have different properties.

The resulting image after applying the four operators $T_\varepsilon^{d,\delta}$, $T_\varepsilon^{a,\delta}$, $T_\varepsilon^{L^1,\delta}$ and $T_\varepsilon^{v,\delta}$ are illustrated in Fig. 3.19, with $\varepsilon = 100$.

On these illustrations, we can see that $T_{100}^{d,\delta}$ leaves intact the most contrasted parts (peaks associated to the most prominent maxima). As compared to the other δ-thresholding operators, only the bright fold at the bottom of the t-shirt is kept intact, because it is brighter than the rest. In all other three cases, the rest of the shirt is almost entirely kept, but not the fold at the bottom. Operators $T^{a,\delta}$, $T^{L^1,\delta}$ and $T^{v,\delta}$ are harder to interpret: as mentioned above, very prominent maxima will be kept, independently of their 'importance' with respect to the decomposition criterion. On the other hand, even an 'important' maximum, whose label will propagate downward for a long time, can be associated a peak with relatively low height. This can be seen on the synthetic example of Sec. 3.4: for the area, L^1 and volume decompositions, the maximum value of the root component is lower than the maximum value of f_{M_1} (strictly, in the L^1 and volume cases).

Nevertheless, as can be seen in Fig. 3.19, these operators do preserve several regions that were erased by $T_{100}^{d,\delta}$: this is particularly visible on the shirt and on the rocks in the middle of the image, as well as in the grass at the bottom.

Because the L^1 and volume decompositions are very similar, the results of $T_{100}^{L^1,\delta}$ and $T_{100}^{v,\delta}$ look the same; the images are actually not exactly identical, but they differ only on 4821 pixels (a little more than 1.5% of the image), and they differ at most by 4 grey levels. The most noticeable difference between those two operators and $T_{100}^{a,\delta}$ is the face, which is erased by $T_{100}^{a,\delta}$ but kept intact by $T_{100}^{L^1,\delta}$ and $T_{100}^{v,\delta}$.

3.7.2 Thresholdings on the α attribute

In this subsection, we illustrate the operators $T_\varepsilon^{c,\alpha}$: peaks are kept if their support is large enough. In particular, $T_\varepsilon^{a,\alpha}$ is comparable to the area opening γ_ε^a, as they make the same maxima disappear, the difference being that the remaining maxima — those with an area extinction value greater than ε — are unaltered by $T_\varepsilon^{a,\alpha}$, whereas they are 'flattened' by γ_ε^a.

The other α thresholding operators keep maxima that are locally the greatest, in the sense that a maximum m and its associated peak are left untouched if and only if there exists a level y such that m is the greatest maximum in its connected component of $X_y^+(f)$, and this connected component has a cardinality of at least ε.

As can be seen in Fig. 3.20, $T_{700}^{d,\alpha}$ preserves several bright spots that the other operators erase, notably snow patches in the background, but also several small dots in the grass.

The other three operators look more similar. A difference can however be seen on the face: the cheekbone is lighter in $T_{700}^{a,\alpha}$ than in the other three cases, whereas the earlobe is kept intact by all

(a) Illustration of $T_{700}^{d,\alpha}$.

(b) Illustration of $T_{700}^{a,\alpha}$.

(c) Illustration of $T_{700}^{L^1,\alpha}$.

(d) Illustration of $T_{700}^{v,\alpha}$.

Figure 3.20: Comparison of the four operators $T_{\varepsilon}^{c,\alpha}$ applied to the original image with $\varepsilon = 700$.

operators except $T^{a,\alpha}_{700}$.

The L^1 and volume cases are again quite similar — they actually differ on only 1816 pixels — but the absolute difference reaches 78. Most differences are hardly visible, but a lighter patch can be seen in $T^{v,\alpha}_{700}$ in the trees in the very background on the top left.

3.7.3 Thresholdings on the λ attribute

Like the volumic extinction value, which was historically introduced as a compromise between area and dynamics, the λ attribute (L^1 norm of a peak) can be seen as a tradeoff between height and support cardinality of a peak. A maximum with high dynamics can be associated a peak with a great δ value but quite narrow, resulting in a smaller λ value than a lower but larger maximum. This can be seen by comparing operators $T^{d,\delta}_{100}$ (Fig. 3.19a) and $T^{d,\lambda}_{8000}$ (Fig. 3.21a): several quite small snow patches on the top, as well as the rightmost Alpine chough's beak, are preserved by the former but erased by the latter; on the contrary, the shirt, forearm and rock above the arm have a lower dynamics but are wide enough to be kept by $T^{d,\lambda}_{8000}$.

A similar statement can be made about the area decomposition: a maximum with a great area extinction value can be associated a peak with a large α value but relatively low, in which case its λ value will be quite small; on the other hand, a maximum with a smaller area extinction value but more prominent can have a larger λ value. In Fig. 3.21a, we can see, for instance, that several snow patches in the background that were erased by $T^{a,\alpha}_{700}$ are preserved by $T^{a,\lambda}_{8000}$. There are actually also 10 peaks that are preserved by $T^{a,\alpha}_{700}$ but erased by $T^{a,\lambda}_{8000}$, but they are less visible because of their low dynamics, although several small differences can be noticed in the grass.

On this image, the result of $T^{L^1,\lambda}_{8000}$ looks quite close to the result of $T^{a,\lambda}_{8000}$, although there are visible small differences, for instance on the rock or in the grass. The reconstruction of the face is also closer to that of $T^{d,\lambda}_{8000}$.

The results of $T^{L^1,\lambda}_{8000}$ and $T^{v,\lambda}_{8000}$ are visually very similar, except for the trees in the background. A small difference can be seen in the grass below the rightmost Alpine chough as well.

3.7.4 Thresholdings on the μ attribute

Since for a peak f^c_l, we have $\mu_c(l) = \lambda_c(l) + \sum_{d \in D(l)} \lambda_c(d)$, thresholding on the μ attribute yields a less active operator than the same thresholding on the λ attribute; for all decompositions:

$$T^{c,\mu}_{\varepsilon}(f) \geq T^{c,\lambda}_{\varepsilon}(f)$$

This can be seen when comparing Figs 3.21 and 3.22: some details are kept by the μ thresholding operators, while erased by the λ thresholding ones.

3.8 Extension to real-valued functions

3.8.1 Decompositions

All previous ideas can be easily extended to real-valued functions by considering their positive and negative parts. A function f can be written as $f = f^+ - f^-$, with $f^+ = f \wedge 0$ and $f^- = (-f) \wedge 0$. Both f^+ and f^- take only nonnegative values, so we can perform their respective decompositions according to some criterion, just as before; we now consider and quantify the positive maxima and negative minima of f.

With $f^+ = \sum_k f^{+c}_k$ and $f^- = \sum_l f^{-c}_l$ the respective decompositions of f^+ and f^- according to criterion c, the decomposition of f is given by:

(a) Illustration of $T_{8000}^{d,\lambda}$.

(b) Illustration of $T_{8000}^{a,\lambda}$.

(c) Illustration of $T_{8000}^{L^1,\lambda}$.

(d) Illustration of $T_{8000}^{v,\lambda}$.

Figure 3.21: Comparison of the four operators $T_{\varepsilon}^{c,\lambda}$ applied to the original image with $\varepsilon = 8000$.

(a) Illustration of $T_{8000}^{d,\mu}$.

(b) Illustration of $T_{8000}^{a,\mu}$.

(c) Illustration of $T_{8000}^{L^1,\mu}$.

(d) Illustration of $T_{8000}^{v,\mu}$.

Figure 3.22: Comparison of the four operators $T_{\varepsilon}^{c,\mu}$ applied to the original image with $\varepsilon = 8000$.

$$f = \sum_k f_k^{+c} - \sum_l f_l^{-c}$$

3.8.2 Self-dual Operators

The definition of the operators $T_\varepsilon^{c,\theta}$ is easily extended to real-valued functions by applying them (with their former definition) to f^+ and f^-:

Definition 8. *Generalized thresholding operators*

$$T_\varepsilon^{c,\theta}(f) = \sum_{k|\theta(k)\geqslant\varepsilon} f_k^{+c} - \sum_{l|\theta(l)\geqslant\varepsilon} f_l^{-c}$$

These operators are self-dual in the sense that applying them to $-f$ is the same as applying them to f, then taking the opposite:

Property 8. *Self duality*
For any (c,θ), for all $\varepsilon > 0$, $T_\varepsilon^{c,\theta}(-f) = -T_\varepsilon^{c,\theta}(f)$.

We obviously lose the antiextensivity property, but the operators are still connected. The operators belonging to the five stable families remain stable: they are still idempotent, and satisfy the absorption property $\forall a$ and $b \geq 0$, $T_a^{c,\theta} \circ T_b^{c,\theta} = T_{a\vee b}^{c,\theta}$.

These families of operators also still define levelings; the only difference in the proof is that we must take into account the case where $g(p) < 0 < g(q)$, in which case we immediately have $f(p) \leq g(p)$ and $g(q) \leq f(q)$.

In order to illustrate this type of operators, let us consider the example image f of Fig. 3.16, which takes values between 0 and 255, and define $g = f - 127.5$, which then takes both negative and nonnegative values. Applying $T_{1000}^{a,\alpha}$ to g preserves both salient maxima (bright structures) and salient minima (dark structures), as shown in Fig. 3.23; on the other hand, applying an area opening of size 1000 to both g^+ and g^- yields the image shown in Fig. 3.24: the snow patches on top are almost filtered out, as well as the lefmost Alpine chough.

Figure 3.23: Result of the self-dual operator $T_{1000}^{a,\alpha}$ applied to the centered image $g = f - 127.5$.

Figure 3.24: Result of applying an area opening of size 1000 to both g^+ and g^-.

3.9 Conclusion and Perspectives

In this article, we present several ways to associate peaks to the maxima of a function, yielding a decomposition of the function as a sum of elementary components. From these decompositions one can attribute several importance measures to each maximum; particular rules of partial summations over these components also define new morphological operators, the properties of which are detailed. Although defined at first from the maxima of a function taking nonnegative values, the decompositions and operators are generalized to real-valued functions, the peaks and trenches corresponding, respectively, to positive maxima and negative minima.

The operators presented in this work do not alter the image in the vicinity of the maxima they preserve; in particular, the families $T^{d,\delta}$, $T^{a,\alpha}$ and $T^{v,\mu}$ define the same extinction values but are less active than their respective counterparts, h-reconstructions, area openings and volumic razings (see Prop. 2). While removing small details, they leave intact the most salient structures (see Fig.s 3.23 and 3.24, as well as [Ala+17b]).

Although this work is mostly theoretical, the concepts and operators introduced offer new possibilities for image simplification. Since the operators introduced here are connected — they do not create new contours — they are particularly well-suited for image segmentation. Figure 3.25 shows two segmentations of the example image, obtained by the watershed algorithm applied to the morphological gradient of Fig. 3.23 (α thresholding of the dual area decomposition) and Fig. 3.24 (area opening of g^+ and g^-). In both cases, the 30 minima of the gradient image with the largest area extinction value were used as markers.

There is no 'better' segmentation in an absolute sense; depending on the application, one can be better than the other, but Fig. 3.25 illustrates that the operators introduced in this work offer new alternatives.

(a) Watershed segmentation using the gradient of $T^{a,\alpha}_{1000}(f - 127.5)$

(b) Watershed segmentation using the gradient of $\gamma^{a}_{1000}(g^{+}) - \gamma^{a}_{1000}(g^{-})$

Figure 3.25: Two segmentations of the original image, obtained by the watershed algorithm using the 30 minima with the largest area extinction value of the gradient of (a) $T^{a,\alpha}_{1000}(g)$ and (b) $\gamma^{a}_{1000}(g^{+}) - \gamma^{a}_{1000}(g^{-})$.

Part II

Retinoptic

Chapter 4

Introduction

Ce chapitre présente brièvement la rétinopathie diabétique et les enjeux de santé publique associés, puis présente le contexte du projet RetinOptic. Nous exposons ensuite les raisons ayant motivé le choix de détecter la macula dans les images rétiniennes, en particulier le rôle de la visibilité de la macula comme estimateur de qualité des images.

4.1 Diabetic Retinopathy

It is estimated that there were 415 million people with diabetes in 2015, and this number is expected to reach 642 million by 2040 [Ogu+17]. A common complication of diabetes is diabetic retinopathy (DR) [EZ16], which is one of the main causes of blindness and visual loss [Sjø+97; Mat+04]. Due to the heterogeneity of protocols between different epidemiology studies, it is hard to give a precise prevalence of DR; however, the percentage of diabetic patients found with DR in recent studies is relatively stable and ranges from 21.9% to 36.8% [DMR09].

DR is detectable and treatable, but regular clinical examination of all patients diagnosed with diabetes is infeasible in practice; in many developing countries, there is a significant lack of ophthalmologists, and in developed countries, the number of people aged 60+ is growing at twice the rate of the profession [Res+12].

In most clinically significant cases, DR is detectable on eye fundus photographs. Telemedicine networks [CC03; Bou+08; Mas+08; TWN15] have been created in various countries in order to perform mass screening; international and local guidelines recommend one fundoscopic examination per year for diabetic patients [Ame12; Mas+08]. Photographs can be taken by technicians in hospitals, specific screening centers, pharmacies or even prisons equipped with mydriatic cameras. In recent years, portable retinographs have been developed, which allow for even more massive screening. Photographs are then sent to ophthalmologists, who grade them and indicate the course of action to be followed. Both patient and practician time can be saved this way, provided that images are of good enough quality.

Due to the increasing amount of data, in conjunction with the limited number of ophthalmologists, computer retinal image understanding is of utmost interest. The literature concerning eye fundus image processing is abundant (a non-exhaustive review can be found in [SG15a]), including many segmentation methods to extract anatomical structures such as the optic disk, the macula or the vascular network, or pathological structures [Agu+14; NAG09; Ver+13; WSM11; Gup+14; Zha+14], as well as automatic predictions of DR severity [Que+17; Xia+17; Gar+96].

4.2 The OPHDIAT Telemedicine Network and OPHDIAT Database

OPHDIAT (for OPHtalmologie DIAbète Télémedecine) [Mas+08] is a telemedicine network created in 2004 by Assistance Publique - Hôpitaux de Paris (APHP) in order to perform DR screening by eye fundus photography in Ile-de-France. In 2018, 17063 examinations were performed on 43 different sites.

An examination in the OPHDIAT database typically consists in two images per eye: one centered on the macula (central image), the other one centered on the optic disk (nasal image). The examination includes some information about the patient: age, sex, diabetes type, current treatment, date of diabetes diagnosis; it also includes the model of the retinograph, the center in which the examination took place and the technician who performed it. All personal data is anonymized.

Each examination is then sent to an ophthalmologist, who performs the diagnosis. He or she provides information about image quality, DR severity, the course of action to be taken, and a short conclusion text. An example is given in Fig. 4.1.

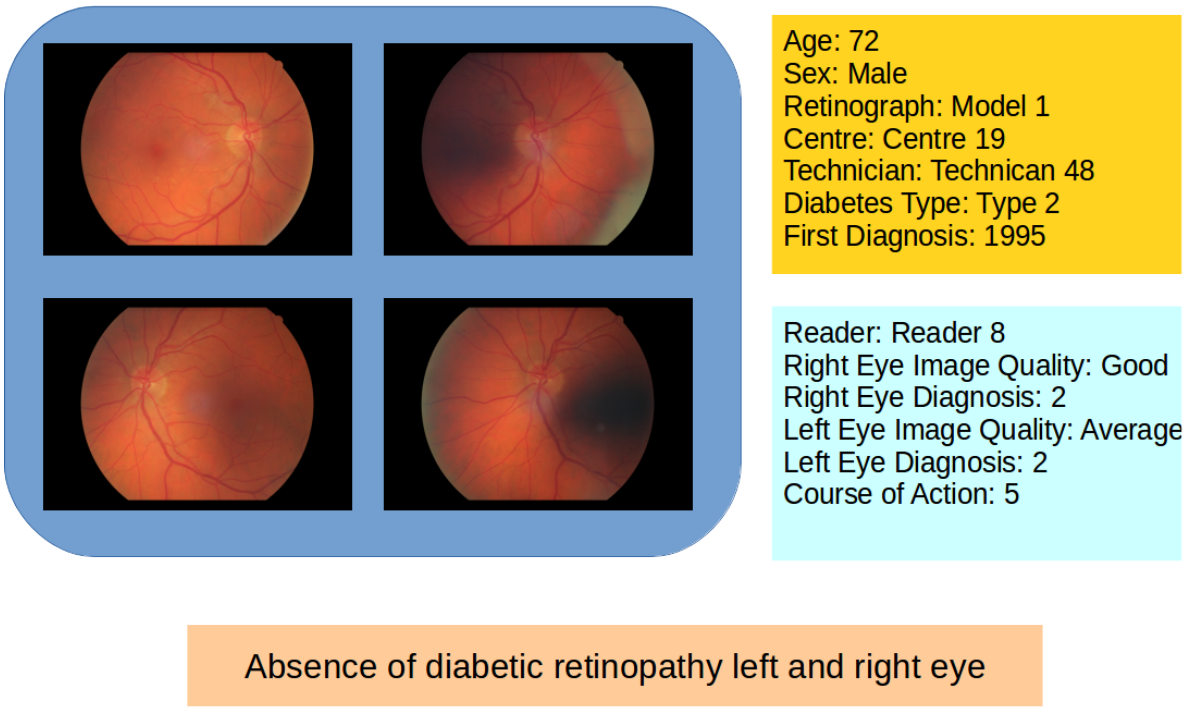


Figure 4.1: Typical examination, containing two images per eye, miscellaneous information and detailed diagnosis

Diagnosis per eye and course of action are coded on the scales summarized in Tables 4.1 and 4.2. For the diagnosis of an eye, code 1 is used when the pictures are uninterpretable, in which case the course of action is almost always 7: the patient must be referred to an ophthalmologist with no emergency (within two months). In special cases, like pregnancy, the course of action can be different and the patient is often referred to an ophtalmologist or asked to take pictures again within a shorter time span. Code 1 for course of action does not appear in any of the 25,702 examinations of the OPHDIAT database.

Code	Meaning
1	?
2	No DR
3	Mild NPDR
4	Moderate NPDR
5	Severe NPDR
6	PDR
7	High Risk PDR

Table 4.1: DR diagnosis code summary. NPDR stands for Nonproliferative Diabetic Retinopathy; PDR stands for Proliferative Diabetic Retinopathy.

Code	Meaning
1	?
2	Eye fundus photographs in 1 month
3	Eye fundus photographs in 3 months
4	Eye fundus photographs in 6 months
5	Eye fundus photographs in 1 year
6	Patient to be referred to an ophthalmologist without emergency (4 months)
7	Patient to be referred to an ophthalmologist without emergency (2 months)
8	Patient to be referred to an ophthalmologist quickly (1 month)
9	Patient to be referred to an ophthalmologist urgently (15 days)
10	DR already known; patient should not be part of the screening

Table 4.2: Course of action code summary.

4.3 The Retinoptic Project

The Retinoptic project is a collaboration between academic and industrial actors, supported by a French "Fonds Unique Interministériel" Grant, in relation with the competitive clusters Systematic and Medicen. The aims of the project included the conception and commercialization of a new portable retinograph, a medical information online platform, and new image processing algorithms.

In this work, we focus on image quality assessment algorithms; about 10% of the examinations of the OPHDIAT network are deemed uninterpretable by ophthalmologists. Some of these uninterpretable examinations are unavoidable, typically when the patient has cataract, or if the pupil is not dilated enough; however, in many cases, the issue is not medical, and it would have been possible to have an interpretable examination by taking another picture. Figure 4.2 illustrates this: both images are blurry, and thus uninterpretable, but it is likely that an image of better quality could have been obtained.

4.4 Image Quality Estimation

Image quality estimation is a necessary preliminary step to most retinal imaging understanding tasks, since it would make little sense applying an automatic diagnosis algorithm to images too noisy, blurred or not contrasted enough. Most publicly available datasets, like the Kaggle Diabetic Retinopathy dataset or the Messidor database [Dec+14], contain only gradable images, and in [NAG09], it is clearly mentioned that for building a local database, "acceptable image quality, as judged by the screening program ophthalmologists, was a selection criterion". A notable exception to this rule is the ARIA database [Dam06], which in particular contains precise annotations of the fovea and macula,

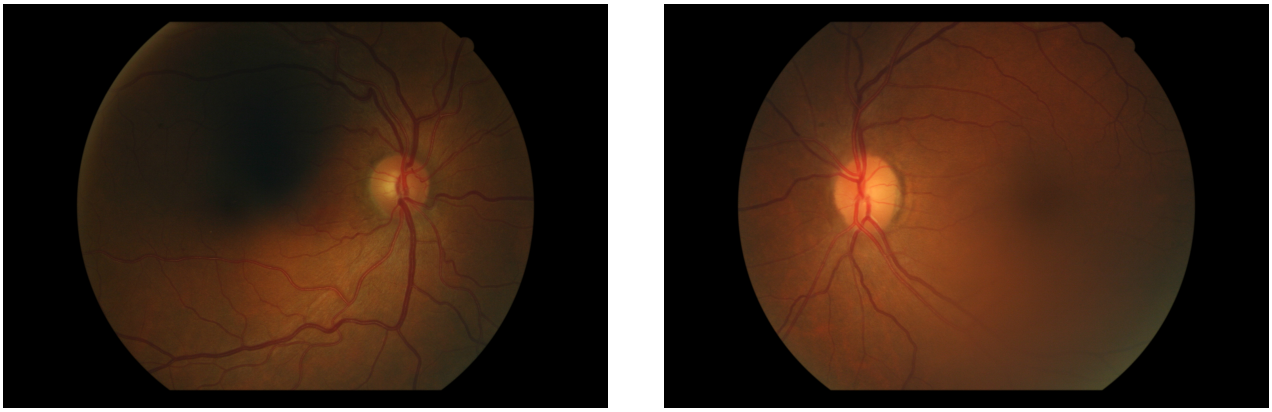


Figure 4.2: Both acquisitions are too noisy to be interpreted, but the issue seems to come from the acquisition, and not from a medical issue like cataract.

even on images where they are hardly — if at all — visible, but their positions can be guessed, using the rest of the image.

In the context of a telemedicine network, the diagnosis is performed by a human expert, but the photographs are taken at a different time and location, by operators whose skill and level of experience can vary. A significant portion of images - around 10% for the OPHDIAT network [Mas+08] - are deemed uninterpretable by ophthalmologists. This could be prevented by automatically estimating the quality at acquisition time, sending a warning to the operator if the photograph should be re-taken. In particular, since an essential requirement is that the region of the macula must be clearly visible for a diagnosis to be made, macula visibility is a relevant criterion for retinal image quality estimation.

4.5 Macula Segmentation

The literature concerning eye fundus image processing is abundant (see for instance [SG15b] for a non-exhaustive review); many segmentation methods have been proposed to extract anatomical structures such as the optic disc, the macula or the vascular network [Agu+14; NAG09; Ver+13; WSM11].

In the present work, we focus on the macula, which is located in the center of the retina, and is responsible for high-resolution, color vision in good light. The fovea, located in its center, is the region with the highest concentration of cone cells. Lesions in the macula impair central vision, and should be detected as soon as possible. When exudates are present on the retina, their distance to the macula determines whether the appropriate treatment is injections or laser therapy.

On eye fundus images, the macula appears as a dark region with low-contrasted borders, that contains no vessels. For hemorrhage detection algorithms [Gup+14; KGU13; Tan+13; Zha14], this often results in a false positive; it is preferrable to first locate the macula, and process it separately.

For all these reasons, it is clear that an accurate macula segmentation algorithm would be of great interest in an automated or semi-automated diagnosis framework. Several designs have been proposed (see [Ver+13] for a review), most of which are loosely based on the same idea: the macula is a reddish region, darker than its neighborhood, and its distance from the optic disc is roughly 2 or 3 times the diameter of said optic disc. Thus, the first step of these methods actually consists in detecting the optic disc, which is a relatively easier task.

In recent years, however, neural networks have been proven to be a very efficient tool for general segmentation; in particular, convolutional neural networks yield impressive results for image classification, outperforming other methods in many complicated computer vision tasks, including a recent Kaggle competition on DR screening [Gra15].

4.6 Outline of this part

The aim of this part is to provide tools for detecting the macula and assessing image quality around it on the images of the OPHDIAT network. Additionally, since our algorithms are to be deployed on embedded systems, we also focus on keeping our solutions as fast and lightweight as possible.

Following the idea of the Kaggle competition for classifying retinal images according to the severity of DR, we first train neural network classifiers for the seemingly simpler task of predicting whether or not the macula is visible or not, in Chapter 5. Although this might comparatively seem like a trivial task, it must be kept in mind that the databases are drastically different: the Kaggle Diabetic Retinopathy databases contains only central images of good enough quality, while ours is a clinical database, containing both nasal and central images, with absolutely no guarantee on their quality.

Assuming we dispose of a good classifier for assessing macula visibility, we can then consider only the images where it is deemed visible, and try to locate it. For this task, we trained regression neural networks, the output of which being the macula's x-y coordinates. We considered two possible kinds of input: either the green channel of the original image, or a morphological decomposition of it, designed so as to potentially compensate illumination artifacts, thus possibly 'helping' the network. We also trained networks with a similar but deeper architecture and compare the results with the shallower networks. Although the database is not the same, as we annotated more and more images during the process, we can safely conclude that, at least for this task, deeper networks perform better. This is the object of Chapter 6.

We then investigate fully-convolutional networks: instead of predicting a score or coordinates like the networks of the previous chapters, they output a real-valued image. The ground-truth is provided as a binary disk centered around the fovea's annotation when annotated, and as a zero-valued image when not. The output of the networks is never, in practice, a binary disk, but a simple post-processing enables us to determine whether the macula is visible or not, and to accurately locate it when it is. Simple features of the network's output image can also be used as quality scores. The work described in this chapter was published in [Ala+20]; Chapter 7 is an extended version of this article.

Chapter 5

Macula Visibility Assessment by Classification Neural Networks

Dans ce premier chapitre consacré aux réseaux de convolution, nous cherchons à répondre à la question "la macula est-elle visible ?" sur des images rétiniennes. Nous détaillons la création de notre première base de données, l'architecture choisie pour les réseaux de classification, et présentons les premiers résultats obtenus. Ayant obtenu des résultats satisfaisants pour cette première tâche, se pose naturellement la question de la localisation de la macula lorsqu'elle est visible, qui est l'objet du chapitre suivant.

5.1 Problem presentation

A first question to answer was whether we could train a neural network for the seemingly simple task of determining whether or not the macula was visible on a retinal image. This may seem quite restrictive, but in some cases, it is possible to obtain a localization heat map from an image classifier. Occlusion sensitivity, as defined in [ZF14], is a way of asserting whether a convolutional neural network is truly identifying an object in an image, or rather using the surrounding context. The idea is to occult the object of interest in an image, then apply the model to this occulted image. If the object of interest is mainly responsible for its high classification score, the score will drop; if the model uses the context information rather than the object itself, it is likely that the score will not change much.

We can monitor the output of the model as we slide the mask over the image, and visualize it as a score heat map. In our case, if the network is actually able to detect the presence of the macula, its location will be given by the region where the score is the lowest on the occlusion heat map.

However, in this preliminary approach, the aim was mostly to investigate how our networks would perform, as well as which input image resolution would be required. Since our algorithms have to be able to run on embedded systems with an acceptable running time, a relatively modest network and a low input resolution are likely preferrable to huge networks requiring high-resolution inputs.

5.2 First Database

The macula was manually annotated by two different operators on 1800 eye fundus images: when visible, the center was labeled by clicking on it. We can see in Fig. 5.1 that the annotations may differ by a few pixels.

This first database actually suffered from several design flaws. It contains the images of the three databases e-ophtha HE, e-ophtha EX and e-ophtha MA, which are annotated databases, respectively for hemorrhage, exudate and microaneurysm detection. There are two sampling biases introduced

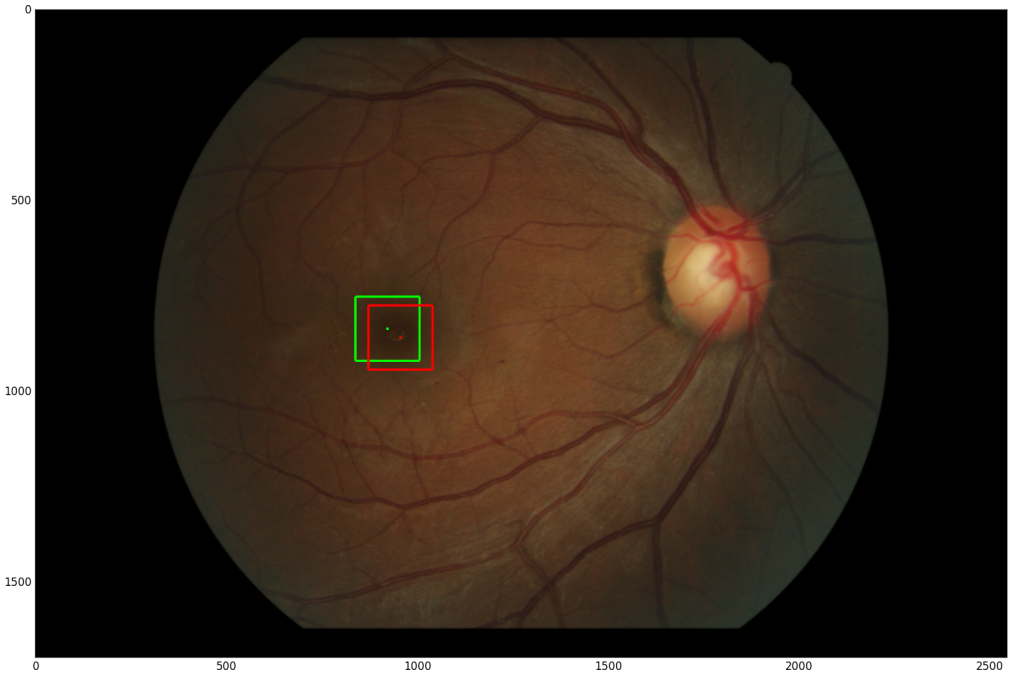


Figure 5.1: Labeled image with the two locations for the macula provided by the two operators

here: since each of these databases contains roughly half pathological images and half healthy ones, the proportion of pathological images is — thankfully — much greater than it is in a real-life application. The other issue is that these annotations were only made on good-quality images, which introduces a second sampling bias regarding image quality.

After the initial annotation of the macula, the two operators agreed on the macula's visibility on only 87% of the images, most of the disagreements happening on images where the macula was either only partially present or localizable but in a image of bad quality. An example of these disagreements can be seen in Fig. 5.3.

This actually raises an interesting question, in that the phrase 'localize the fovea', although seemingly simple, needs to be clarified in light of the pursued goal. For instance, in both images of Fig. 4.2, it is easy to deduce the macula's localization using the rest of the image, but the macula itself is hardly visible, if at all.

In the literature, some algorithms exist that aim at localizing structure, including the macula/fovea, no matter how poor the image quality is. The ARIA database provides pixel-precise annotations for the fovea and macula on several retinal images, including the one depicted in Fig. 5.2. At this point, it is highly debatable whether it is sensible to evaluate the performance of a localization algorithm on this kind of image.

In order to reduce the gap between the two operators, it was decided that the macula should be annotated only if it was entirely visible and if the corresponding zone of the image was of good enough quality to distinguish it as well as the small blood vessels surrounding it. After looking a second time at all images for which there was a disagreement, it was decided which annotation to keep.

The original images come from different retinographs and have various resolutions: in order to harmonize the data, each image was cropped around the region of interest, and rescaled. The new resolution must be relatively small, in order for the network not to be too large, but not small to the point the visibility of the macula cannot be assessed. In first attempts, we used 256x256, and later 128x128 images. Only the green channel of the images was used.

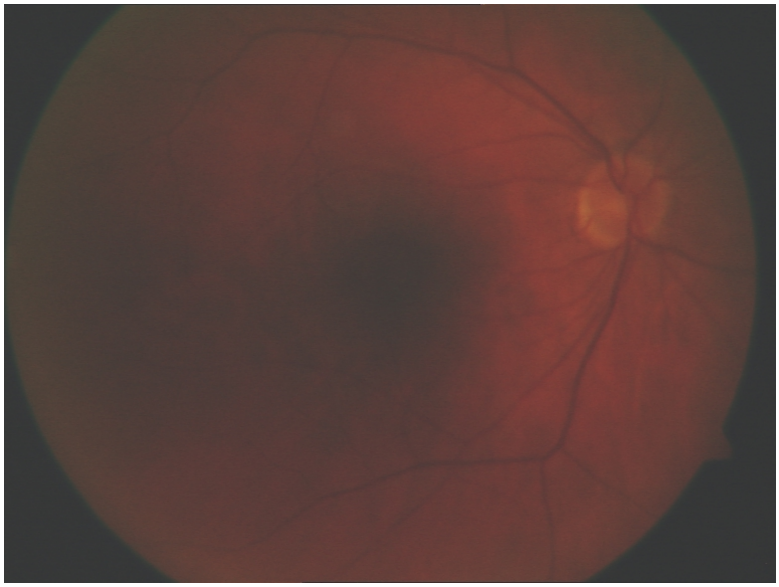


Figure 5.2: Sample image from the ARIA database. Despite its poor quality, a pixel-wise annotation for the macula and fovea are provided.

Figure 5.3: Example of disagreement between the two operators: the macula can be seen but is in a dark zone of the image.

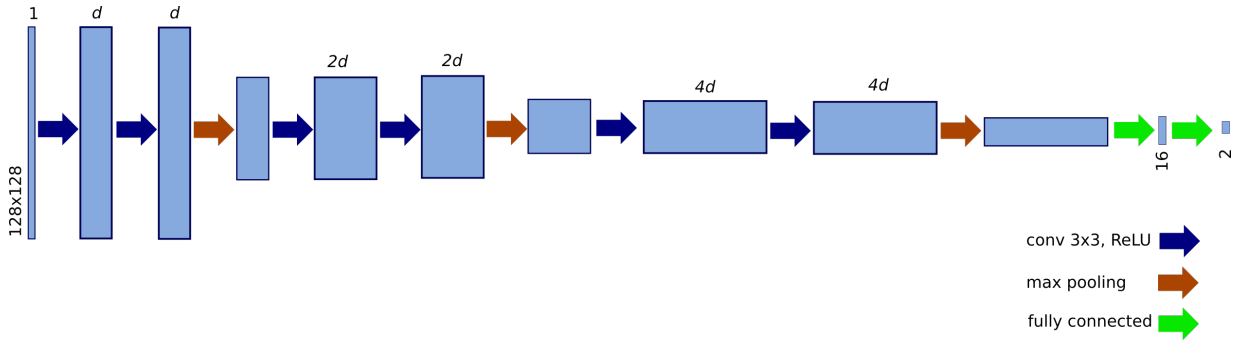


Figure 5.4: Network architecture

5.3 Network Architecture

The database was split in three parts: a training (80% of images), validation (10%) and test (10%) set. Images from a given examination were attributed to a same set, in order to avoid similar images in the training and validation, or training and test sets.

The networks we used consist in convolutional and max-pool layers at the beginning, ending with a fully-connected layer. Each convolutional layer is followed by a Rectified Linear Unit(ReLU). More precisely, for the networks with a 128x128 input, there are two 3x3 convolutional layers with d channels, followed by a 2x2 max-pooling, followed by two new 3x3 convolutional layers, this time with $2d$ channels, followed by a 2x2 max-pool layer and two new 3x3 convolutions, this time with $4d$ channels. The last convolutional block is then the input of a fully-connected layer of size 16, and the final output is a logistic unit. This architecture is summarized in Fig. 5.4. We also trained networks with 256x256 input images, with the same architecture except there is an additional [CONV-CONV-POOL] block; in this case, the last convolutional blocks have depth $8d$. In our experiments, we set $d = 8$. The networks were trained using the RMSProp [TH12] optimizer, and with a dropout [Sri+14] keep probability of 0.7.

5.4 Classification Results

For each architecture, the networks were trained several times, with different random initializations; we then selected the one with the best logistic loss on the validation set, which turned out to be one of the 128x128 input networks. On the test set, thresholding the logistic score at 0.5, we obtain 97.8% accuracy. Several classification results are presented below.

5.5 Conclusion

In this chapter, we saw that we are able to determine whether the macula is visible with great accuracy, with a relatively simple network, and after rescaling images to a modest 128x128 resolution. The necessity of providing a precise meaning to the phrase 'the macula is visible' has been assessed: our criterion is that the macula must be entirely within the image and that the image quality in its surrounding region must be good enough. When this is the case, it would be useful to predict the macula's location: if we consider that this is equivalent to predicting the fovea's coordinates, can we actually use the same network architecture for this task, simply by replacing the logistic classification loss by the L^2 regression loss? This is the object of the next chapter.

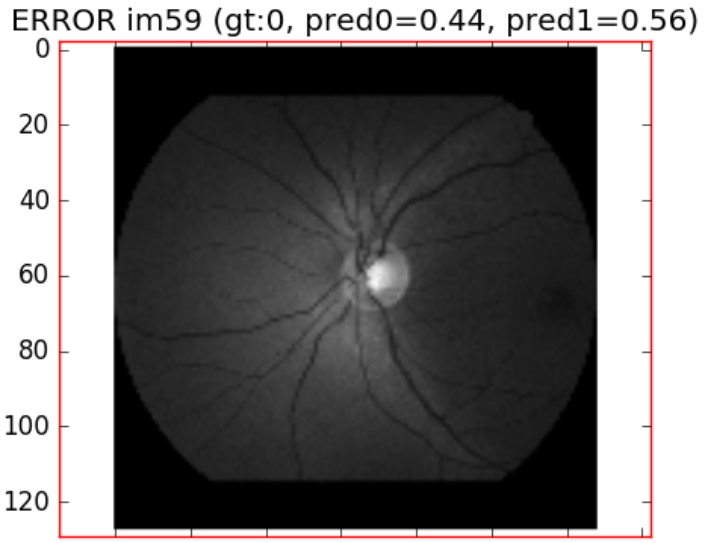


Figure 5.5: The macula was annotated as visible, but the networks predicts that it is not: the score is only 0.44, which falls below the 0.5 threshold.

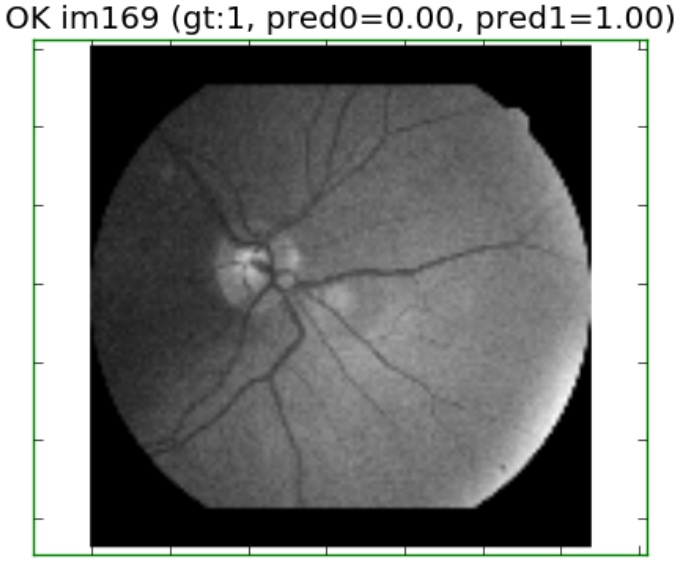


Figure 5.6: The image is correctly classified as 'macula not visible' (score: 0).

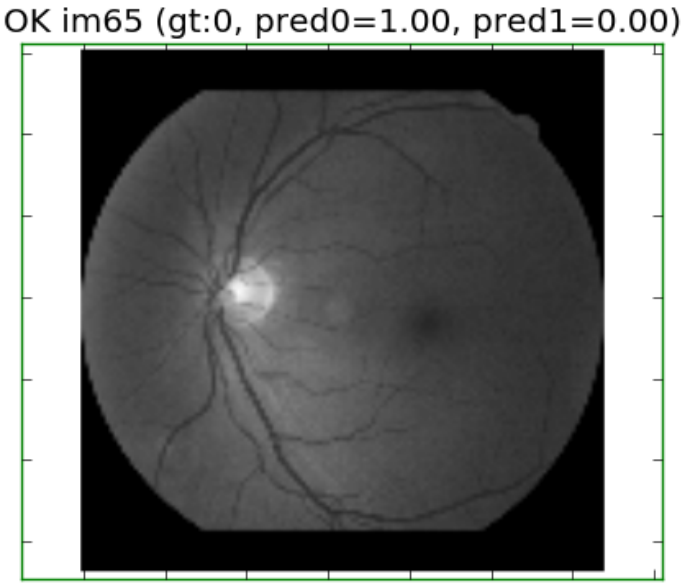


Figure 5.7: The macula is correctly predicted as 'macula visible' (score: 1).

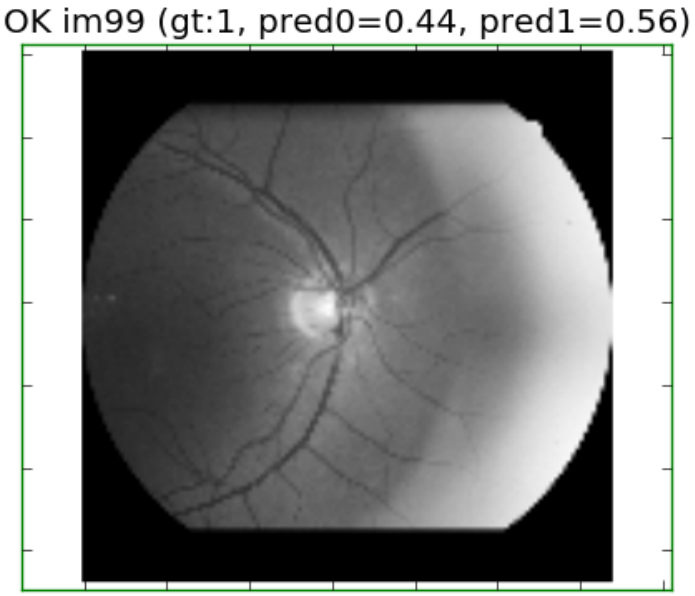


Figure 5.8: The image is correctly classified as 'macula visible' but with a relatively high visibility score (0.44).

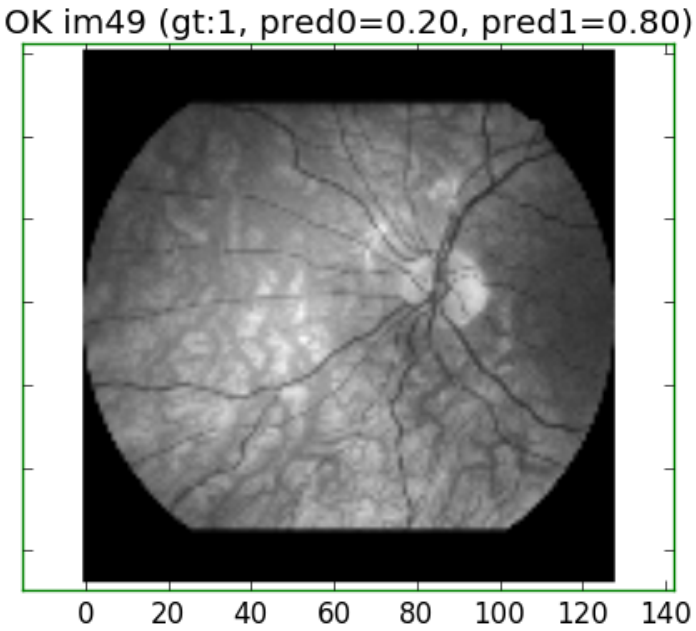


Figure 5.9: The image is correctly classified as 'macula not visible', but with a non-zero visibility score (0.2).

Chapter 6

Macula Localization by Regression Neural Networks

Dans ce chapitre, nous proposons de localiser la macula sur des images rétiniennes à l'aide de réseaux de convolution de régression. Ceci suppose que nous avons à disposition un classificateur afin de s'assurer au préalable que la macula est bien visible sur les images où nous cherchons à la localiser.

Nous envisageons deux types d'entrée possibles pour nos réseaux : le canal vert des images, ou une décomposition morphologique de celui-ci visant à corriger les artefacts d'illumination présents sur certaines images. Les résultats obtenus semblent montrer que ce pré-traitement n'améliore pas la performance de localisation. Par ailleurs, les premiers réseaux, peu profonds, ne semblent prédire correctement que la coordonnée horizontale. La normalisation de la sortie atténue cet effet mais dégrade la performance globale.

Des réseaux plus profonds donnent de meilleurs résultats, même avec moins de neurones dans la dernière couche complètement connectée (et donc moins de paramètres). Cependant, la limitation principale reste la difficulté à détecter les erreurs de localisation. Dans le chapitre suivant, nous investiguons une autre approche du problème, par des réseaux de segmentation, afin d'obtenir à la fois la classification (macula présente ou non), la localisation lorsque la macula est visible, ainsi qu'un score de confiance de la prédiction.

6.1 Introduction

Assuming we have a good classifier for assessing whether or not the macula is visible on retinal images, if we want to locate it, since it can be considered approximately constant in shape and size, the problem of macula segmentation can be reduced to that of fovea localization.

We have seen that with the architecture introduced in the precedent chapter, we are able to answer the first question, 'is the macula visible?', with great accuracy. Now, can the same architecture be used in order to answer the second one: what are the fovea's coordinates?

We also wanted to investigate the influence of the network's complexity on the results: in order to do so, we tried different values for the number of neurons n in the last, fully-connected layer of our network. In a subsequent step, we also investigate a deeper architecture.

In our database, we noticed that some images present large horizontal luminosity gradients, which, at least intuitively, could hinder the localization of a dark structure, like the macula. If we somehow correct this luminosity artifact, could this 'help' the network and improve its performance?

We saw in the previous chapter that our first database possibly suffered from some biases; we first annotated 2700 more images, thus hopefully obtaining a more accurate representation of reality. Before training the deeper networks on the second half of this chapter, we again annotated more images, for a total 6098.

6.2 Morphological Decomposition

As previously, a possible input for a learning algorithm is the resized green channel, as shown in Fig. 6.1; however, we also used a morphological decomposition in order to correct a possible non-uniform illumination, and to separate bright and dark objects.

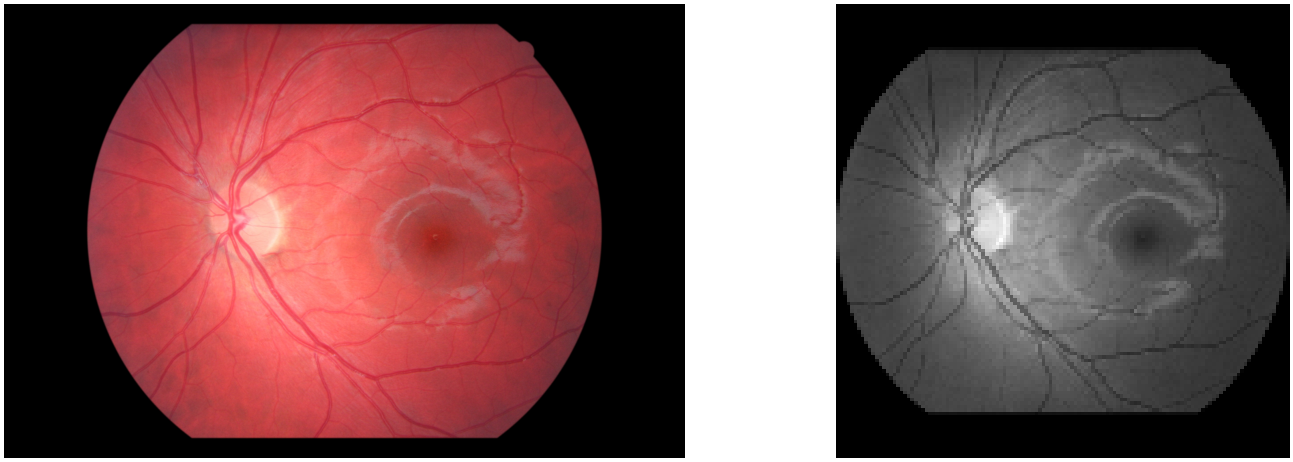


Figure 6.1: The original color image on the left has been cropped around the region of interest, zero-padded on top and bottom so as not to introduce distortion, then resized to 128x128.

The decomposition is obtained by applying hexagonal alternating sequential filters [SV92] to the green channel, the size of the largest hexagon being the approximate size of the optic disc. The resulting image is an estimation of the illumination effects; the positive and negative residues respectively contain the bright and the dark structures. This decomposition is illustrated in Fig. 6.2.

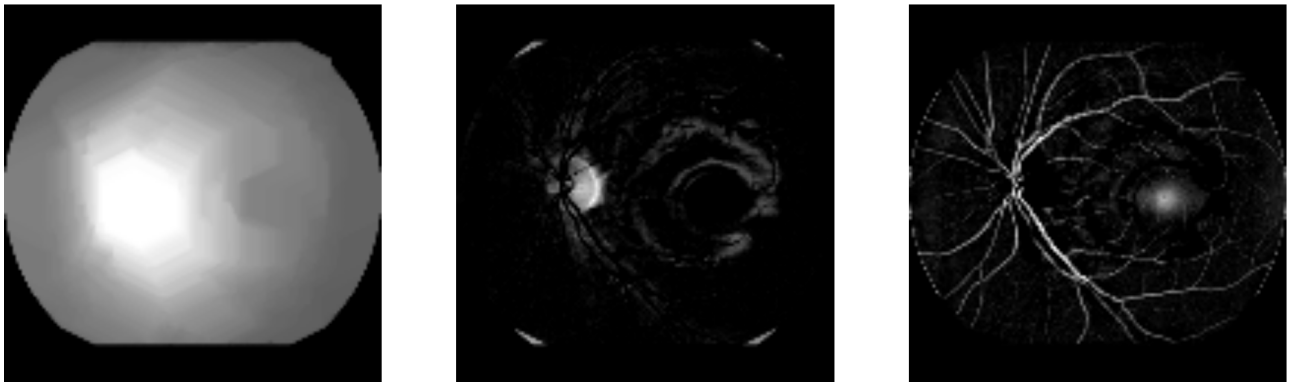


Figure 6.2: From left to right: result of the alternating sequential filters, positive residue (bright structures), negative residue (dark structures). The macula appears as a bright circle on the third image.

Since the macula and vessels are darker than their surroundings, they should appear as bright structures on the negative residue. Furthermore, the macula is an avascular region, meaning there should be no blood vessels around. In Fig. 6.2, it can be seen that it does appear as a bright circle surrounded by a darker region. The positive residue contains bright structures, among which the optic disc, which is helpful for localizing the macula. The first channel, that we call the *horizon*, should contain little to no information. We include it nevertheless in our inputs, because it guarantees that no information is lost in the transformation. In fact, as explained in section 6.3, any convolutional neural network taking the green channel as input can be represented as a convolutional neural network

taking the decomposition as input.

6.3 Network architecture

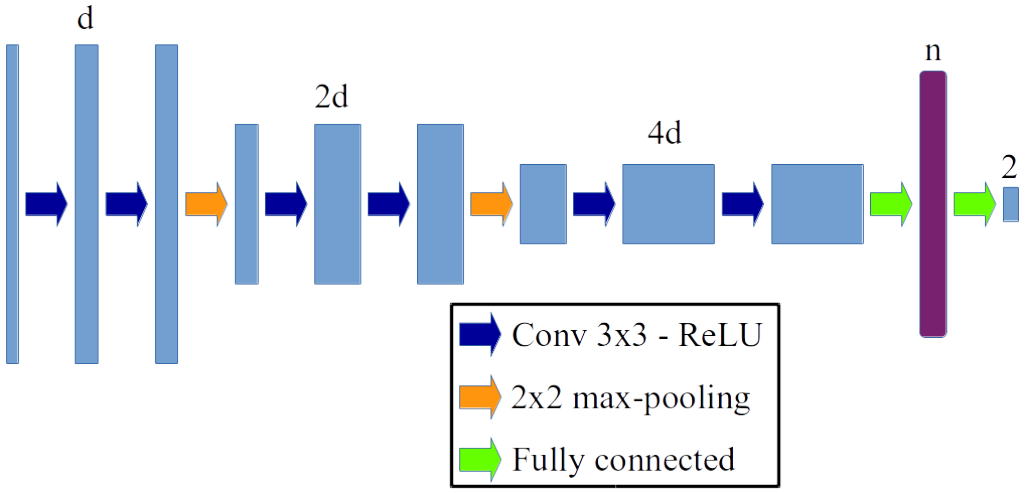


Figure 6.3: Architecture of the network: the input image can have either 1 or 3 channels. The last layer before the output is fully connected and contains most of the network's parameters.

The network we used is almost the same as in the previous chapter, except for the last two layers: the last max-pooling layer is suppressed, and the last one is linear instead of logistic. Its architecture is recalled in Fig. 6.3. The depth of the first two convolutional layers, denoted by d in the figure, was set to 8 in all our experiments. After each max-pooling layer, the depth is doubled. We have tried different values of n, the number of neurons in the fully-connected layer, ranging from $n = 6$ to $n = 32$. This meta-parameter controls the degree of complexity of the network.

We actually defined two types of networks, depending on the number of channels of the inputs (1 for the green channel, 3 for the morphological decomposition), but only their first convolutional layer is different. In the case of green channel inputs, this layer is described by d 3x3 filters (matrices) $\mathbf{A}_k, k = 1 \dots d$ and d scalar biases $b_k, k = 1 \dots d$. If we consider an input (single-channel) image $\mathbf{I}$, the k-th channel of the intermediary image after the first layer is the convolution $\mathbf{I} \star \mathbf{A}_k + b_k$.

Now let us consider the decomposition of the image into its horizon, positive and negative residue: $\mathbf{I} = \mathbf{H} + \mathbf{P} - \mathbf{N}$, as described in section 6.2. In order to describe the first convolutional layer of the second type of network, we now need to define d triplets $(\mathbf{A}_k^h, \mathbf{A}_k^p, \mathbf{A}_k^n), k = 1 \dots d$, and d biases $b_k, k = 1 \dots d$. The layer's output's k-th channel is given by: $\mathbf{H} \star \mathbf{A}_k^h + \mathbf{P} \star \mathbf{A}_k^p + \mathbf{N} \star \mathbf{A}_k^n + b_k$.

Let us consider now the particular case where for each k there exists a matrix $\mathbf{A}_k$ such that $\mathbf{A}_k^h = \mathbf{A}_k^p = -\mathbf{A}_k^n = \mathbf{A}_k$.

Clearly $\mathbf{H} \star \mathbf{A}_k^h + \mathbf{P} \star \mathbf{A}_k^p + \mathbf{N} \star \mathbf{A}_k^n = (\mathbf{H} + \mathbf{P} - \mathbf{N}) \star \mathbf{A}_k = \mathbf{I} \star \mathbf{A}_k$.

This equality shows that the class of models described by the first type of networks (green channel input) is strictly included in the class of models described by the second type (3-channel decomposition input). This is achieved via 144 extra parameters, but as previously mentioned, the model's complexity is driven by the number of neurons in the fully connected layer, n: the 1D-input network already has $32771n + 18042$ parameters.

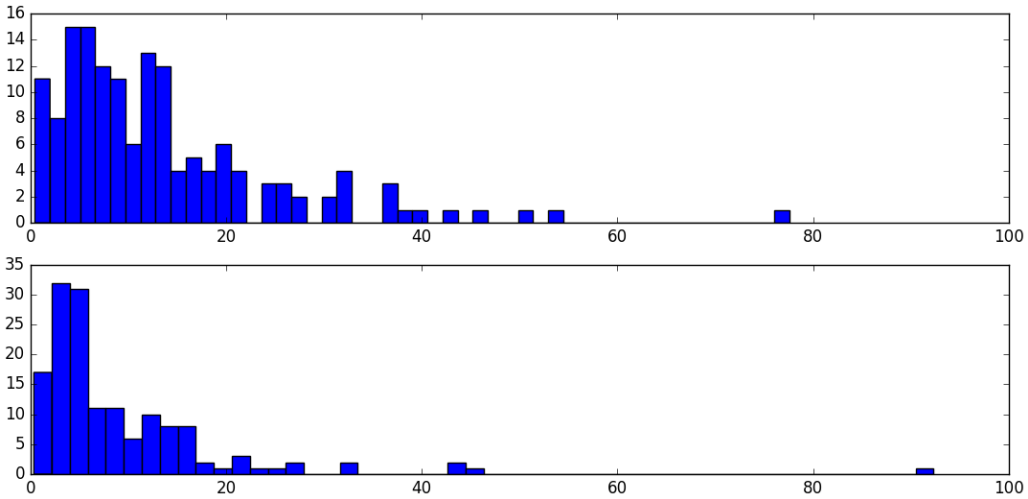


Figure 6.4: Histograms of distances between predicted location and ground truth in the preliminary experiment. Top: green channel input, bottom: morphological decomposition 3-channel input.

6.4 Preliminary Results

Our first database, introduced in the previous chapter, contained 1800 retinal images. On 1155 of those, the macula was deemed visible and entirely within the image by both annotators. Using 1000 images for training and the rest as validation set, we trained two networks, one with the green channel as input, the other one with the morphological decomposition as input.

The first results were in favor of the morphological decomposition: the average distance to the ground truth was 9.2 pixels in this case, as opposed to 13.5 pixels for the green channel input network. The histograms of distances are shown in Fig. 6.4.

From this first experiment, it would be tempting to assume that providing the three channels of the morphological decomposition instead of the raw green channel makes the task easier; the network possibly can design some features it cannot with the single channel input. It is even possible that the decomposition-input network ignore the first two channels altogether and creates features based on the third one only.

In order to either confirm or deny these intuitions, we conducted a more thorough investigation, presented in the next section.

6.5 New database

We annotated another 2700 images, all taken at random from the OPHDIAT database. The annotation was again performed by the same two annotators, independently from each other, by clicking on the fovea, at the condition that the macula be clearly visible and entirely within the image. In a second phase, a consensus was found between the annotators when there was a disagreement on the visibility of the macula.

This time, we did not include the images from the e-ophtha databases, in order not to introduce a quality bias. A first remark that can be made about the new database concerns the proportion of images with visible macula: it was considered visible on 2035 images out of a total 3843, which is a little less than 53%, as compared to the almost 64% (1150 out of 1800) of the initial database.

The database on which our experiments were performed consists in the 2035 images where the macula was deemed clearly visible and within the field of view by both annotators.

This database (containing the images where the macula was visible) was split into three different

sets: one for training, one for validation and one for test. Our base contains 2035 images, but they come from 1650 patients only; we made sure in splitting the database that no patient appeared in two different sets. We selected 1300 patients for training, 200 for validation and 150 for test, resulting in respectively 1587, 253 and 195 images.

6.6 Results on the new database

All networks were trained using the RMSProp [TH12] optimizer with a dropout of 0.7, and L^2 loss. Let (x, y) be the ground truth coordinates of the fovea on a particular image and $(\hat{x}, \hat{y})$ the predicted values: the associated loss is $L = (x - \hat{x})^2 + (y - \hat{y})^2$.

Learning was performed over 15000 epochs; at each epoch, validation loss was evaluated and the model was saved if it improved over the best value so far.

For each value of $n \in \{6, 8, 12, 16, 32\}$, three networks of each type with n neurons in the fully-connected layer were trained with different random initializations. A fourth one of each type was trained for $n = 6$ because two out of the first three 'green-channel' networks performed significantly worse than the third one.

Results are presented in tables 6.5 and 6.6, respectively for green channel and morphological decomposition, in terms of root mean square error. Minimizing the L^2 loss is equivalent to minimizing the root mean square error, but the latter is homogeneous to a distance, and is an upper bound of the mean distance error, making it somewhat more easily interpretable.

n=32	n=16	n=12	n=8	n=6
3.17	3.45	4.69	5.37	5.41
3.48	4.74	4.94	5.67	6.07
3.66	5.65	6.42	12.44	11.42
				11.53

Figure 6.5: Best validation error across 15000 epochs for green channel input networks.

n=32	n=16	n=12	n=8	n=6
3.23	4.48	5.36	4.81	5.66
3.33	5.01	5.75	5.15	5.96
3.80	5.81	5.84	5.67	6.16
				6.71

Figure 6.6: Best validation error across 15000 epochs for morphological decomposition input networks.

The two types of networks yield fairly similar performances, even if for all values of n except $n = 8$, the lowest loss is achieved by a green channel input network. Three of the smallest networks of this type, though, significantly underperform, which does not happen when the morphological decomposition is used as input.

Since the database was annotated by two people, we can also compare these results to human performance: the mean distance between the two annotators is 1.59, about half the validation loss of the best network.

The L^2 loss (or equivalently, the root mean square error) is only one, scalar, evaluation metric, and does not capture all the information about the predictions. In the following section, we take a closer look at the predictions.

6.7 Problem anisotropy

The OPHDIAT database contains two types of images: acquisitions centered on the macula, and acquisitions centered on the optic disc. In both cases, the patient's eye during the acquisition is horizontal; he is never supposed to look up or down. This means that the fovea's vertical position in our database is always roughly the same, as illustrated in Fig. 6.7. Let us consider a model that always predict the correct horizontal position, and always predicts a constant value, equal to the mean on the training set, for the vertical position. This model would have a root mean square error of 2.98, which is lower than all of the networks we trained.

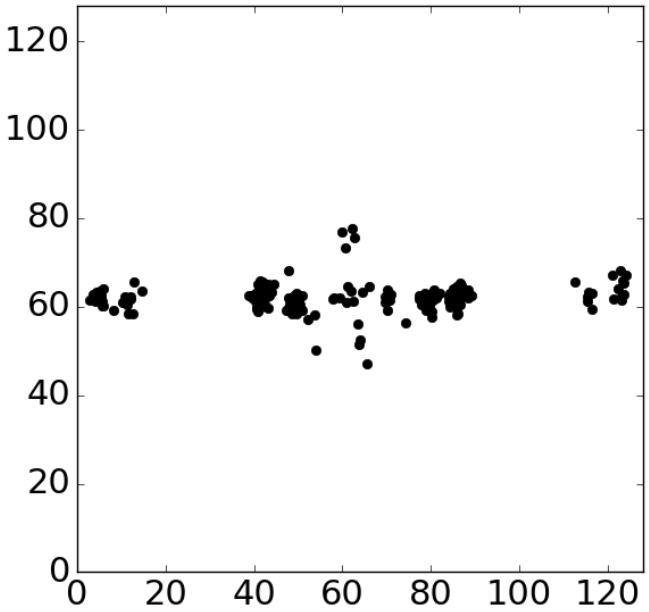


Figure 6.7: Ground truth positions of the fovea in the validation base.

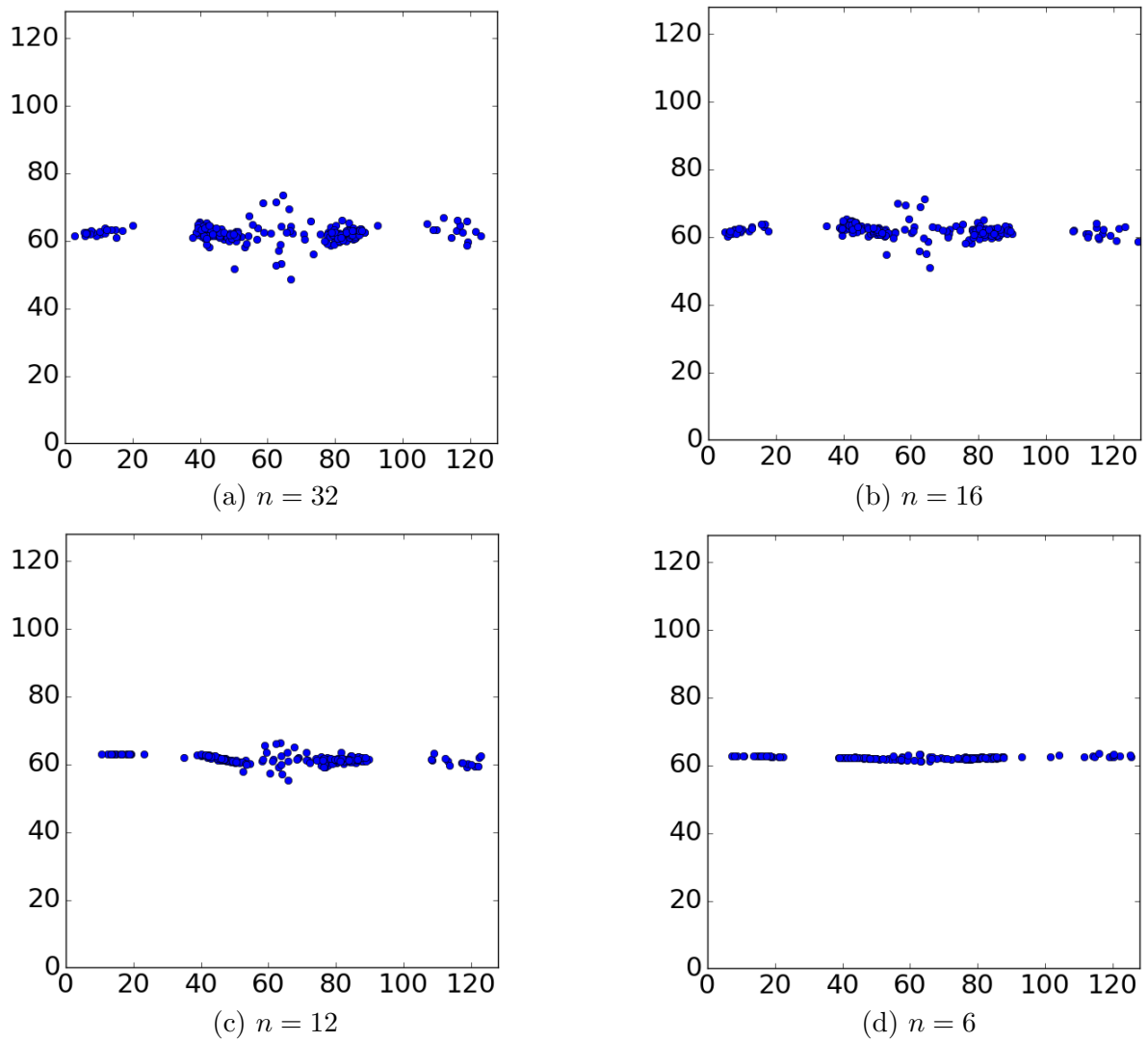
Figure 6.8 shows that this is precisely what our networks tend to do: the smaller the network is, the smaller the vertical variance of the predictions. The predictions shown here are those of the best green channel input network with n neurons in the last layer, for n in $\{6, 12, 16, 32\}$; the sets of predicted locations looked the same for the decomposition input networks.

This can be interpreted by considering that the output layer of the neural network is a linear regression from $\mathbb{R}^n$ to $\mathbb{R}^2$, and that all the rest of the network is learning n 'good' features for this regression task. But since the vertical position is almost constant, it is easily approximated with no feature at all; the smallest networks, with $n = 6$, seem to learn 6 features that are as relevant as possible for predicting the horizontal position, and zero feature for predicting the vertical one. Although for higher values of n, the vertical variance of the predictions increases, it remains about half that of the ground truth even for $n = 32$.

6.8 Output Normalization

In order to better predict the vertical positions, we normalized the outputs. The raw coordinates (x, y) were replaced by $(a, b) = (\frac{x-\mu_x}{\sigma_x}, \frac{y-\mu_y}{\sigma_y})$, where μ_x, σ_x are the mean and standard deviation of the x coordinate on the training set (similarly for μ_y, σ_y).

Given a prediction $(\hat{a}, \hat{b})$, we can of course retrieve the corresponding predicted coordinates $(\hat{x}, \hat{y}) = (a\sigma_x + \mu_x, b\sigma_y + \mu_y)$.



(a) $n = 32$ (b) $n = 16$

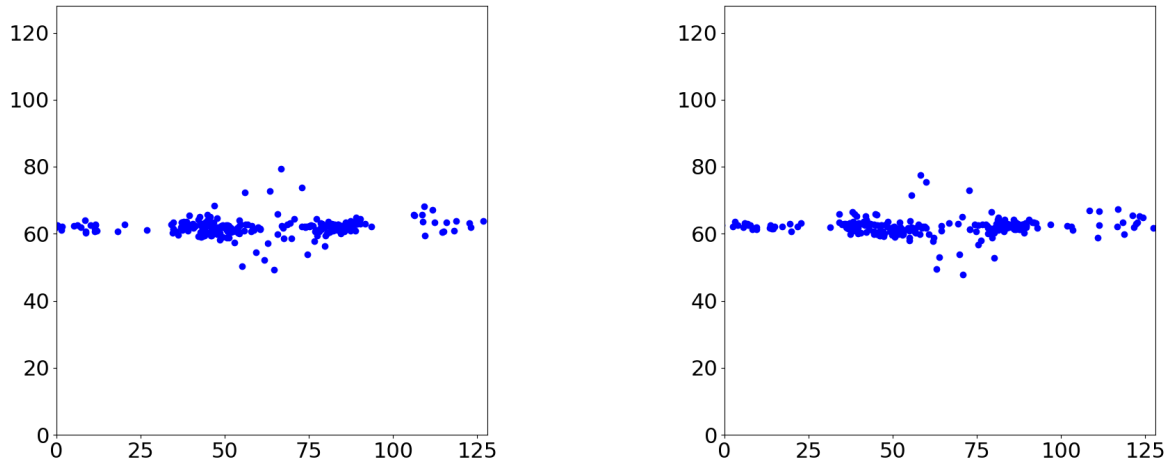
(c) $n = 12$ (d) $n = 6$

Figure 6.8: Fovea location predictions on the validation database for different green-channel input networks with decreasing number n of neurons in the last layer.

Note that minimizing the L^2 loss between the prediction $(\hat{a}, \hat{b})$ and ground truth (a, b) is equivalent to minimizing the modified L^2 loss $L = (x - \hat{x})^2 + \lambda(y - \hat{y})^2$, where $\lambda = \sigma_x/\sigma_y$.

The predictions of new networks, with normalized outputs, are shown in Fig. 6.9. Because of the larger penalty for vertical errors, for a given number of neurons n in the fully-connected layer, the variance of the y coordinate is greater than before, which was expected: for $n = 16$, the vertical variance is 7.34 with output normalization, and only 3.72 without. For $n = 32$, the vertical variance is 7.41 with output normalization, and 5.70 without. For reference, the vertical variance of the annotations is 8.79.

As mentioned in the previous section, the output layer of our networks can be seen as a linear regression with n features, and the rest of the network as a mapping between the input image and those features. By normalizing the outputs, or equivalently by using a modified loss function, it does seem that we can coerce the network into designing features that are useful for predicting the fovea's vertical position. Unsurprisingly, however, this is at the expense of overall precision, when measured in terms of the usual euclidian norm. Indeed, with normalized outputs, the mean error (distance between prediction and annotation) is 4.34 pixels for the network we learned with $n = 32$ (4.59 pixels



(a) $n = 32$ (b) $n = 16$

Figure 6.9: Predictions of the regression networks with output normalization on the validation database.

for $n = 16$), which is almost twice as high as the mean errors obtained without normalization (2.43 pixels for $n = 32$, and 2.71 pixels for $n = 16$).

The fact that, for a given value of n, normalizing the outputs worsens the network's performance could be expected. For the sake of simplicity, let us make the overly simplistic assumption that each of the n learned features is either useful for predicting the x coordinate but entirely uninformative for predicting the y coordinate, or useful for predicting y but entirely uninformative for predicting x. Given the anisotropy of the problem, it can be expected that if the outputs are not normalized, most or all of the n features would be of the first kind, and that few to none would be of the second kind: Fig. 6.8 tends to comfort this view, notably in the $n = 6$ case.

If the outputs are normalized, we could expect a more balanced ratio of horizontal and vertical features; out of $n = 32$, for instance, we could imagine that about 16 features would be useful for predicting the x coordinate, and 16 would be useful to predict the y coordinate. Since the actual L^2 loss is mostly driven by the horizontal precision, this would make this network at a disadvantage when compared to the unnormalized network with $n = 32$ features, but it should have a couple more horizontal, and way more vertical, features than the unnormalized network with $n = 16$ features.

Actually, the performance of the normalized network with $n = 32$ (4.59 pixel mean error) is significantly worse than what we obtained with $n = 16$ in the unnormalized case (2.71 pixel mean error). We trained a network with $n = 64$ neurons with normalized outputs, and its performance (3.85 pixel mean error) is not significantly better than one of the unnormalized networks with only $n = 6$ (Fig. 6.8(d)), which had a mean error of 3.89 pixels.

There are two possible explanations: either coercing the network into learning better vertical features makes it harder for it to learn good horizontal ones, or the network does learn quite good horiztonal features, but the weights of the linear regression are suboptimal when measuring the performance in terms of the usual distance.

The hypothesis that features are either 'purely vertical' or 'purely horizontal' is clearly simplistic (although this is debatable in the $n = 6$ case), but the general idea remains the same, even taking into account hybrid features that have some predictive power for both coordinates.

Contrary to our intuition that providing a more 'sensible' input could lead to better performance,

and despite promising preliminary results, it seems that the morphological decomposition preprocessing we introduced does not, in fact, help the network learn better features for localization.

Using the L^2 loss with a very anisotropic output distribution leads to an unwanted behavior of our prediction models: it is in a sense 'not worth' predicting the vertical position, which is almost constant. Normalizing the outputs leads to more variance in the distribution of the y coordinates, but this is at the expense of overall performance in terms of mean error. If we consider the n neurons in the last hidden layer as features designed by the first layers of the network, and the output layer as a linear regression model, it is unclear whether the bad performance of the networks with normalized outputs is due to learning bad features or to learning bad weights in the linear regression part.

We next investigate a deeper architecture, which may be able to learn better features than the shallower one used so far.

6.9 Final database

With the same methodology described previously, we annotated more images; our final database totals 6098 images, the macula considered visible and entirely within the field of view on 3142 of which. In the rest of this chapter, since we focus on the regression task, we only focus on these 3142 images, although, as we shall see, fully-convolutional networks could handle all images, contrary to the regression ones.

We used 2193 images as our training set, 639 as our validation set and 310 images as our test set. The positions of the validation dataset are shown in Fig. 6.10.

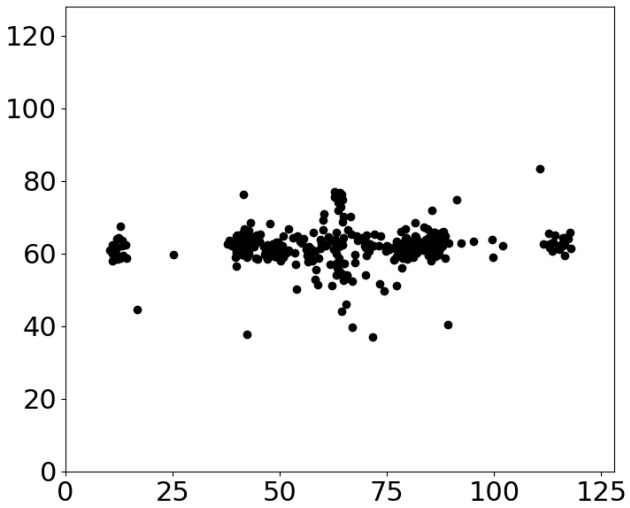


Figure 6.10: Positions of the fovea in the new validation dataset.

6.10 Deeper Regression Networks

On this new database, we tried a deeper architecture, illustrated in Fig. 6.12. As in the previous section, it consists in a succession of convolution-ReLU/convolution-ReLU/max-pool segments, the difference being in the number of these segments (five instead of three). Each max-pooling layer is a 2x2 pooling with stride 2, meaning that the resulting block is half the height and half the width. After each pooling, the depth of the following convolutional block is doubled. We kept $d = 8$ for our

initial depth; the last convolutional block before the fully-connected layer is thus 256 channels deep, but each channel is only 4x4.

The number of parameters of these networks is $2051n + 294906$; compared to the previous architecture, there are many more weights in the convolutional layers, but since the last block before the fully-connected layer is much smaller, the number of parameters grows much more slowly with n.

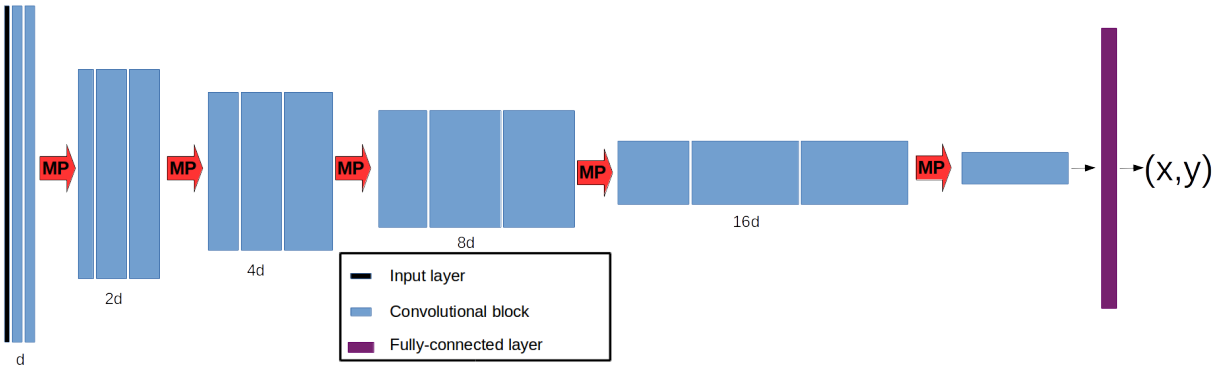


Figure 6.11: Architecture of the deeper regression networks. Each convolutional layer is followed by a Rectified Linear Unit (ReLU). The red arrows labeled MP indicate a 2x2 max-pooling with stride 2, dividing the height and width of the block by 2. After each max-pooling, the width of the convolutional block is doubled.

6.11 Results

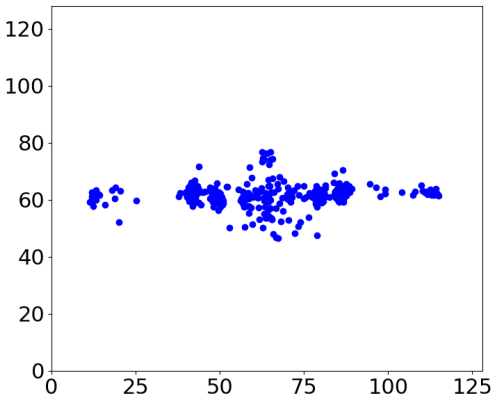
We trained three networks, with $n = 8$, $n = 32$ and $n = 64$ neurons in the fully-connected layer, without output normalization. The predictions on the validation set are shown in Fig. 6.12. Even with n as small as 8, the networks do not exhibit the same behavior — all predictions on the same horizontal line, or very close to it — as the ones in the previous section. The vertical variances of the three models are almost the same: 10.29, 10.51 and 10.30 for $n = 8$, $n = 32$ and $n = 64$, respectively. The vertical variance on the validation set is actually a bit higher (15.28), but the difference is less drastic than with the shallower networks.

In terms of localization, all three networks also perform much better: the mean distances between prediction and annotation are 1.41, 1.29 and 1.34 pixels, for $n = 8, 32, 64$ respectively. The network with $n = 32$ performs better than the network with $n = 64$, indicating a probable overfitting of the latter.

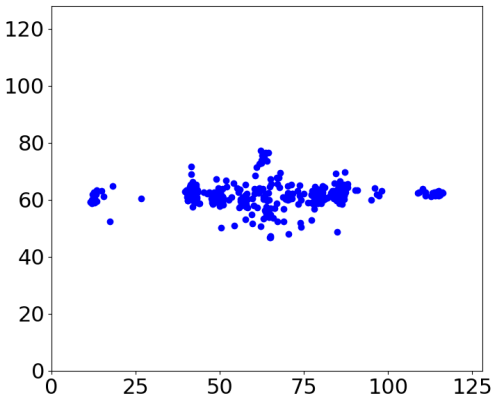
If we consider again that the first layers, up to the fully-connected one, design the features of a linear regression model (the output layer), the features learned by the deeper networks seem to be much more meaningful than the ones learned by the shallower ones. In particular, the deeper network with only $n = 8$ performs better than the shallower ones with $n = 32$, despite a smaller number of parameters ('only' 311314 against more than a million).

6.12 Limitations

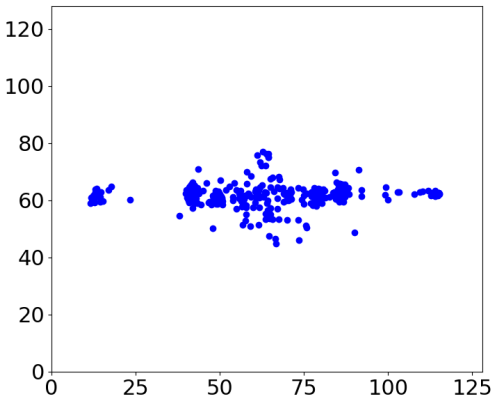
Although the mean error of the networks, of the order of the pixel, is very low, they still make large errors on some images: the best network on the validation set, in terms of mean error, is the one with $n = 32$ neurons in the last layer. If we consider that at this resolution, the macula is approximately a circle of radius 10, and that the localization fails when the predicted location for the fovea is outside



(a) $n = 8$



(b) $n = 32$



(c) $n = 64$

Figure 6.12: Predictions of the three deeper networks on the validation set.

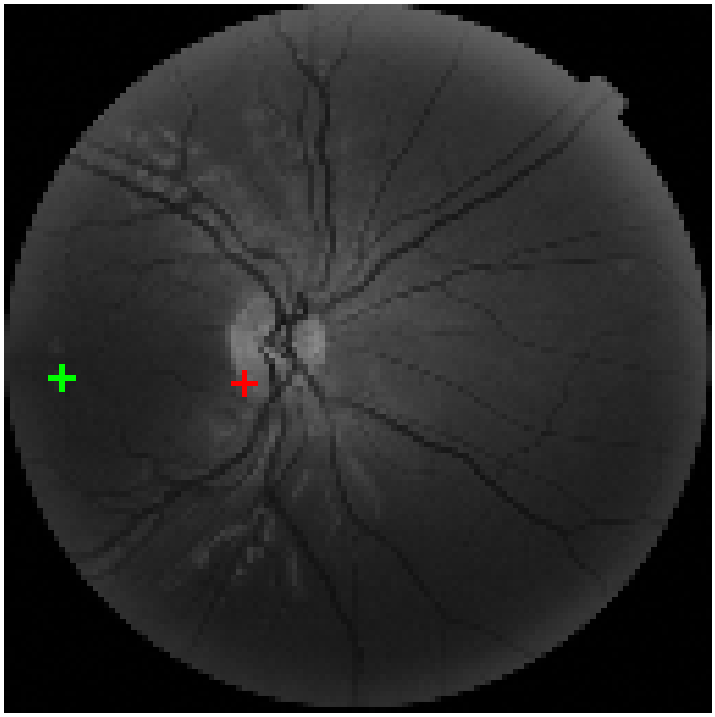


Figure 6.13: Example of localization failure. The green cross indicates the manual annotation of the fovea, the red cross is the prediction of the network.

the macula, this network fails on 5 images. An example can be seen in Fig. 6.13: the prediction of the network is actually on the optic disk, 33 pixels away from the fovea's manual annotation.

6.13 Conclusion

There are two issues with approaching macula/fovea localization as a regression problem: the model, in this case a neural network, always outputs a pair of coordinates, no matter the input. In particular, if we use a classifier (such as the one described in Chapter 5) to discriminate between images where the macula is visible and images where it is not, whenever there is a classification mistake, the system will predict a location even though the macula is not visible. The other issue, even assuming that we dispose of an ideal classifier, is that there is no way of knowing when the model might be making a problematic mistake; ideally, it would be useful to have a confidence score associated to a prediction, and to trigger a warning when this score is too low.

In this chapter, we have seen that our first, 'shallow', networks, have trouble dealing with the anisotropy of the outputs, and that normalizing them does not solve this problem. The deeper architecture performs better than the shallower ones, even with fewer parameters. The databases were not exactly the same, which has to be taken into consideration, but the difference in terms of mean error seems to large for this phenomenon to have happened purely by chance. The best network has a very low mean error, of the order of the pixel on 128x128 input images, and makes very few problematic errors.

Coupled with a classifier such as the one described in Chapter 5, we can obtain a system able to determine with great accuracy whether the macula is visible and entirely within the image, and to localize it if it is. The main issue is that errors, although rare, are hard to detect. The classifier does predict a soft score between 0 and 1 but in practice it is very close to 0 or to 1 most of the time; the regression network returns a prediction but no confidence score.

As an alternative to this system, we also trained fully-convolutional networks, which are able to perform both tasks, as we shall see in the next chapter.

Chapter 7

Fully-Convolutional Networks for Segmentation and Image Quality Estimation

Ce chapitre est une version plus détaillée de l'article Fast macula detection and application to retinal image quality assessment *(Robin Alais, Petr Dokládal, Ali Erginay, Bruno Figliuzzi, Etienne Decencière), publié dans Biomedical Signal Processing and Control 55 (2020) [Ala+20].*

Après avoir procédé en deux temps (classification puis localisation) dans les chapitres précédents, nous cherchons à présent à obtenir une segmentation directe de la macula. Pour ce faire, nous utilisons des réseaux entièrement convolutionnels dont la sortie est une image de même taille que l'image d'entrée.

La visibilité de la macula est un critère essentiel pour les ophtalmologistes pour juger de la qualité d'une image centrale : la sortie de notre réseau, moyennant un post-traitement morphologique, permet de fournir non seulement la localisation de la macula mais également un score de qualité.

Par ailleurs, l'algorithme que nous proposons ici est très léger (peu de paramètres) et rapide, ce qui le rend particulièrement adapté à un déploiement sur des architectures embarquées.

Les résultats obtenus en terme de localisation de la macula sont meilleurs qu'avec les réseaux de régression des chapitres précédents, et la performance de classification des images (la macula est-elle visible ou non ?) est proche de la performance humaine. Bien que la visibilité de la macula ne soit pas le seul critère pour déterminer si une image rétinienne est de bonne qualité ou non, l'algorithme proposé pourrait permettre de réduire significativement le nombre d'examens non évaluables dans un réseau de télémédecine et peut être complété par d'autres critères de qualité.

This chapter is an extended version of the article *Fast macula detection and application to retinal image quality assessment* (Robin Alais, Petr Dokládal, Ali Erginay, Bruno Figliuzzi, Etienne Decencière), published in Biomedical Signal Processing and Control 55 (2020) [Ala+20].

7.1 Problem Presentation

Image quality estimation is a necessary preliminary step to most retinal image understanding tasks, since it would make little sense applying an automatic diagnosis algorithm to images too noisy, blurred or not contrasted enough. Most publicly available datasets, like the Kaggle Diabetic Retinopathy dataset or the Messidor database [Dec+14], contain only gradable images, and in [NAG09], it is clearly mentioned that for building a local database, "acceptable image quality, as judged by the screening program ophthalmologists, was a selection criterion".

In the context of a telemedicine network, the diagnosis is performed by a human expert, but the

photographs are taken at a different time and location, by operators whose skill and level of experience can vary. A significant portion of images - around 10% for the OPHDIAT network [Mas+08] - are deemed uninterpretable by ophthalmologists. This could be prevented by automatically estimating the quality at acquisition time, sending a warning to the operator if the photograph should be re-taken.

For a diagnosis to be made, an essential requirement is that the region of the macula must be clearly visible. In the present chapter, we focus on this task and present a lightweight CNN architecture capable of evaluating if the macula is 1) clearly visible 2) entirely within the field of view. In addition, our algorithm is able to accurately locate the macula if both conditions are met. This position can be used as well to check if a central image is correctly centered on the macula, or as part of an automated diagnosis algorithm, where the distance between lesions and the macula is an important information.

7.2 Related Work

So far, proposed methods for locating the macula either require images to be of sufficient quality [NAG09], or attempt at providing a location based on contextual information like the optic disk and vascular network, even if the macula itself is not visible in the image [Ver+13; WSM11]. The originality of our approach lies in the fact that we use real clinical data, including very low-quality images; we automatically assess image quality, and we deliberately aim at detecting the macular region *only* if its visibility is sufficient.

The earliest attempts at defining a score for eye fundus image quality relied on properties of intensity histograms [LW99; LGB01]. Image structure clustering, introduced in [NAG06], applies a bank of filters in order to perform unsupervised segmentation into several clusters roughly corresponding to anatomical structures of the retina, such as optic disk, vessels or retinal background. In this article, authors summarize a retinal photograph as a 20-dimensional vector composed of the histogram of the image structure clusters (5 bins), along with the three histograms of each color plane (5 bins each); good quality images are then separated from bad quality images by means of a Support Vector Machine. A similar approach was used in [Pau+10], although for images centered on the optic nerve of only 22.5° field of view. The authors also used a Support Vector Machine as their final classifiers, but they use Haralick [H+73] and sharpness features.

Other features have been proposed, such as the density of visible blood vessels, either in the whole image [Gia+08; Gia+10] or near the macula [Hun+11]. Measures of clarity [Fle+12] and blurring [Pir+12] have also been defined. Finally, some methods combine both general image features and retina-specific ones, making use of vessel density, histogram, textural, and local sharpness [Yu+12; POS14; DOC12].

In recent years, convolutional neural networks have been proven very efficient on difficult computer vision tasks, notably winning the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012 competition [KSH12]. CNNs have then been applied to various tasks, including segmentation of retinal images [Man+16; AlB+18], glaucoma and DR grading [Pra+16; Gul+16; Gra15].

In this context, estimating retinal image quality with convolutional neural networks is an interesting research direction. In [Mah+16a], a convolutional network is trained in order to discriminate gradable images from artificial ungradable images obtained by adding noise to the original images of the DRISHTI dataset [Siv+15], which contains 101 acquisitions centered on the optic disk, with a 30° field of view, all images being taken with the pupils dilated. In [Ten+16], the same authors have experimented with both a 'shallow' network (the total number of weights cannot be calculated, since the number of neurons in the two fully-connected layers is not given, but the weights in the convolutional layers alone exceed 1 million) trained from scratch, and AlexNet [KSH12] fine-tuned on a dataset consisting in 908 ungradable and 944 gradable non-mydriatic images. On a larger dataset (9653 ungradable retinal images and 11347 gradable images), they also evaluated the possibility of using a hybrid method combining saliency maps and CNNs [Mah+16b]. Finally, [Sun+17] compare the

performance of fine-tuning four CNN architectures - AlexNet [KSH12], GoogLeNet [Sze+15], VGG-16 [SZ14] and ResNet-50 - on a 3000-image subset of the Kaggle database. These preliminary studies report that large networks are hard to train, and must deal with overfitting issues, due to the huge amount of parameters. In an attempt to overcome this problem, data augmentation is extensively used, to the point where so much distortion is introduced that most of the training data is not constituted of real or even realistic examples. An example of this is generating artificial new data by applying large rotations (up to 210° !) to images; the resulting images are unrealistic, and it makes the network more or less rotational invariant, which is not a desirable feature for analyzing eye fundus images.

Another drawback of very large networks is their integration in embedded systems. Several million weights can constitute a significant fraction of the available memory: 233MB for AlexNet, 528MB for VGGNet for weights and biases alone, and the prediction times on embedded CPUs can be on the order of a second [Lu+17].

In this work, we propose a lightweight solution, with only 8329 parameters, and a reduced number of convolutions to be performed, meaning low power consumption as well. A comparison of memory requirements and computation times on embedded systems between our algorithm and other classic convolutional networks is given in Table 7.1. Our algorithm, including disk access and post-processing, was benchmarked on a Raspberry Pi; timings for the other networks come from [Lu+17] and were obtained on a more powerful 1.9GHz quad-core ARM Cortex-A57 64bit CPU (NVIDIA TX1). For a given task, we should expect it to be performed faster on the TX1 than on the Raspberry Pi. Despite this, our algorithm is the fastest, being more than three times as fast as ResNet, and more than 17 times as fast as VGGNet. It is also the lightest by far, requiring only 98kB for storing the network's weights. We have also benchmarked a modified version of the segmentation network U-Net [RFB15], whose performance for our task is evaluated in section 7.5.

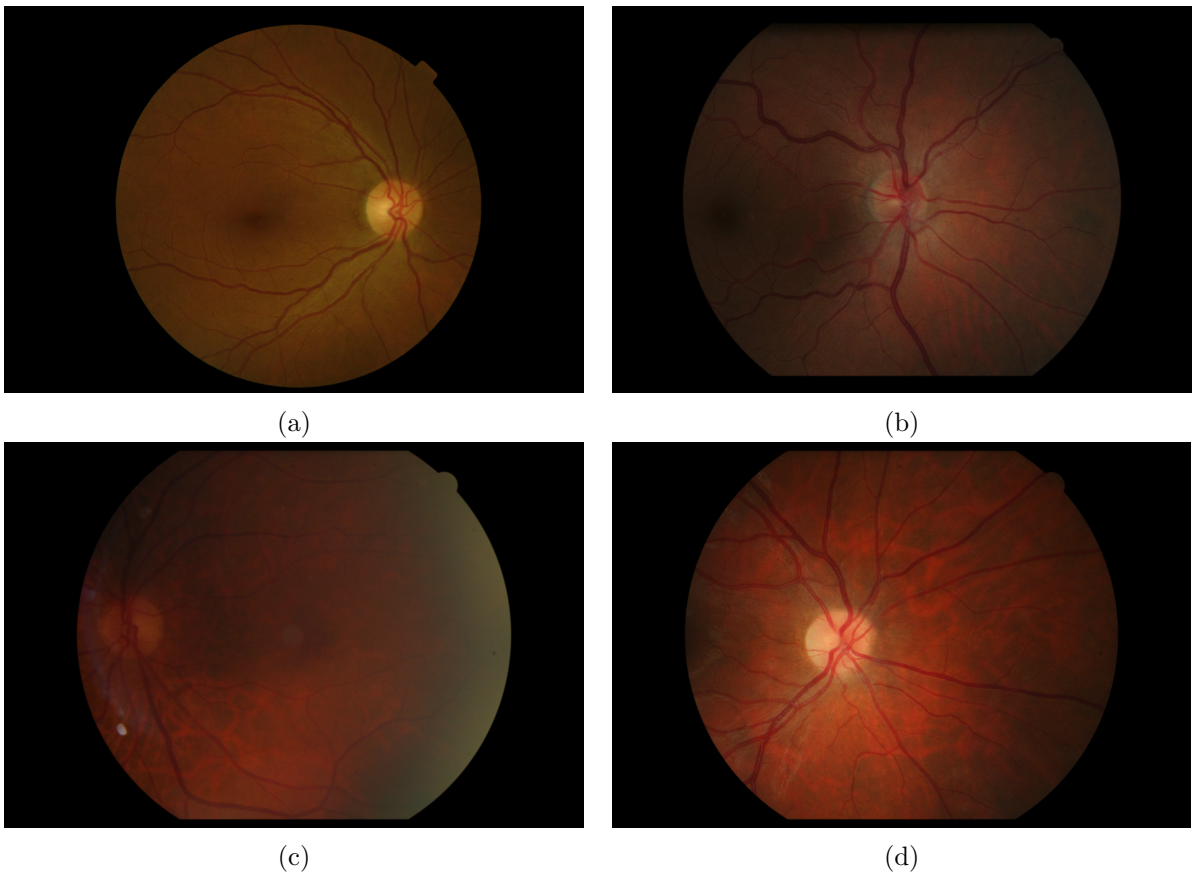
Network	Time (ms) TX1 (CPU)	Time (ms) Rasp.Pi	Weights (MB)
AlexNet	893		233
VGGNet	2809		528
GoogLeNet	638		26
ResNet	567		97
Us		161	0.1
U-Net		92	0.6

Table 7.1: Computation times and memory use of various convolutional networks on embedded systems. The first four networks have been benchmarked in [Lu+17] on a NVIDIA Jetson TX1. Our network and our implementation of U-Net were benchmarked on a Raspberry Pi; it should be expected that they would have run faster on the TX1.

7.3 Database

We extracted 6098 eye fundus images from the e-ophtha database [Dec+13]. This database has itself been extracted from the OPHDIAT telemedicine network for DR screening. These images are either *central* (centered on the macula), or *nasal* (centered on the optic disk). Different retinographs were used for the acquisitions, with resolutions ranging from 1440x960 to 3504x2336 pixels. Two different readers, independently from each other, indicated whether or not the macula was both entirely within the field of view, and clearly visible, meaning that the fovea and the small vessels around the avascular region can be seen. When that was the case, the fovea's position (x- and y- coordinates) was labeled.

When there was a disagreement on the macula's visibility, a decision was made on which annotation to use. Out of the 6098 images that were considered, the macula was deemed visible by both readers on 3142. The remaining images correspond either to bad quality central images or to nasal images where the macula is at least partly outside the field of view (FOV). Sample images are shown in Fig. 7.1.



(a) (b)

(c) (d)

Figure 7.1: Sample images from the e-ophtha dataset: the macula is considered clearly visible and entirely within the FOV in both (a) and (b). It is in the center of image (c) but the quality is insufficient for grading, and it is partly outside the FOV in (d), which is a good quality nasal image.

7.4 Methodology

7.4.1 Image Preprocessing

Briefly put, the idea is to train a network to segment the macular region: with an ideal algorithm, if the segmentation is unsuccessful, it means that no macula is visible; if the segmentation succeeds, it means that image quality in the macular region was sufficient, and we immediately get fovea localization as a byproduct.

To train the network, we used the green channel, which was cropped, zero-padded in order to retain a square frame when necessary, and resized to a 128x128 image (see Fig. 7.2). This can seem aggressive, and some details might be lost, but approaches for assessing the severity of DR use networks with inputs as small as 512x512 [Que+17], even though the task is much more complex and can rely on the presence of small structures like microaneurysms. A previous work for localizing both macula and optic disk in good-quality images used as input the green channel resized to 256x256 [AlB+18]. The aim of this work is a bit different, since we are less interested in predicting a precise location -

although our algorithm never predicts a location outside the macula (see Sec. 7.5.2) - but rather in estimating the quality of the macular region. A 128x128 resolution is sufficient to visually assess the visibility of the macula, and means that smaller networks can be used, easier to train and less costly to run in an embedded framework, in real time.

Where the macula was visible, we used the mean of the two annotations as reference for the fovea's position; in 128x128 resolution, the average distance between readers was 1.25 pixels. We considered the macula to be about 20 pixels wide, and we used as ground truth a disk of ones of radius 10 pixels, the rest of the image being set to zero. When the macula was not visible because of low image quality or because it was at least in part outside the FOV (Fig. 7.1c and 7.1d), the ground truth was an image of zeros.

The only pre-processing consisted in dividing the (8-bit) images by 255, in order to get images valued between 0 and 1. No contrast enhancement or filtering was applied, since we wanted to evaluate the quality and macula visibility of raw images. Data augmentation was used by applying vertical symmetries (transforming a right eye into a left eye or the other way around), but no horizontal symmetries or rotations were used, to avoid creating unrealistic training data.

The dataset was split in three parts: a training set, used for learning network parameters, a validation set, used to estimate network performance during learning and setting some hyperparameters, and a test set, exclusively used to assess the method's performance. Images corresponding to the same patient are necessarily in the same set in order to avoid any evaluation bias. The number of images in each set is given in Table 7.2.

Additionally, in order to further estimate the generalizability of localization performance, we will use the ARIA Database C (Fig. 7.18), for which fovea localization ground truth is available, as an additional test set.

	Training	Validation	Test	All
Visible	2193	639	310	3142
Not visible	2056	596	304	2956
Total	4249	1235	624	6098

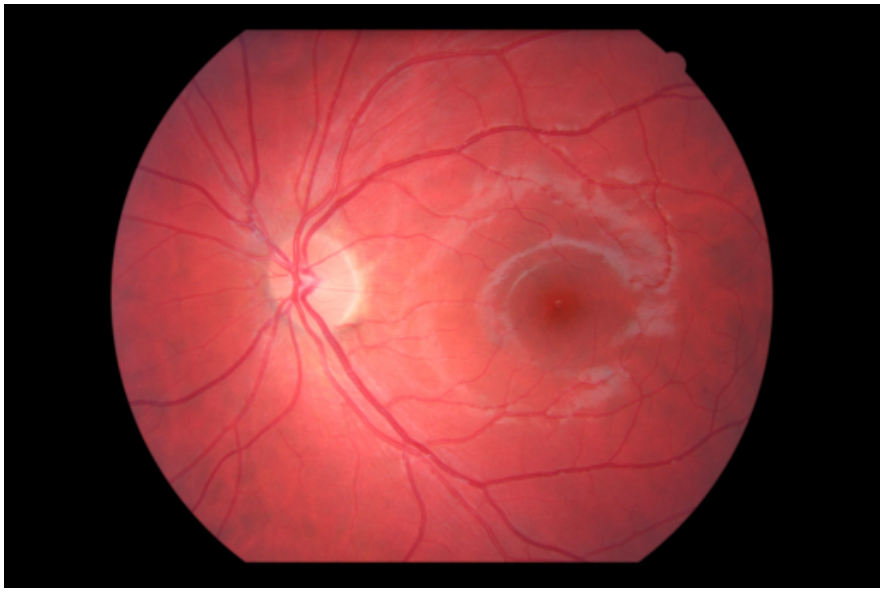
Table 7.2: Number of images in the training, validation and test sets.

7.4.2 Network architecture

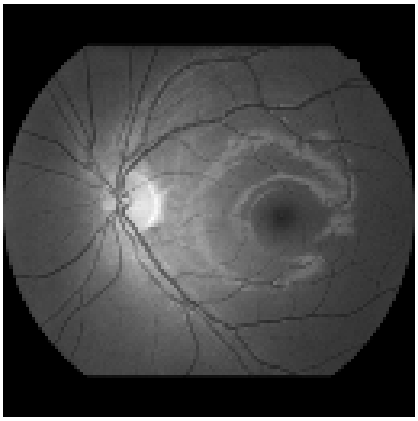
We trained a fully-convolutional network consisting only of 3x3 convolutional layers (each followed by a rectified linear unit) piled up, with all convolutional layers having 8 channels. We used zero-padding at every step, so that the output image is the same size as the input one. A similar architecture has been shown to provide good results for cell nucleus segmentation [Pan+10].

The number of convolutional layers we pile up determines the receptive field of each pixel in the output image: that is, if there are L layers in the network (including the output layer), the value of one pixel in the output image depends on the values in a $2L + 1$ by $2L + 1$ square centered at this position in the input image. This is illustrated in Fig. 7.3. We tried different values of L, ranging from 10 to 20; the best validation loss was achieved for networks with $L = 16$. The receptive field is then 33x33 pixels. When centered on a pixel at the edge of the macula, it contains the fovea and most - but not all - of the ground truth macular region.

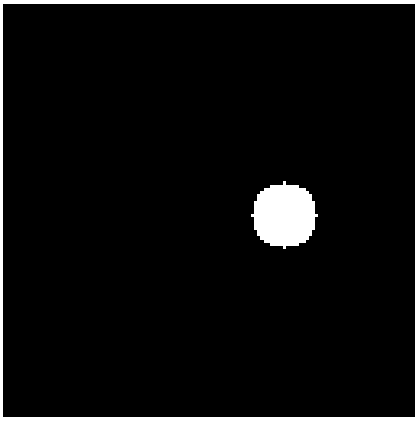
This network has very few parameters: each channel of the first layer is defined by a 3x3 convolutional kernel and a bias term. In the following layers, each channel is defined by a 3x3x8 convolutional kernel and an extra bias term (73 parameters). Finally, the output layer is defined by a 3x3x8 convolutional kernel and a bias term. The network has a total 8329 parameters, which is extremely few (in comparison, AlexNet has 60 million parameters).



(a)



(b)



(c)

Figure 7.2: The green channel of the original color image (a) has been cropped around the region of interest, zero-padded on top and bottom so as not to introduce distortion, then resized to 128x128 (b). Image (c) is the ground truth used to train our convolutional networks.

The network was initialized with truncated normal distributions with standard deviation $\sigma = 0.1$ for the convolution weights and zeros for the biases. The objective function we minimized was the L^2 distance between images. The gradient was estimated at each step on a mini-batch of 8 images, using the RMSProp optimizer [HSS12]. The network was trained for 4000 epochs.

7.4.3 Network Visualization

It is notoriously hard to interpret neural networks; however, for convolutional networks like this one, we can look at the learned filters [ZF14] or at the activations at a given layer. In this particuliar case, the convolutional filters are 3x3 (for the first layer) or 3x3x8, making the former quite uninformative and the latter hard to visualize. We can, however, look at the activations after each layer: for the input image in Fig. 7.4, we show in Figures 7.5, 7.6, 7.7, 7.8 the activations of layers 1,5,10 and 15, respectively.

Most operations of the first layer look like basic oriented edge detectors, or in the case of the

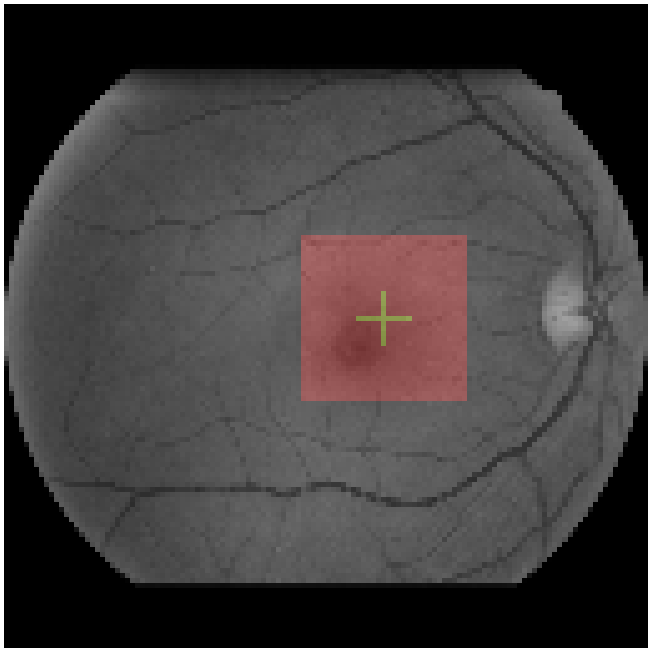


Figure 7.3: Illustration of the network's receptive field. The output value of the pixel at the center of the green cross depends on the input image's values in the red square.

Figure 7.4: Input image

first one, possibly approximately a mean filter. The activations of the following layers are harder to interpret, since they result from a succession of nonlinear operations, but as soon as the fifth layer, we can see that part of the macula contrasts with its surroundings, either as a bright spot over a dark background, or as a dark structure surrounded by a lighter zone. With this particular input image, we could devise a simple post-processing algorithm applied to the fourth channel of the tenth layer and obtain the fovea's localization. On the fifteenth layer, we can clearly see the macula's segmentation on most channels, although the fifth one seems to be uninformative in this case. The major veins,

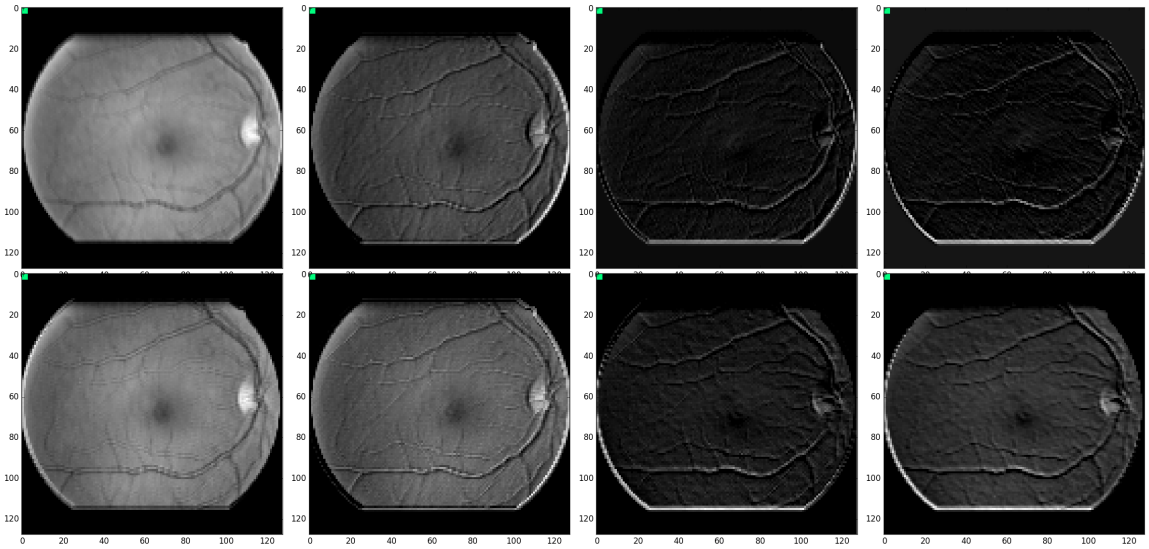


Figure 7.5: Activations of the first layer. The green square on the top left of the images represents the size of the receptive field at this layer.

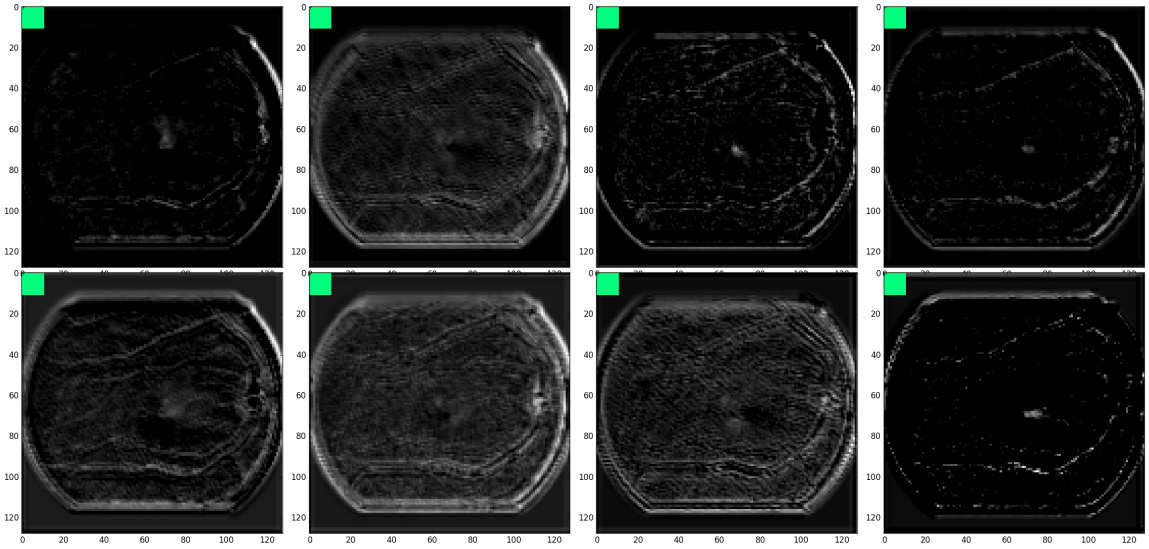


Figure 7.6: Activations of the fifth layer. The green square on the top left of the images represents the size of the receptive field at this layer.

which are the darkest structures of the original image, along with the macula, can still be seen as well. The output of the network is shown in Fig. 7.9: the non-zero values are in the right place but the segmentation is clearly not perfect.

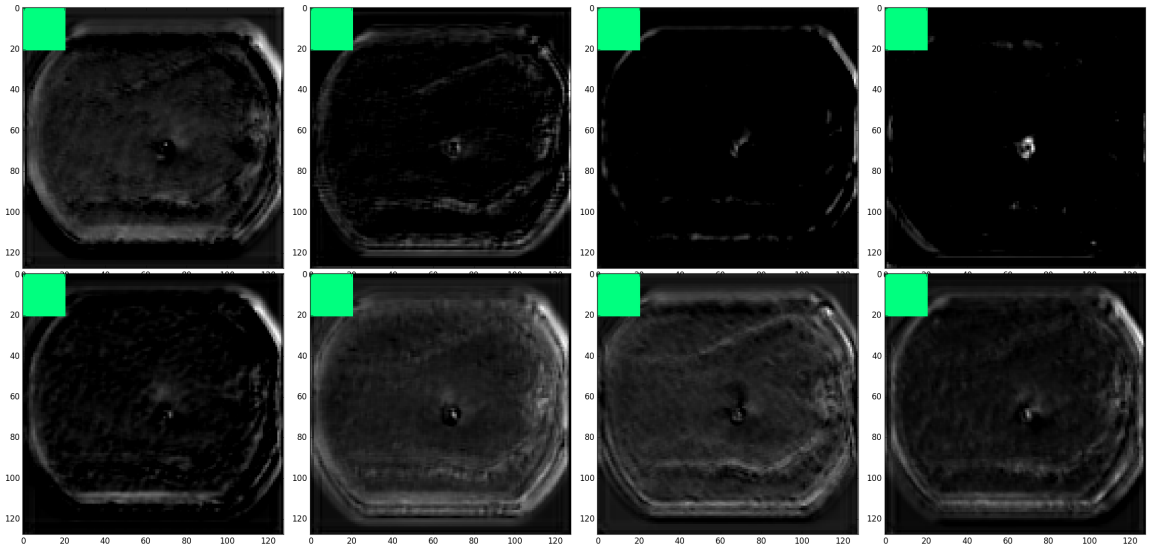


Figure 7.7: Activations of the tenth layer. The green square on the top left of the images represents the size of the receptive field at this layer.

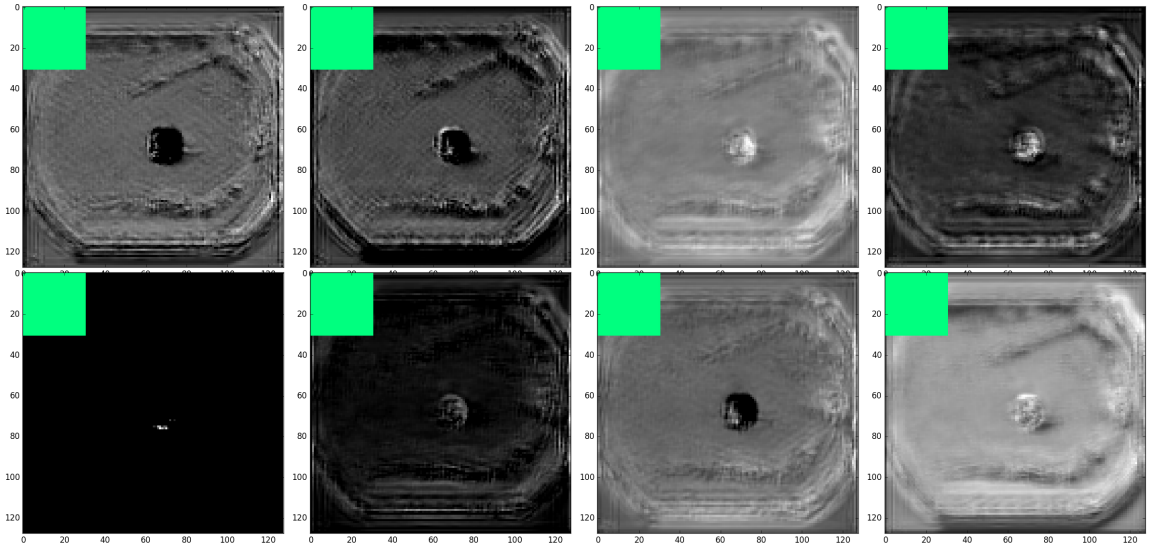


Figure 7.8: Activations of the fifteenth layer (second-to-last). The green square on the top left of the images represents the size of the receptive field at this layer.

7.4.4 Network Output Post-processing

Given an input 128x128 gray-scale image, the network outputs a 128x128 nonnegative array. A rectified linear unit is used in the output layer. We did also experiment using a sigmoid output activation, along with logistic loss, but it turned out that network convergence was harder to reach, and the

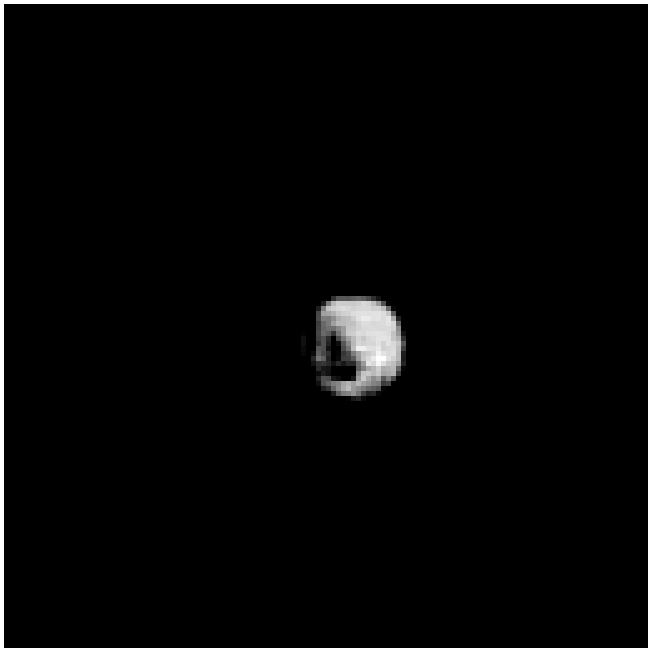


Figure 7.9: Output of the network.

obtained performance was lower.

Based on an output image, we have to answer the question "Is the macula visible in the original image ?". Since the ground truth image for a positive instance is a binary disk of radius 10, this is equivalent to answering the (ill-defined) question "Does the output image look like a binary disk of radius 10 ?".

The strategy we implemented goes as follows: the output image is thresholded above a value t in order to obtain a binary image. Then, only connected components of area greater than a value A are kept. If there is exactly one component remaining, we assume that the macula is visible, and the fovea can be located as the centroid of this component. If there is zero component of large enough area, we assume that the macula is not visible. If there are two or more components with area greater than A, this means that the algorithm is behaving oddly, and since our application is sensitivity-driven (it is important to correctly identify ungradable images), as a measure of precaution, we also consider that the macula is not clearly visible. Examples of network outputs and illustration of our post-processing are given in figure 7.10.

7.5 Results

7.5.1 Macula Visibility Estimation

Parameter Influence

There are two parameters to be set: the threshold t and minimum area A. In order to pick the best values for these parameters, we looked at specificity (the fraction of images where the macula was annotated as visible correctly classified), sensitivity (the fraction of images where nothing was annotated classified as such) and overall accuracy on the validation set. Doing so on the training set could lead to overfitting, while doing so on the test set would give a biased estimation of the algorithm's ability to generalize.

The specificity, sensitivity and accuracy curves are shown in Fig. 7.11. Unsurprisingly, the higher the threshold, the higher the sensitivity, but even taking $t = 0.5$ and $A = 1$ (i.e. demanding there

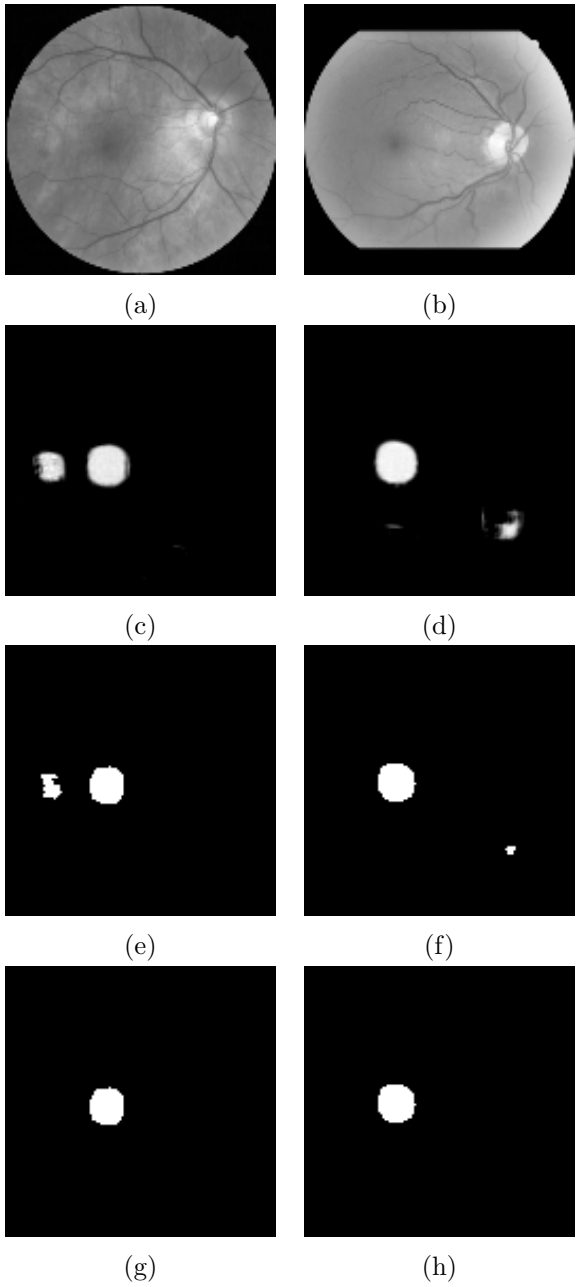
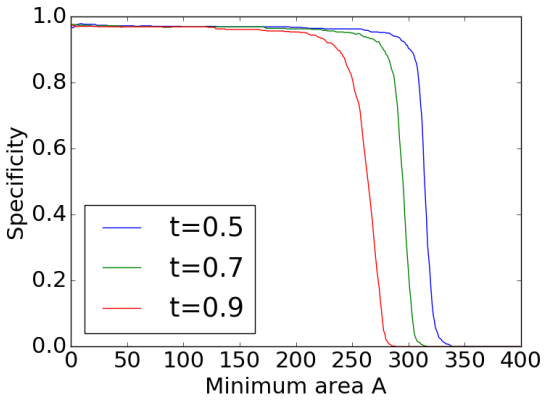
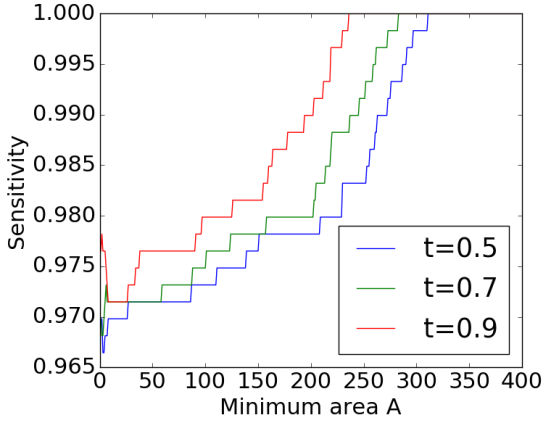


Figure 7.10: (a-b): input images, (c-d): corresponding network outputs, (e-f): thresholded network outputs ($t = 0.9$), (g-h): connected components of area greater than $A = 200$ pixels. These images belong to the test set and were chosen specifically because more than one component remained after thresholding, but this actually happens in only about 5% cases.

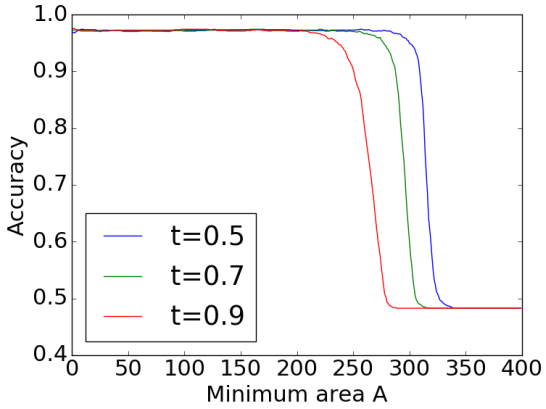
is only one connected component after thresholding), almost 97% images where macula visibility was considered insufficient by human readers are correctly classified. The main observation is that for $A \leq 200$, the choice of parameters actually has very little influence on the results. For the three considered thresholds, there is a steep fall when setting A above a certain value: in order to maximize sensitivity while keeping a certain margin to this critical value, we chose to set $t = 0.9$ and $A = 200$, which leads to a reasonable tradeoff, with 99% sensitivity, 95.3% specificity and 97.1% overall accuracy. We use those parameters in the following.



(a) Algorithm specificity on the validation dataset plotted against A for different threshold values.



(b) Algorithm sensitivity on the validation dataset plotted against A for different threshold values.



(c) Algorithm accuracy on the validation dataset plotted against A for different threshold values.

Figure 7.11: Influence of the threshold t and the minimum area A on the classification performance.

Test Set Results

Accuracy on the test set reaches 96.4%. In comparison, the agreement ratio between the two annotators before a consensus was made was only of 89.9%. Sensitivity reaches 98.7%; in other words, out of 304 images where the macula was not annotated in the ground truth, the algorithm makes 4 mistakes. The images on which these errors are made can be seen in Fig. 7.12. As can be seen on the figure, these correspond, if not to annotating errors, at least to borderline cases. In two out of the four images, the small vessels around the macula can even be distinguished.

Specificity reaches 94.2%: out of 310 images where the macula was annotated as clearly visible, the algorithm correctly classifies 292.

Performance on Pathological Images

Although we obtain a very good performance on our test set in terms of accuracy, sensitivity and specificity, it is interesting to look more specifically at images where automatic macula detection would be expected to be hard. Figures 7.13, 7.14, 7.15, 7.16 show examples of network outputs for images where lesions are present. Although there is one case of unsuccessful detection, we can see that the network is able to detect the macula even when there are hemorrhages or exudates in the macular

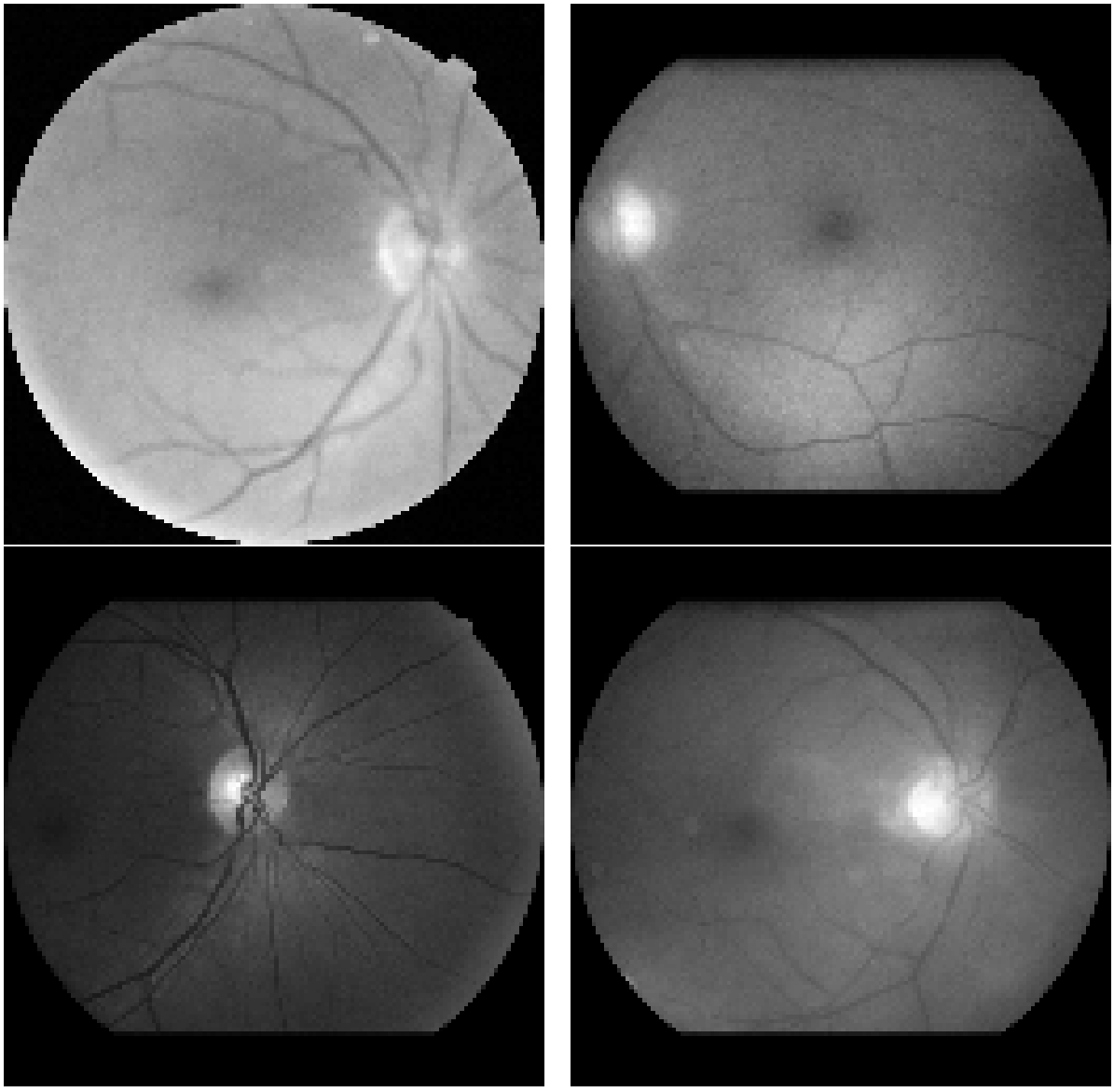
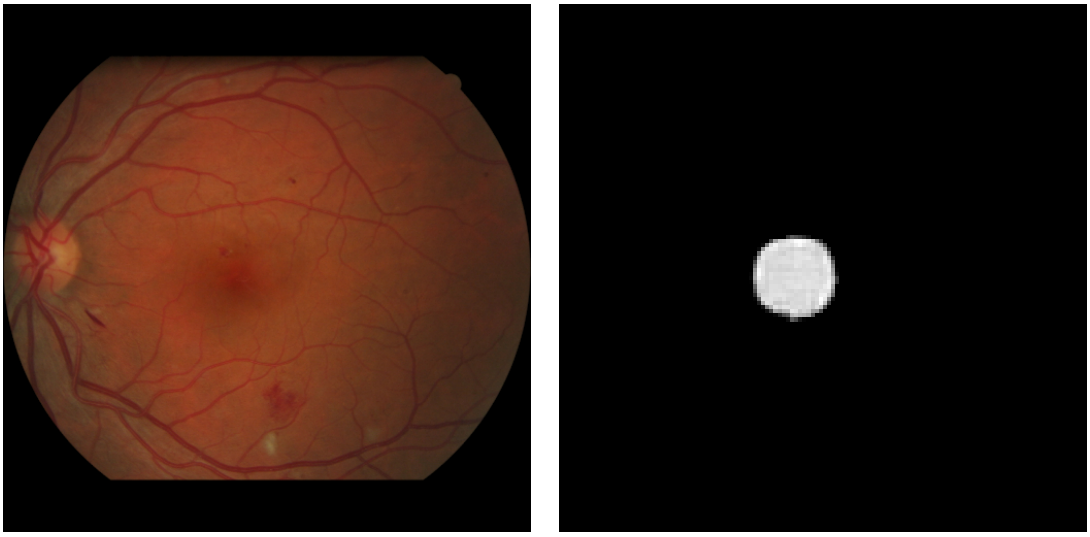


Figure 7.12: The four test set images for which the algorithm incorrectly predicts that the zone of the macula is of good quality.

zone.

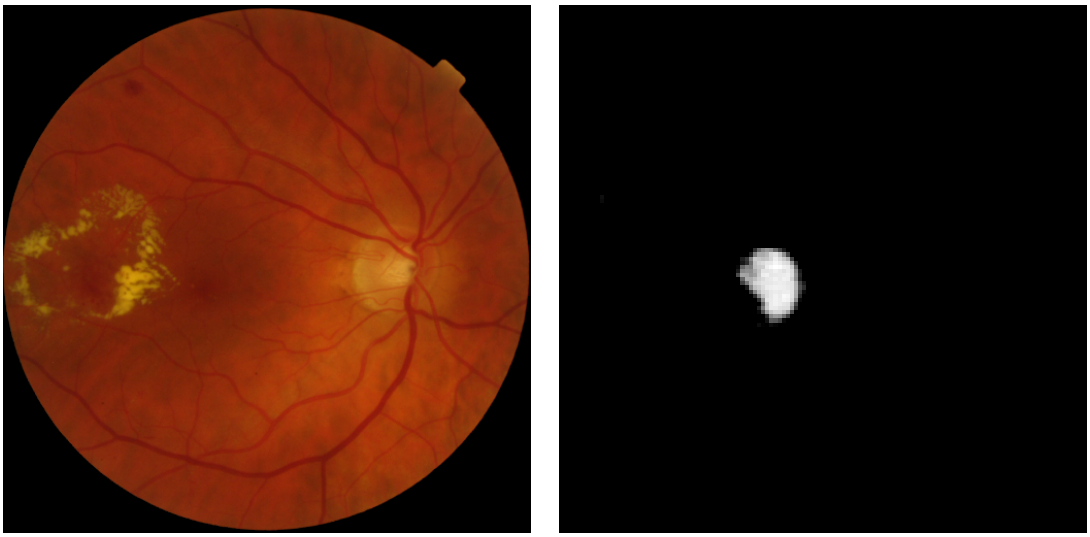
U-Net Comparison

U-Net [RFB15] is a popular network for segmentation and has been successfully used in a variety of applications. It makes perfect sense trying to apply it to our problem, however we have found it to slightly underperform compared to our network, in terms of both specificity and sensitivity. Several configurations of U-Net were tested. The best one had 4 filters in the initial convolutional layer, and a gaussian noise layer at the end. On the test set, this network incorrectly predicts that the macula is not visible on 19 images, which is similar to our network (18 false negatives). The main drawback is that it has more than twice as many false positives (9 versus 4). It also has many more parameters (122,953) than our network. As for fovea localization, which is detailed in the next section, the average error for U-Net is 1.22 pixels, which is again more than our network's error (0.95 pixel).



(a) (b)

Figure 7.13: Successful detection despite hemorrhages in the macular region



(a) (b)

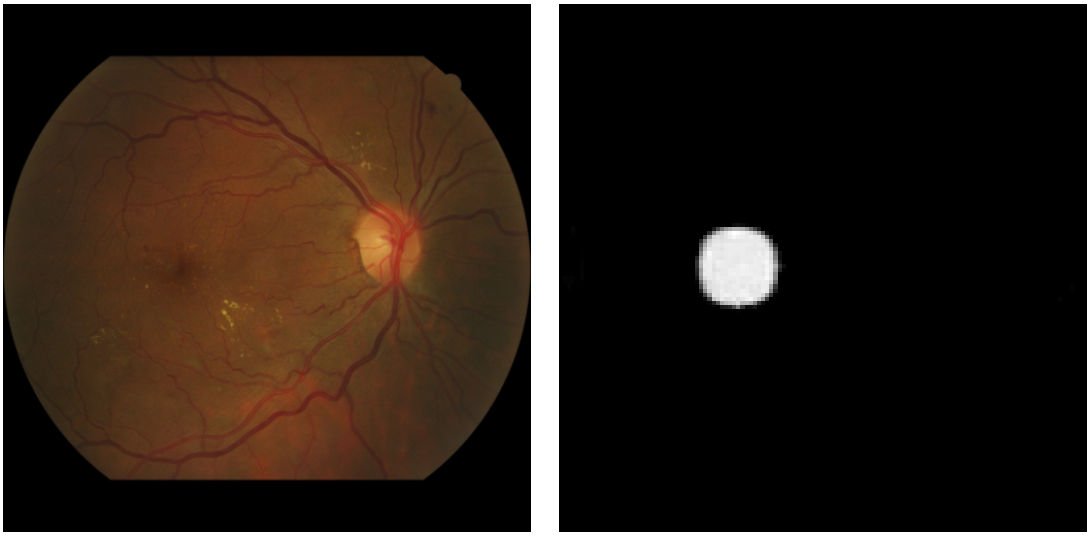
Figure 7.14: Partial detection despite exudates near the macula

7.5.2 Fovea Localization Results

Although the main task our algorithm addresses is assessing the quality of the macula region, it can also be used to segment the macula, or localize the fovea, as mentioned in section 7.4.4. In this section, we evaluate the performance of our algorithm for localizing the fovea, on the database we extracted from the e-ophtha database and on the ARIA database.

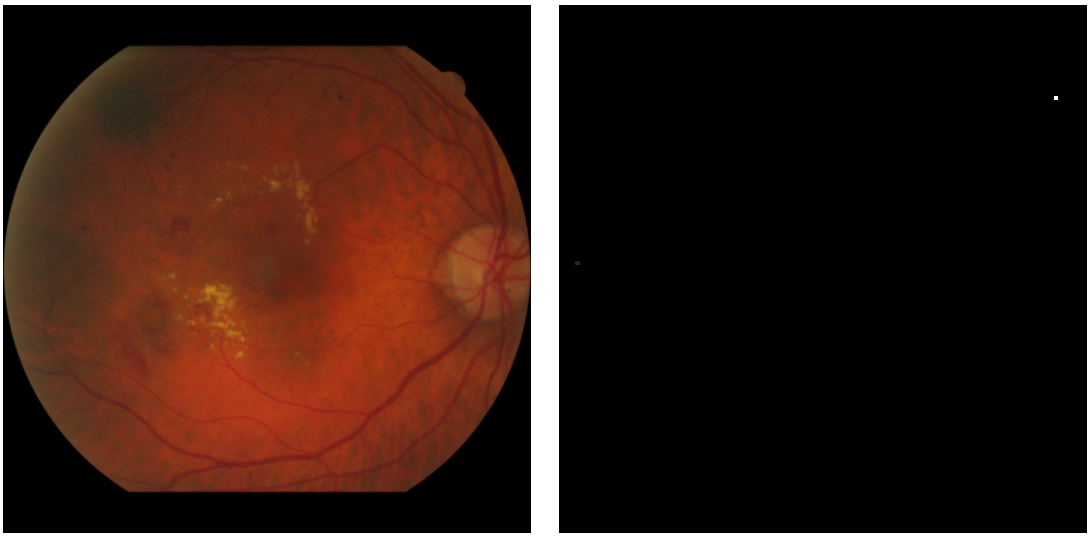
e-ophtha Database

As mentioned in the previous section, if we use $t = 0.9$ and $A = 200$, our algorithm predicts 292 images where the macula is visible, out of the 310 images where it was annotated. By lowering either parameter or defining another strategy, we could predict a location for the macula for more images, but it would make little sense in this context trying to localize it if we are not even confident it is in



(a) (b)

Figure 7.15: Successful detection despite exudates near the macula



(a) (b)

Figure 7.16: Detection failure. In addition to exudates and hemorrhages near the macula, there seems to be a reflexion in the center of the image.

the FOV. In order to remain consistent, we leave the parameters unchanged and present localization results only on the images identified by our algorithm.

As previously mentioned in section 7.4.4, we use the centroid of the (only) connected component of the thresholded output image as our estimation for fovea location. The mean of the two annotators' positions is used as ground truth. The histogram of distances to the ground truth is shown in Fig. 7.17.

The average test error is 0.95 pixel, which is less than the average distance between the annotators. One pixel in 128x128 resolution represents 0.075 mm. The largest test error is 4.85 pixels, or 0.33 mm; it has to be noted that, since the macula was considered to be 10 pixels in radius, all of the predicted values lie within the macula.

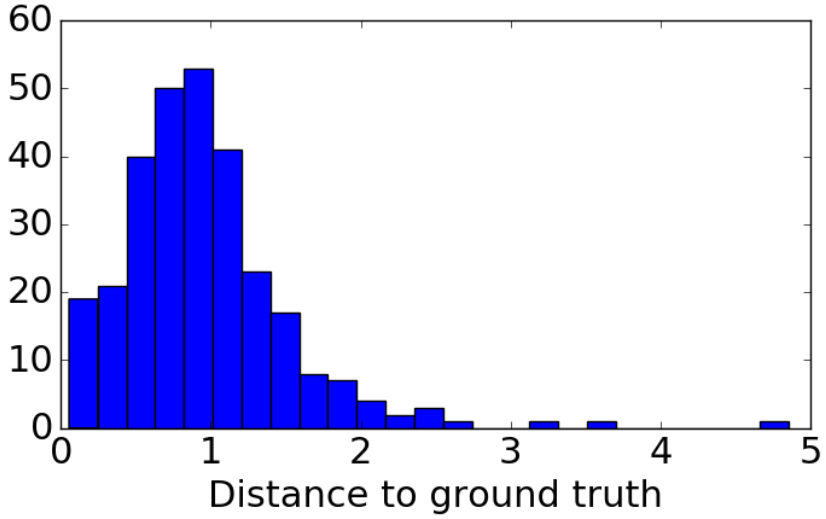


Figure 7.17: Histogram of distances to ground truth on the test base (in pixels after resizing to 128x128).

ARIA Database C

The ARIA database C contains 61 central eye fundus images with corresponding ground truth annotations for optic disk and fovea. Our network, with the same post-processing parameters as before, predicts a position for 46 of them, with a mean error of 1.4 pixel (0.1 mm) and a maximum error of 6 pixels (0.44 mm). The average error is again comparable to the average distance between two human readers (1.25 pixels, as mentioned in Sec. 7.3) and the maximum error is about half the macula's radius.

The 15 remaining images correspond to poor quality acquisitions; some examples can be seen in Fig. 7.18. Although it is quite easy for an experimented human observer to approximately locate the fovea region, the macular regions of these images are clearly of limited to no interest for an ophthalmologist.

7.6 Discussion

Guidelines for teleophthalmology recommend taking two photographs per eye, one centered on the macula, the other centered on the optic disk. A mandatory condition for a couple of images to be gradable by an ophthalmologist is that the macula is clearly visible in at least one of the two images. In practice, a significant proportion of examinations - about 10% for the OPHDIAT network - does not meet this required quality criterion. The algorithm detailed in the present work could significantly reduce the fraction of ungradable central images. Since the algorithm also provides the location of the macula, it can also be used in order to assert that a central image is indeed well centered around it. It can also be used as part of an automated diagnosis algorithm, for which macula segmentation is often a crucial step [Zha14].

Eye fundus photographs so far are often made by healthcare professionals, using tabletop non-mydriatic cameras; however, recent years have seen the emergence of portable, handheld retinographs, cheaper and allowing for screening in remote locations. These generally produce images of lower quality than tabletop retinographs. The algorithm described in the present work is very lightweight: it requires very little memory for storage, is very fast, and consumes little energy, compared to the much larger state-of-the-art networks. It can easily be integrated in an embedded device, telling in

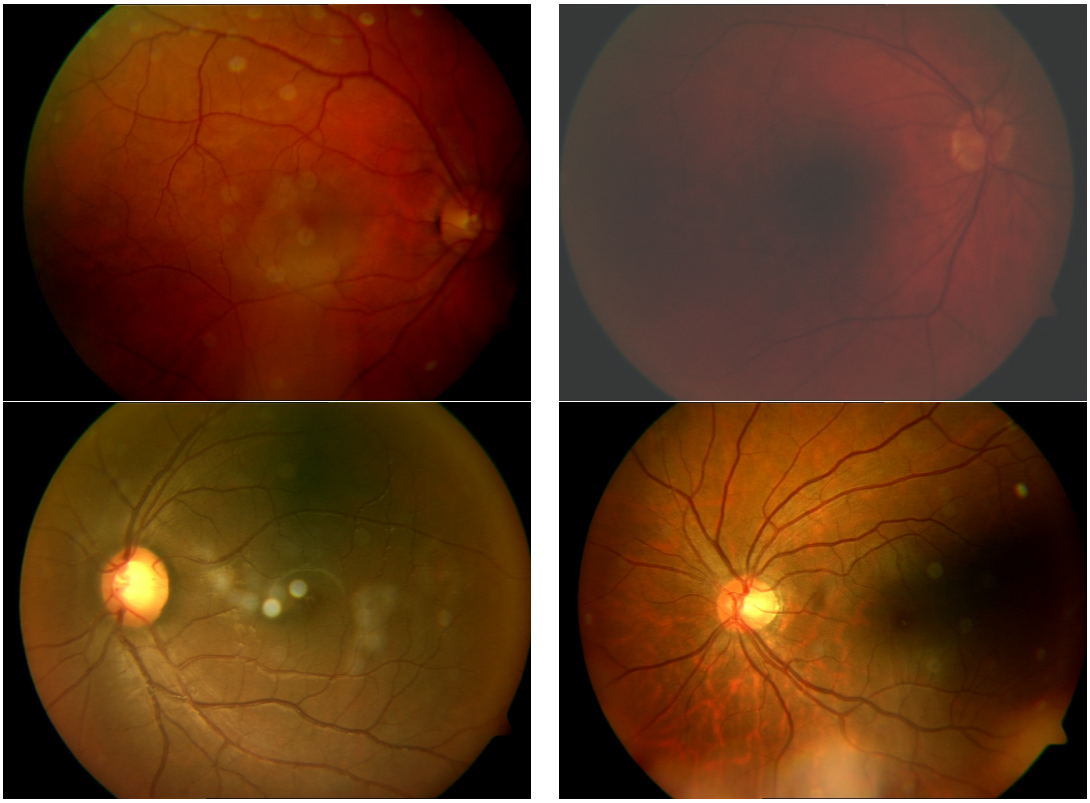


Figure 7.18: Examples of images from the ARIA database for which the network does not return a fovea position.

real time the operator whether another acquisition should be made.

7.7 Conclusion

In this chapter, we use a fully-convolutional network to segment the macular region, in order to assess the quality of eye fundus images. Our algorithm is also able to provide fovea localization, within 0.1 mm of human performance in average, in the case image quality is deemed sufficient.

Although macula visibility is not the only requirement for a retinal image to be gradable for DR, the method we propose is able to significantly reduce the number of ungradable images sent to medical experts in teleophthalmology networks, saving both patient and physician time, when the number of ophthalmologists is insufficient for current needs and the diabetic population is expected to grow much faster than that of the profession. It can also be combined with other quality criteria, such as contrast and sharpness, to build a complete quality assessment system.

Conclusion

Cette thèse s'articule autour de la localisation d'objet et de la segmentation. Il s'agit d'une tâche courante en traitement d'image, mais il n'existe pas de méthode universelle ; ces dernières années, les réseaux de neurones convolutionnels sont devenus l'outil le plus populaire pour ces tâches, et dans de nombreux cas, ils ont obtenu de meilleures performances que les techniques précédentes de l'état de l'art. Cependant, ces méthodes manquent d'interprétabilité et nécessitent des bases de données de grande taille pour l'apprentissage. Dans ce manuscrit, nous avons approché de problème de détection d'objet de plusieurs manières, en fonction du nombre de données annotées disponibles et des contraintes des projets, comme l'interprétabilité ou le besoin de caractériser les objets une fois segmentés.

Dans la première partie, consacrée au projet ATHENA, l'algorithme de détection de défauts que nous avons créé est dans l'immédiat l'outil le plus utile pour les utilisateurs finaux, mais il pourrait s'avérer inutilement complexe si le processus d'acquisition des images était amélioré. Les images sur lesquelles nous avons travaillé au début du projet étaient souvent bruitées ou peu contrastées, et une grande partie de l'information était perdue à cause de la compression JPEG — qui par ailleurs causait des artefacts de saturation. Sur les acquisitions plus tardives, sauvegardées en TIFF 16 bits, la différence de niveaux de gris entre les défauts et le bruit de fond était tellement importante qu'un simple seuillage au dessus et en dessous de valeurs bien choisies serait sans doute suffisante pour obtenir une bonne segmentation.

Cependant, même avec une acquisition idéale, du bruit subsisterait, à cause de la rugosité, de variations locales du facteur d'absorption ou de l'émissivité, ou à cause de rayures superficielles : notre estimation du niveau de bruit global reste pertinente. La méthode de seuillages à 3σ successifs que nous proposons avant d'estimer l'écart-type pourrait s'avérer utile dans d'autres applications, bien qu'une analyse mathématique plus poussée serait sans doute nécessaire (nous avons prouvé que la procédure converge, mais pour certaines distributions, la valeur finale pourrait ne pas être une estimation pertinente du niveau de bruit). Dans le cas particulier de l'acquisition par thermographie aller-retour, nous avons aussi donné des définitions mathématiques de la symétrie et du rapport signal/bruit des défauts, valeurs jusqu'ici dépendantes de l'opérateur, ce qui harmonise les rapports de détection.

Dans la deuxième partie, consacrée au projet RetinOptic, nous avons utilisé plusieurs types de réseaux de neurones convolutionnels afin de détecter la macula sur des images de fond d'œil. Il s'avère qu'un aspect important du problème consiste à trouver une bonne manière de poser le problème. La première approche a consisté à séparer la tâche en deux : une tâche de classification pour déterminer si la macula était présente, et une tâche de régression pour la localiser. Bien que nous obtenions une bonne localisation dans la majorité des cas avec cette stratégie, il est difficile de détecter d'éventuelles erreurs. Le but consistait à proposer un algorithme pouvant être intégré dans un rétinographe portable : en estimant la qualité de l'image au moment de l'acquisition, le système peut alerter l'opérateur si la qualité est insuffisante ; en localisant la fovéa, il permet de s'assurer qu'au moins une image par œil est correctement centrée sur la macula. En reformulant notre tâche comme un problème de segmentation, nous avons pu utiliser des réseaux complètement convolutionnels, lesquels fournissent à la fois un score de qualité et la localisation demandée. De plus, ces réseaux ont très peu de paramètres et sont très

rapides, ce qui les rend idéaux pour une intégration dans des systèmes embarqués.

Ces dernières années, de nombreux articles de recherche se sont concentrés sur le diagnostic automatique de la rétinopathie diabétique, avec des résultats prometteurs ; cependant, il reste plusieurs limitations : il existe peu de bases de données publiques, et les plus populaires ne contiennent que des images centrales de bonne qualité. L'idée d'utiliser à terme ces algorithmes, non pas pour aider, mais pour remplacer les médecins, soulève également des questions légales et morales qui sortent du cadre de cette thèse. Dans l'immédiat, le défaut le plus problématique concerne la gestion des pathologies graves rares. Il est sans doute possible d'obtenir des performances proches de l'humain pour détecter les maladies les plus communes (comme la rétinopathie diabétique, mais aussi la dégénérescence maculaire liée à l'âge ou l'œdème papillaire), mais les cas de mélanome ou d'autres conditions rares sont trop peu présents dans les bases de données, et un mauvais diganostic dû à un faux positif a des conséquences sévères.

L'évaluation de la qualité des images que nous proposons n'est pas un outil diagnostic, mais aide les opthalmologues de manière indirecte. Lorsqu'une acquisition est jugée ininterprétable, le patient doit consulter un spécialiste, ce qui annule l'intérêt d'un réseau de dépistage ; le fait de s'assurer de la qualité des images réduit le nombre de consultations inutiles. Notre solution peut être améliorée en y intégrant d'autres critères de qualité, comme l'estimation du flou ou du contraste, ou d'autres critères spécifiques aux images rétiniennes. Un avantage de se concentrer sur la macula dans ce manuscrit est le fait que l'annotation peut être faite relativement rapidement et par des non experts. Cependant, une base de données comprenant à la fois des images centrales et nasales, dont la qualité serait jugée par des ophtalmologues, serait utile.

Bien que n'étant pas directement reliés à la segmentation, les décompositions en pics, les nouvelles valeurs d'extinction et les opérateurs présentés dans ce manuscrit ouvrent de nouvelles possibilités. En particulier, puisque ces nouveaux opérateurs sont connectés, ils peuvent être utilisés préalablement à un algorithme de segmentation sans créer de nouveaux contours. Le choix des marqueurs dans plusieurs algorithmes de segmentation de type ligne de partage de eaux ou waterpixels, est souvent basée sur les valeurs d'extinction des minima du gradient ; les nouveaux attributs que nous avons définis peuvent servir de critères de sélection alternatifs.

Conclusion

This thesis is centered around object localization and segmentation. This is a common task in image processing, but there is no universal algorithm or method for it; in recent years, convolutional neural networks have become, by far, the most popular tool for these tasks as well as for other image processing problems, and in many cases, they have been shown to outperform former state-of-the-art techniques. However, these methods have the main drawbacks of lacking interpretability, and of requiring large enough datasets for training. In the present work, we tackled this problem of object detection in different ways, depending both on the availability of annotated data and specific requirements of real-life applied problems, like interpretability or the need for further characterization of segmented objects.

In the first part, concerning the ATHENA project, the automatic defect detection algorithm we devised is the most immediately useful tool for end users of this technology, but it might actually become unnecessarily complex if the acquisition process is improved. The images we worked with at the beginning of the project were often noisy or poorly contrasted, and a lot of information was lost when saving them in JPEG format — which additionally caused saturation artifacts. On later acquisitions, saved in 16-bit TIFF, the difference in grayscale values between defects and background noise was so huge that a simple thresholding over and under well-chosen values would likely be enough to provide a good segmentation.

Even with an ideal acquisition technique, though, some noise is bound to remain, due to rugosity, local variations in absorptance and emissivity, or superficial scratches: our global noise estimation should remain relevant. The simple iterative 3σ thresholding method we propose prior to standard deviation estimation could likewise be useful in a variety of applications, although it would benefit from a more thorough mathematical analysis (we have proven that the procedure eventually reaches convergence, but for some distributions, the estimated value could be meaningless). In the specific case of the two-scan thermography acquisition, we also provided mathematical definitions of defect symmetry and signal/noise ratio that were so far operator-dependent, which harmonizes detection reports.

In the second part, concerning the RetinOptic project, we used several kinds of convolutional neural networks for detecting the macula on eye fundus images. An important aspect of this problem, it turns out, consists in finding a good way of formulating the problem. The first approach consisted in splitting the problem in two: a classification task to determine if the macula was present, and a regression task to locate it. Although we can achieve a good localization with this strategy in a majority of cases, it is hard to detect possible mistakes, in either step. The aim was to propose an algorithm to be integrated in a portable retinograph: by estimating image quality at the time of acquisition, the system can warn the operator if the quality is insufficient, and by localizing the fovea, it could also assess that there is at least one image per eye correctly centered on the macula. By reformulating our task as a segmentation problem, we were able to use fully-convolutional networks, which provide both a quality score and the desired localization. Additionally, these networks have very few parameters and are very fast, making them ideal for integration in embedded systems.

In recent years, many research articles have focused on automated diagnosis of diabetic retinopathy,

with promising results; however, there remain several limitations: there are few publicly available databases for this task, and the most popular ones only contain central images of good quality. The idea of eventually using these algorithms, not as an aid, but as a replacement for practicians, also raises legal and moral issues that are beyond the scope of this thesis. An objective problematic drawback of these methods, for now, is their handling of rare, severe cases. Obtaining close to human performance for detecting the most common diseases (not only diabetic retinopathy, but also age-related macular degeneration or papillar edema) is likely achievable, but instances of melanoma or other rare conditions are too few, and a false negative has dire consequences.

The image quality evaluation we propose is not a diagnosis tool, but helps ophthalmologists in an indirect way. When an acquisition is deemed ungradable, the patient has to be referred to a physician, defeating the purpose of a screening network; assessing image quality reduces the number of unnecessary appointments. Our solution can be improved by integrating other quality criteria, either general, like blur estimation or contrast, or specific to retinal images. An advantage of focusing on the macula in this work is that the annotation can be done relatively quickly and by non-experts. However, a database containing both central and nasal images, with a ground truth quality provided by ophthalmologists, would be a valuable resource.

Although not directly related to segmentation, the decompositions in peaks, novel extinction values and operators introduced in this work open new possibilities. In particular, since these new operators are connected, they can be applied before a segmentation algorithm without creating new contours. The choice of markers in several segmentation algorithms, like watershed or waterpixels, is often based on extinction values of the gradient's minima; the new features we defined can be used as alternative selection criteria.

Bibliography

[Agu+14] Carla Agurto, Victor Murray, Honggang Yu, Jeffrey Wigdahl, Marios Pattichis, Sheila Nemeth, E. Simon Barriga, and Peter Soliz. "A Multiscale Optimization Approach to Detect Exudates in the Macula". In: *IEEE Journal of Biomedical and Health Informatics* 18.4 (July 2014), pp. 1328–1336. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2013.2296399.

[Ala+17a] Robin Alais, Petr Dokládal, Etienne Decencière, and Bruno Figliuzzi. "Automatic Detection of Cracks and Delaminations in Thermal Images". In: *IXth Workshop "NDT in Progress"*. Prague, 2017, pp. 2–8. ISBN: 978-80-87012-63-5.

[Ala+17b] Robin Alais, Petr Dokládal, Etienne Decencière, and Bruno Figliuzzi. "Function Decomposition in Main and Lesser Peaks". In: *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*. Springer. 2017, pp. 319–330.

[Ala+20] Robin Alais, Petr Dokládal, Ali Erginay, Bruno Figliuzzi, and Etienne Decencière. "Fast macula detection and application to retinal image quality assessment". In: *Biomedical Signal Processing and Control* 55 (2020), p. 101567.

[AlB+18] Baidaa Al-Bander, Waleed Al-Nuaimy, Bryan M. Williams, and Yalin Zheng. "Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc". In: *Biomedical Signal Processing and Control* 40 (Feb. 2018), pp. 91–101. ISSN: 17468094. DOI: 10.1016/j.bspc.2017.09.008.

[Ame12] American Diabetes Association. "Executive Summary: Standards of Medical Care in Diabetes–2012". In: *Diabetes Care* 35.Supplement_1 (Jan. 2012), S4–S10. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/dc12-s004.

[AS03] Jesus Angulo and Jean Serra. "Automatic analysis of DNA microarray images using mathematical morphology". In: *Bioinformatics* 19.5 (Mar. 2003), pp. 553–562. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btg057.

[Beu01] Serge Beucher. "Geodesic reconstruction, saddle zones and hierarchical segmentation". In: *Image Analysis & Stereology* 20.2 (2001), pp. 137–141.

[Bou+08] Marie Carole Boucher, Gilles Desroches, Raul Garcia-Salinas, Amin Kherani, David Maberley, Sébastien Olivier, Mila Oh, and Frank Stockl. "Teleophthalmology screening for diabetic retinopathy through mobile imaging units within Canada". In: *Canadian Journal of Ophthalmology / Journal Canadien d'Ophtalmologie* 43.6 (Dec. 2008), pp. 658–668. ISSN: 00084182. DOI: 10.3129/i08-120.

[CC03] Johanna Choremis and David R. Chow. "Use of telemedicine in screening for diabetic retinopathy". In: *Canadian Journal of Ophthalmology / Journal Canadien d'Ophtalmologie* 38.7 (2003), pp. 575–579. ISSN: 0008-4182. DOI: https://doi.org/10.1016/S0008-4182(03)80111-4.

[Che67] Pafnutii Lvovich Chebyshev. “Des valeurs moyennes, Liouville’s”. In: *J. Math. Pures Appl.* 12 (1867), pp. 177–184.

[Dam06] F Damian. *Aria online, retinal image archive*. 2006.

[Dec+13] Etienne Decencière, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, Jean-Claude Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, Damien Elie, Pascale Massin, Zeynep Viktor, Ali Erginay, Bruno Laÿ, and Agnès Chabouis. “TeleOphta: Machine learning and image processing methods for teleophthalmology”. In: *IRBM* 34.2 (Apr. 2013), pp. 196–203. ISSN: 19590318. DOI: 10.1016/j.irbm.2013.01.010.

[Dec+14] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein. “Feedback on a publicly distributed image database: the Messidor database”. In: *Image Analysis & Stereology* 33.3 (Aug. 2014), p. 231. ISSN: 1854-5165, 1580-3139. DOI: 10.5566/ias.1155.

[DMR09] Cecile Delcourt, Pascale Massin, and Myriam Rosilio. “Epidemiology of diabetic retinopathy: expected vs reported prevalence of cases in the French population”. In: *Diabetes & metabolism* 35.6 (2009), pp. 431–438.

[DOC12] João Miguel Pires Dias, Carlos Manta Oliveira, and Luís A. da Silva Cruz. “Evaluation of Retinal Image Gradability by Image Features Classification”. In: *Procedia Technology* 5 (2012), pp. 865–875. ISSN: 22120173. DOI: 10.1016/j.protcy.2012.09.096.

[EZ16] Andreas Ebneter and Martin S. Zinkernagel. “Novelties in Diabetic Retinopathy”. In: *Endocrine Development*. Ed. by C. Stettler, E. Christ, and P. Diem. Vol. 31. S. Karger AG, Jan. 2016, pp. 84–96. ISBN: 978-3-318-05638-9 978-3-318-05639-6. DOI: 10.1159/000439391.

[Fle+12] Alan D. Fleming, Sam Philip, Keith A. Goatman, Peter F. Sharp, and John A. Olson. “Automated clarity assessment of retinal images using regionally based structural and statistical measures”. In: *Medical Engineering & Physics* 34.7 (Sept. 2012), pp. 849–859. ISSN: 13504533. DOI: 10.1016/j.medengphy.2011.09.027.

[Gar+96] G G Gardner, D Keating, T H Williamson, and A T Elliott. “Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool.” In: *British Journal of Ophthalmology* 80.11 (Nov. 1996), pp. 940–944. ISSN: 0007-1161. DOI: 10.1136/bjo.80.11.940.

[GB92] C Gruss and D Balageas. “Theoretical and experimental applications of the flying spot camera”. In: *ONERA TP* (1992).

[Gia+08] Luca Giancardo, Michael D. Abràmoff, E. Chaum, T. P. Karnowski, F. Meriaudeau, and K. W. Tobin. “Elliptical local vessel density: a fast and robust quality metric for retinal images”. In: *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 3534–3537.

[Gia+10] Luca Giancardo, Fabrice Meriaudeau, Thomas P. Karnowski, Edward Chaum, and Kenneth Tobin. “Quality assessment of retinal fundus images using elliptical local vessel density”. In: *New developments in biomedical engineering*. InTech, 2010.

[GLB93] C Gruss, F Lepoutre, and D Balageas. “Nondestructive evaluation using a flying-spot camera”. In: *ONERA TP* 1 (1993).

[Gra15] Ben Graham. *Kaggle Diabetic Retinopathy Detection competition report*. Tech. rep. Coventry, UK: Department of Statistics and Centre for Complexity Science, 2015.

[Gri92] Michel Grimaud. "New measure of contrast: the dynamics". In: *San Diego'92*. International Society for Optics and Photonics. 1992, pp. 292–305.

[Gul+16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22 (2016), pp. 2402–2410.

[Gup+14] Garima Gupta, Keerthi Ram, S. Kulasekaran, Niranjan Joshi, Mohanasankar Sivaprakasam, and Rashmin Gandhi. "Detection of retinal hemorrhages in the presence of blood vessels". In: *Proceedings of the Ophthalmic Medical Image Analysis First International Workshop*. Boston, Massachusetts, 2014, pp. 105–112.

[H+73] Robert M Haralick, Karthikeyan Shanmugam, et al. "Textural features for image classification". In: *IEEE Transactions on systems, man, and cybernetics* 6 (1973), pp. 610–621.

[Har+94] J Hartikainen, R Lehtiniemi, J Rantala, J Varis, and M Luukkala. "Fast infrared line-scanning method and its applications". In: *Review of Progress in Quantitative Nondestructive Evaluation* 13 (1994), pp. 401–401.

[HSS12] G Hinton, N Srivastava, and K Swersky. "RMSProp: Divide the gradient by a running average of its recent magnitude". In: *Neural networks for machine learning, Coursera lecture 6e* (2012).

[Hun+11] Andrew Hunter, James A. Lowell, Maged Habib, Bob Ryder, Ansu Basu, and David Steel. "An automated retinal image quality grading algorithm". In: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 5955–5958.

[Kau+87] Irving Kaufman, Pan-Tze Chang, Hsueh-Shun Hsu, Wen-Yuan Huang, and Daw-Yang Shyong. "Photothermal radiometric detection and imaging of surface cracks". In: *Journal of Nondestructive Evaluation* 6.2 (June 1987), pp. 87–100. ISSN: 1573-4862. DOI: 10.1007/BF00568887.

[KGU13] Nutnaree Kleawsirikul, Smith Gulati, and Bunyarit Uyyanonvara. "Automated Retinal Hemorrhage Detection Using Morphological Top Hat and Rule-based Classification". In: *3rd International Conference on Intelligent Computational Systems (ICICS 2013)*. 2013, pp. 39–43.

[Kra+98] J.-C. Krapez, L. Legrandjacques, F. Lepoutre, and D.L. Balageas. "Optimization of the photothermal camera for crack detection". In: *Proceedings of the 1998 International Conference on Quantitative InfraRed Thermography*. QIRT Council, 1998. DOI: 10.21611/qirt.1998.048.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[Kub68] Edward J Kubiak. "Infrared detection of fatigue cracks and other near-surface defects". In: *Applied Optics* 7.9 (1968), pp. 1743–1747.

[Leh+92] R. Lehtiniemi, J. Hartikainen, J. Varis, and M. Luukkala. "Induction Heating as a Selective Heat Source in Fast Thermal Non-Destructive Evaluation". In: *Photoacoustic and Photothermal Phenomena III*. Ed. by Dane Bićanić. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 512–515. ISBN: 978-3-540-47269-8.

[Leh+94] Lehtiniemi, R., Rantala, J., Hartikainen, J., Varis, J., and Luukkala, M. "Performance of a line-scanning thermal nondestructive evaluation system applied to low-diffusivity materials". In: *J. Phys. IV France* 04.C7 (1994), pp. C7–587–C7–590. DOI: 10.1051/jp4:19947138.

[LGB01] Marc Lalonde, Langis Gagnon, and Marie-Carole Boucher. "Automatic visual quality assessment in optical fundus images". In: *Proceedings of vision interface*. Vol. 32. Ottawa, 2001, pp. 259–264.

[Lu+17] Zongqing Lu, Swati Rallapalli, Kevin Chan, and Thomas La Porta. "Modeling the Resource Requirements of Convolutional Neural Networks on Mobile Devices". In: *Proceedings of the 2017 ACM on Multimedia Conference - MM '17* (2017). arXiv: 1709.09503, pp. 1663–1671. DOI: 10.1145/3123266.3123389.

[LW99] S Lee and Yiming Wang. "Automatic retinal image quality assessment and enhancement". In: *Proceedings of SPIE Image Processing*. Vol. 3661. 1999, pp. 1581–1590.

[Mac+15] Vaia Machairas, Matthieu Faessel, David Cardenas-Pena, Théodore Chabardes, Thomas Walter, and Etienne Decencière. "Waterpixels". In: *IEEE Transactions on Image Processing* 24.11 (Nov. 2015), pp. 3707–3716. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2015.2451011.

[Mah+16a] Dwarikanath Mahapatra, Pallab K. Roy, Suman Sedai, and Rahil Garnavi. "A CNN based neurobiology inspired approach for retinal image quality assessment". In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 1304–1307.

[Mah+16b] Dwarikanath Mahapatra, Pallab K. Roy, Suman Sedai, and Rahil Garnavi. "Retinal Image Quality Classification Using Saliency Maps and CNNs". In: *Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings*. Ed. by Li Wang, Ehsan Adeli, Qian Wang, Yinghuan Shi, and Heung-Il Suk. DOI: 10.1007/978-3-319-47157-0_21. Cham: Springer International Publishing, 2016, pp. 172–179. ISBN: 978-3-319-47157-0.

[Man+16] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. "Deep retinal image understanding". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 140–148.

[Mas+08] P. Massin, A. Chabouis, A. Erginay, C. Viens-Bitker, A. Lecleire-Collet, T. Meas, P.-J. Guillausseau, G. Choupot, B. André, and P. Denormandie. "OPHDIAT©: A telemedical network screening system for diabetic retinopathy in the Île-de-France". In: *Diabetes & Metabolism* 34.3 (2008), pp. 227–234. ISSN: 1262-3636. DOI: https://doi.org/10.1016/j.diabet.2007.12.006.

[Mat+04] David R Matthews, Irene M Stratton, Stephen J Aldington, Rury R Holman, Eva M Kohner, and UK Prospective Diabetes Study Group. "Risks of progression of retinopathy and vision loss related to tight blood pressure control in type 2 diabetes mellitus: UKPDS 69". In: *Archives of ophthalmology (Chicago, Ill. : 1960)* 122.11 (Nov. 2004), pp. 1631–1640. ISSN: 0003-9950. DOI: 10.1001/archopht.122.11.1631.

[Mey04] Fernand Meyer. "Levelings, image simplification filters for segmentation". In: *Journal of Mathematical Imaging and Vision* 20.1-2 (2004), pp. 59–72.

[Mey98] Fernand Meyer. "From connected operators to levelings". In: *Computational Imaging and Vision* 12 (1998), pp. 191–198.

[NAG06] Meindert Niemeijer, Michael D. Abràmoff, and Bram van Ginneken. "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening". In: *Medical Image Analysis* 10.6 (Dec. 2006), pp. 888–898. ISSN: 13618415. DOI: 10.1016/j.media.2006.09.006.

[NAG09] Meindert Niemeijer, Michael D. Abràmoff, and Bram van Ginneken. "Fast detection of the optic disc and fovea in color fundus photographs". In: *Medical Image Analysis* 13.6 (Dec. 2009), pp. 859–870. ISSN: 13618415. DOI: 10.1016/j.media.2009.08.003.

[Ogu+17] K. Ogurtsova, J.D. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N.H. Cho, D. Cavan, J.E. Shaw, and L.E. Makaroff. "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040". In: *Diabetes Research and Clinical Practice* 128 (June 2017), pp. 40–50. ISSN: 01688227. DOI: 10.1016/j.diabres.2017.03.024.

[Pan+10] Baochuan Pang, Yi Zhang, Qianqing Chen, Zhifan Gao, Qinmu Peng, and Xinge You. "Cell nucleus segmentation in color histopathological imagery using convolutional networks". In: *Chinese conference on pattern recognition*. 2010, pp. 1–5.

[Pau+10] Jan Paulus, Jörg Meier, Rüdiger Bock, Joachim Hornegger, and Georg Michelson. "Automated quality assessment of retinal fundus photos". In: *International Journal of Computer Assisted Radiology and Surgery* 5.6 (Nov. 2010), pp. 557–564. ISSN: 1861-6410, 1861-6429. DOI: 10.1007/s11548-010-0479-7.

[Pet80] Y Petunin. "Justification of the three-sigma rule for unimodal distributions". In: *Theory of Probability & Mathematical Statistics* (1980).

[Pir+12] Ramon Pires, Herbert F. Jelinek, Jacques Wainer, and Anderson Rocha. "Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection". In: IEEE, Aug. 2012, pp. 229–236. ISBN: 978-0-7695-4829-6 978-1-4673-2802-9. DOI: 10.1109/SIBGRAPI.2012.39.

[POS14] João Miguel Pires Dias, Carlos Manta Oliveira, and Luís A. da Silva Cruz. "Retinal image quality assessment using generic image quality indicators". In: *Information Fusion* 19 (Sept. 2014), pp. 73–90. ISSN: 15662535. DOI: 10.1016/j.inffus.2012.08.001.

[Pra+16] Harry Pratt, Frans Coenen, Deborah M. Broadbent, Simon P. Harding, and Yalin Zheng. "Convolutional Neural Networks for Diabetic Retinopathy". In: *Procedia Computer Science* 90.Supplement C (2016). 20th Conference on Medical Image Understanding and Analysis (MIUA 2016), pp. 200–205. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.07.014.

[Que+17] Gwenolé Quellec, Katia Charrière, Yassine Boudi, Béatrice Cochener, and Mathieu Lamard. "Deep image mining for diabetic retinopathy screening". In: *Medical Image Analysis* 39 (July 2017), pp. 178–193. ISSN: 13618415. DOI: 10.1016/j.media.2017.04.012.

[Res+12] Serge Resnikoff, William Felch, Tina-Marie Gauthier, and Bruce Spivey. "The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners". In: *British Journal of Ophthalmology* 96.6 (June 2012), pp. 783–787. ISSN: 0007-1161, 1468-2079. DOI: 10.1136/bjophthalmol-2011-301378.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Vol. 9351. DOI: 10.1007/978-3-319-24574-4_28. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24573-7 978-3-319-24574-4.

[SG15a] K. S Sreejini and V. K Govindan. "A Review of Computer Aided Detection of Anatomical Structures and Lesions of DR from Color Retina Images". In: *International Journal of Image, Graphics and Signal Processing* 7.11 (Oct. 2015), pp. 55–69. ISSN: 20749074, 20749082. DOI: 10.5815/ijigsp.2015.11.08.

[SG15b] K. S Sreejini and V. K Govindan. "A Review of Computer Aided Detection of Anatomical Structures and Lesions of DR from Color Retina Images". In: *International Journal of Image, Graphics and Signal Processing* 7.11 (Oct. 2015), pp. 55–69. ISSN: 20749074, 20749082. DOI: 10.5815/ijigsp.2015.11.08.

[She97] Steven M Shepard. "Introduction to active thermography for non-destructive evaluation". In: *Anti-Corrosion Methods and Materials* 44.4 (1997), pp. 236–239.

[Siv+15] Jayanthi Sivaswamy, SR Krishnadas, Arunava Chakravarty, GD Joshi, A Syed Tabish, et al. "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis". In: *JSM Biomedical Imaging Data Papers* 2.1 (2015), p. 1004.

[Sjø+97] Anne Katrin Sjølie, Judith Stephenson, Steve Aldington, Eva Kohner, Hans Janka, Lynda Stevens, John Fuller, B. Karamanos, C. Tountas, and A. Kofinis. "Retinopathy and vision loss in insulin-dependent diabetes in Europe: the EURODIAB IDDM Complications Study". In: *Ophthalmology* 104.2 (1997), pp. 252–260.

[SOG98] Philippe Salembier, Albert Oliveras, and Luis Garrido. "Antiextensive connected operators for image and sequence processing". In: *Image Processing, IEEE Transactions on* 7.4 (1998), pp. 555–570.

[Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[Str+13a] M. Streza, D. Dadarlat, Y. Fedala, and S. Longuemart. "Depth estimation of surface cracks on metallic components by means of lock-in thermography". In: *Review of Scientific Instruments* 84.7 (July 2013), p. 074902. ISSN: 0034-6748, 1089-7623. DOI: 10.1063/1.4813744.

[Str+13b] M Streza, Y Fedala, J P Roger, G Tessier, and C Boue. "Heat transfer modeling for surface crack depth evaluation". In: *Measurement Science and Technology* 24.4 (Mar. 2013), p. 045602. DOI: 10.1088/0957-0233/24/4/045602.

[Sun+17] Jing Sun, Cheng Wan, Jun Cheng, Fengli Yu, and Jiang Liu. "Retinal Image Quality Classification Using Fine-Tuned CNN". In: *Fetal, Infant and Ophthalmic Medical Image Analysis*. Ed. by M. Jorge Cardoso, Tal Arbel, Andrew Melbourne, Hrvoje Bogunovic, Pim Moeskops, Xinjian Chen, Ernst Schwartz, Mona Garvin, Emma Robinson, Emanuele Trucco, Michael Ebner, Yanwu Xu, Antonios Makropoulos, Adrien Desjardin, and Tom Vercauteren. Vol. 10554. DOI: 10.1007/978-3-319-67561-9_14. Cham: Springer International Publishing, 2017, pp. 126–133. ISBN: 978-3-319-67560-2 978-3-319-67561-9.

[SV92] Jean Serra and Luc Vincent. "An overview of morphological filtering". In: *Circuits, Systems and Signal Processing* 11.1 (1992), pp. 47–108. ISSN: 1531-5878. DOI: 10.1007/BF01189221.

[SZ14] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014).

[Sze+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[Tan+13] Li Tang, Meindert Niemeijer, Joseph M. Reinhardt, Mona K. Garvin, and Michael D. Abramoff. "Splat Feature Classification With Application to Retinal Hemorrhage Detection in Fundus Images". In: *IEEE Transactions on Medical Imaging* 32.2 (Feb. 2013), pp. 364–375. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2012.2227119.

[Ten+16] Ruwan Tennakoon, Dwarikanath Mahapatra, Pallab Roy, Suman Sedai, and Rahil Garnavi. "Image Quality Classification for DR Screening Using Convolutional Neural Networks". In: University of Iowa, Oct. 2016, pp. 113–120. DOI: 10.17077/omia.1054.

[TH12] Tijmen Tieleman and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2 (2012).

[Thi17] Abdoulahad Thiam. "Contribution à l'étude et à l'optimisation du procédé de thermographie active appliquée à la détection de défauts surfaciques". PhD thesis. Université Bourgogne Franche-Comté, 2017.

[TWN15] Kevin Tozer, Maria A. Woodward, and Paula A. Newman-Casey. "Telemedicine and Diabetic Retinopathy: Review of Published Screening Programs." In: *Journal of endocrinology and diabetes* 2.4 (2015).

[Vac95] Corinne Vachier. "Extraction de caractéristiques, segmentation d'image et morphologie mathématique". PhD thesis. Dec. 1995.

[Var+92] J. Varis, J. Hartikainen, R. Lehtiniemi, J. Rantala, and M. Luukkala. "Transferable Measurement System for Fast Non-Destructive Evaluation". In: *Photoacoustic and Photothermal Phenomena III*. Ed. by Theodor Tamir, Helmut K. V. Lotsch, and Dane Bićanić. Vol. 69. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 565–567. ISBN: 978-3-662-13876-2 978-3-540-47269-8. DOI: 10.1007/978-3-540-47269-8_145.

[Var+95] Jussi Varis, Markku Oksanen, Jukka Rantala, and Mauri Luukkala. "Observations on image formation in the line scanning thermal imaging method". In: *Review of Progress in Quantitative Nondestructive Evaluation*. Springer, 1995, pp. 447–452.

[Ver+13] Rodrigo Veras, Fatima Medeiros, Romuere Silva, and Daniela Ushizima. "Assessing the accuracy of macula detection methods in retinal images". In: *Digital Signal Processing (DSP), 2013 18th International Conference on*. IEEE, 2013, pp. 1–6.

[Vin93] Luc Vincent. "Grayscale area openings and closings, their efficient implementation and applications". In: *First Workshop on Mathematical Morphology and its Applications to Signal Processing*. 1993, pp. 22–27.

[VM95] Corinne Vachier and Fernand Meyer. "Extinction value: a new measurement of persistence". In: *IEEE Workshop on Nonlinear Signal and Image Processing*. Neos Marmaras, Greece, June 1995, pp. 254–257.

[VS91] Luc Vincent and Pierre Soille. "Watersheds in digital spaces: an efficient algorithm based on immersion simulations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.6 (June 1991), pp. 583–598. ISSN: 1939-3539. DOI: 10.1109/34.87344.

[Wan+90] YQ Wang, PK Kuo, LD Favro, and RL Thomas. "A novel "flying-spot" infrared camera for imaging very fast thermal-wave phenomena". In: *Photoacoustic and Photothermal Phenomena II*. Springer, 1990, pp. 24–26.

[WSM11] Daniel Welfer, Jacob Scharcanski, and Diane Ruschel Marinho. "Fovea center detection based on the retina anatomy and mathematical morphology". In: *Computer Methods and Programs in Biomedicine* 104.3 (Dec. 2011), pp. 397–409. ISSN: 01692607. DOI: 10.1016/j.cmpb.2010.07.006.

[Xia+17] Zhitao Xiao, Xinpeng Zhang, Lei Geng, Fang Zhang, Jun Wu, Jun Tong, Philip O. Ogunbona, and Chunyan Shan. "Automatic non-proliferative diabetic retinopathy screening system based on color fundus image". In: *BioMedical Engineering OnLine* 16.1 (Dec. 2017). ISSN: 1475-925X. DOI: 10.1186/s12938-017-0414-z.

[Yu+12] Honggang Yu, Carla Agurto, Simon Barriga, Sheila C. Nemeth, Peter Soliz, and Gilberto Zamora. "Automated image quality evaluation of retinal fundus photographs in diabetic retinopathy screening". In: *Image analysis and interpretation (SSIAI), 2012 IEEE southwest symposium on*. IEEE, 2012, pp. 125–128.

[ZF14] Matthew D. Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer vision–ECCV 2014*. Springer, 2014, pp. 818–833.

[Zha+14] Xiwei Zhang, Guillaume Thibault, Etienne Decencière, Beatriz Marcotegui, Bruno Laÿ, Ronan Danno, Guy Cazuguel, Gwénolé Quellec, Mathieu Lamard, Pascale Massin, Agnès Chabouis, Zeynep Victor, and Ali Erginay. "Exudate detection in color retinal images for mass screening of diabetic retinopathy". In: *Medical Image Analysis* 18.7 (Oct. 2014), pp. 1026–1043. ISSN: 1361-8415. DOI: 10.1016/j.media.2014.05.004.

[Zha14] Xiwei Zhang. "Image processing methods for computer-aided screening of diabetic retinopathy". PhD thesis. École nationale supérieure des mines Paris, 2014.

RÉSUMÉ

Cette thèse traite de la détection et de la localisation d'objets, dans le contexte de deux projets : ATHENA et RetinOptic. Ces dernières années, les réseaux de neurones convolutionnels sont devenus l'approche prédominante pour ces tâches; cependant, ces techniques ont leurs inconvénients, et peuvent dans certains cas ne pas être appliquables. Lorsque peu d'exemples sont disponibles, ou que l'interprétabilité du modèle est essentielle, des méthodes de traitement d'image plus traditionnelles peuvent être plus adaptées. Dans ce manuscrit, plusieurs approches de détection sont décrites, selon la disponibilité des données et les contraintes spécifiques des problèmes.

Dans la première partie, consacrée au projet ATHENA, le but est de détecter et caractériser des défauts sur des images thermiques de pièces métalliques. La solution proposée, reposant sur l'analyse de certains extrema, nous permet de fournir une segmentation des défauts, et des définitions rigoureuses des concepts de rapport signal/bruit ou de symétrie du signal, jusqu'ici mal définis et dépendants de l'utilisateur.

Une analyse théorique plus détaillée des extrema est ensuite présentée, étendant la notion classique de valeurs d'extinctions. Nous introduisons et illustrons plusieurs nouvelles décompositions morphologiques, ainsi que de nouveaux opérateurs morphologiques, et de nouveaux attributs des extrema.

Dans la seconde partie, consacrée au projet RetinOptic, l'objectif est de détecter et localiser la macula sur des images de fond d'œil, à l'aide d'une solution suffisamment rapide et légère pour être intégrée à un système embarqué. Nous avons constitué une base annotée de plus de 6000 images, et utilisé différents types de réseaux de convolution, correspondant à différentes formulations de notre problème : classification, régression ou segmentation. Via un post-traitement de la sortie de notre réseau de segmentation, nous fournissons un score de qualité de l'image reposant sur la visibilité de la macula.

L'objectif d'un réseau de télémedecine est d'éviter les consultations médicales non nécessaires, dans un contexte où les ophtalmologistes sont peu nombreux par rapport au nombre croissant de personnes diabétiques; en cas d'acquisition ininterprétable, il s'agit d'un échec du réseau de télémédecine. Grâce à une vérification automatique de la qualité de l'image telle que nous le proposons, l'opérateur peut être prévenu si l'acquisition doit être refaite. En limitant le nombre d'examens ininterprétables, le nombre de consultations non nécessaires peut également être limité, améliorant l'efficacité du système de dépistage.

MOTS CLÉS

Segmentation d'images ; Thermographie ; Valeurs d'extinction ; Opérateurs morphologiques ; Réseaux de neurones convolutionnels ; Imagerie rétinienne

ABSTRACT

This thesis is centered around object detection and localization, in the context of two projects: ATHENA and RetinOptic. Over the past few years, convolutional neural networks have become the predominant approach for these tasks; however, these techniques have their drawbacks, and may not be applicable, in certain cases. When few examples are available, or when model interpretability is required, more traditional image processing methods may be better suited to the task. In the present work, object detection is addressed in different ways, depending both on the availability of data, and problem-specific requirements.

In the first part, concerning the ATHENA project, the aim is to detect and characterize defects on thermal images of metallic pieces. The solution we propose, based on the analysis of certain extrema of interest, enables us to provide a segmentation of defects, and proper definitions to the concepts of signal/noise ratio, or signal symmetry, which were so far ill-defined and operator dependent.

A more thorough theoretical analysis of extrema is then presented, expanding on the classic notion of extinction values. We introduce and illustrate several new morphological decompositions, as well as new morphological operators, and new features of extrema.

In the second part, concerning the RetinOptic project, the aim is to detect and localize the macula on eye fundus images, with a solution light and fast enough to be integrated in an embedded system. We constituted an annotated database of more than 6000 images, and used several kinds of convolutional networks, corresponding to different ways of formulating the problem, as a classification, regression, or segmentation task. By post-processing the output of our segmentation network, we provide an image quality score based on the visibility of the macula.

The purpose of a telemedicine network is to avoid unnecessary medical appointments, in a context where ophthalmologists are few, compared to the growing number of diabetic people; that purpose is defeated if the acquisition is uninterpretable. With an integrated image quality assessment like the one we propose, the operator can be told whether the acquisition should be re-done. By limiting the number of uninterpretable examinations, the number of unnecessary appointments can be limited as well, improving the efficiency of the screening network.

KEYWORDS

Image segmentation; Thermography; Extinction values; Morphological operators; Convolutional neural networks; Retinal imaging