Machine Learning for Simplifying the Use of Cardiac Image Databases

Ján Margeta

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n° 84 :

Sciences et technologies de l'information et de la communication

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité " Contrôle, optimisation et prospective "

présentée et soutenue publiquement par

Ján MARGETA

le 14 Décembre 2015

Apprentissage automatique pour simplifier l'utilisation de banques d'images cardiaques

Machine Learning for Simplifying the Use of Cardiac Image Databases

Directeurs de thèse : **Nicholas AYACHE** et **Antonio CRIMINISI**

Jury

M. Patrick CLARYSSE, DR, Creatis, CNRS, INSA Lyon	Rapporteur
M. Bjoern MENZE, Professeur, ImageBioComp Group, TU München	Rapporteur
M. Hervé DELINGETTE, DR, Asclepios Research Project, Inria Sophia Antipolis	Président
M. Antonio CRIMINISI, Chercheur principal, MLP Group, Microsoft Research Cambridge	Examinateur
M. Hervé LOMBAERT, Docteur, Asclepios Research Project, Inria Sophia Antipolis	Examinateur
M. Alistair A. YOUNG, Professeur, Auckland Bioengineeding Institute, University of Auckland	Examinateur
M. Nicholas AYACHE, DR, Asclepios Research Project, Inria Sophia Antipolis	Examinateur

T H È S E

MINES ParisTech
Centre de Mathématiques Appliquées
Rue Claude Daunesse B.P. 207, 06904 Sophia Antipolis Cedex, France

Acknowledgments

This has been an incredible journey. A journey of learning and self-discovery. Many people, papers, or books have helped to shape the thesis and my thinking. I have been lucky to have had amazing mentors and have worked with many incredibly smart people. A complete list would easily take up another chapter. It has been a great pleasure and honour to work and spend time with you. Thank you all!

First, I would like to thank my PhD supervisors and mentors. I fell many times but they never gave up on me, always helped me to stand back up and to pull myself forward. To Nicholas Ayache, not only for inviting me to join Asclepios, but also for sharing his vision with me and for his inspiring leadership of our vibrant team. Thank you for helping me to get better clarity in cloudy times. I am also immensely grateful to Antonio Criminisi. For his encouragement, for the discussions that helped to spark new ideas. For showing me how to practically do machine learning and the importance of visualisation. I am also indebted to you for the opportunity to visit MSR in Cambridge and spend a fantastic summer on a fascinating project.

I am grateful to Hervé Lombaert for his enthusiasm that never vaned, for thoroughly rereading the thesis and helping me to find the right voice. For all the unforgettable moments at and outside of Asclepios. I would like to express my gratitude to Peter Kontschieder, my MSR summer mentor, for sharing his endless passion and energy, for pushing me into challenging problems and for helping me to improve my focus and iteration speed. I would like to thank to all permanent researchers at Asclepios. To Maxime Sermesant, for all cardiac discussions, to Hervé Delingette for inspiring work and presiding my jury, to Xavier Pennec for inspiring scientific rigor, to Olivier Clatz for sharing his enterpreneurial spirit with me.

I would like to thank to all of my jury members and my thesis reviewers for joining us in Sophia Antipolis from faraway lands, for thoroughly reading the thesis and for sharing their feedback with me. I enjoyed the pre-defence chat with Patrick Clarysse, thank you for the encouraging words. I am grateful to Bjoern Menze, for the discussions we have had during his visits to Asclepios. They helped me better understand what content retrieval should be about. I cannot thank enough to Alistair Young. Not only for providing us with the fantastic data resource but also for making the world of cardiac imaging a better place. For the care put into the Cardiac atlas project and for the number of inspiring papers that influenced my work. My big thanks and kudos go to Catalina Tobon and Avan Suinesiaputra, for making science more objective and for organising two incredibly fun challenges that I had the chance to participate in.

I am grateful to all cardiologists I have met and had the chance to work with. In particular, my thanks go to Daniel C Lee, for input on our papers, I learned a lot from your feedback. To Philipp Beerbaum for talking us through real MRI acquisitions at St. Thomas Hospital, and for patiently responding to a number of naive cardiac questions. To Andrew Taylor and the whole GOSH team. I truly

enjoyed the exchanges we have had. They helped me to get better understanding of the clinical challenges.

This would be so much more difficult without the flawless support of Isabelle Strobant. I would also like to thank Scarlet Schwiderski-Grosche and Laurent Massoulié for making the joint PhD with MSR possible. I am grateful to Valérie Roy, for her incredible help in the final and most important stage of the thesis. To Vibot, the finest master in computer vision, for the foundations it gave me. To Prof. Hamed Sari-Sarraf for an impeccable machine learning course.

I have had the chance to share the office with some truly remarkable people. Thank you all for your friendships and for the good times we have had. To Stéphanie and Loïc, for their kindness and patience with many of my questions. To Thomas for his wit, cheerfulness and for always being there. To Clair for his curiosity. To Darko for your inspiring healthier research lifestyle. Thanks to all PhDs, Post-Docs and engineers at Asclepios. For providing the foundation for all that is so great about the team. My huge thanks go to Chloé, Krissy (also for being a fantastic scientific wife), Marine, Rocío, and Flo for the great times together in and out of the lab in the nature. To Ezequiel, for helping me out with my first academic paper. To Matthieu, Raphael, Sophie, Marc-Michel, for selflessly being there till the last rehearsal. To Nicolas and Vikash, for sharing laughs and the path till the end of the Mesozoic era. To Adityo, Alan, Barbara, Bishesh, Christof, Federico, Hakim, Héloïse, Hugo, Irina, John, Loïc, Marco, Marzieh, Mehdi, Michael and Roch. I had an incredible summer internship at MSR in Cambridge. I am grateful to Yani and Sam for the "deep talks" and crawls "under the table", to Qinxun for great time and culinary experiences (keep rolling Jerry!), Ina, Diana and Jonas. To Rubén and Oscar.

To my friends, to Kos and Amanda, Natalka and Zhanwu, Alex, Beto, Emka, Tjaša and Xtian for their undisputed friendship. To Math, Issis, Hernán and Arjan for the wonderful time during my visits of Paris, Barcelona and London. To Zoom, Nico, Clém, Francis for their hospitality. To my kayak friends, for their camaraderie, and for keeping me physically alive and mentally sane. To Oli for showing me another perspective on the world. To Raph, Laurent, Hervé, Thierry, Fred, Eric and the whole SPCOCCK. To Steffi and Pascal. I am indebted to Mumu and the teams at CHU Nice and Bratislava for saving my life.

Last, but not least, none of this would be possible without the love and support of my family, I am deeply grateful for always believing in me. To my dad, a tremendous source of inspiration to me, to my mum for constantly pushing me till the finish line, and to Karol, the best brother ever! To Magdalénka, my (not only) tireless copywriter, for her endless love and patience along this journey.

This work was supported by Microsoft Research through its PhD Scholarship Programme and ERC Advanced Grant MedYMA 2011-291080. The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 611823 (VP2HF).

Table of Contents

Abbreviations and Acronyms

AHA	American Heart Association
Ao	aorta
AoD	descending aorta
AV	aortic valve
BOW	Bag of words
CAP	Cardiac atlas project
CBIR	Content based image retrieval
CMR	Cardiac magnetic resonance imaging
CNN	Convolutional neural network
CT	Computed tomography
DE-MRI	Delayed enhancement MRI
DF	Decision forest
DICOM	Digital Imaging and Communications in Medicine
DTW	Dynamic time warping
ECG	Electrocardiogram
ED	end diastole
EDV	end diastolic volume
EF	ejection fraction
EHR	Electronic health record
ES	end systole
ESV	end systolic volume
GPU	Graphical processing unit
GRE	Gradient echo

HIPAA	Health Insurance Portability and Accountability Act
LA	left atrium
LoG	Laplacian of Gaussian
LRN	Local response normalisation
LV	left ventricle
LVC	left ventricular cavity
LVOT	Left ventricular outflow tract
ML	machine learning
MR	Magnetic resonance
MRI	Magnetic resonance imaging
MV	mitral valve
NAF	Neighbourhood approximating forest
NPV	Negative predictive value
PACS	Picture archiving and communication system
PCA	Principal Component Analysis
PET	Positron emission tomography
PM	papillary muscle
PPM	posterior papillary muscle
PPV	Positive predictive value
RA	right atrium
ReLU	Rectified linear unit
RGB	Red-Green-Blue
RV	right ventricle
SAX	short axis
SCMR	Society for Cardiac Magnetic Resonance
SGD	Stochastic gradient descent

SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
SSFP	Steady state free precession
STACOM	Statistical Atlases and Computational Modeling of the Heart
SVM	Support vector machine
TOF	Tetralogy of Fallot
TV	tricuspid valve
US	Ultrasound

Chapter 1

Introduction

Contents

1.1 The dawn of the cardiac data age

The developments in cardiology over the last century (Cooley and Frazier, 2000; Braunwald, 2014) have been quite spectacular. Many revolutions have happened since the first practical Electrocardiogram (ECG) by Einthoven in 1903. These include cardiac catheterization (1929), heart and lung machine and first animal models in the 1950s, minimally invasive surgeries (1958), and drug development (β blockers (1962), statins (1971), and angiotensins (1974)). The diagnostic imaging of the heart has also vastly improved. The post-war development of cardiac Ultrasound (US), Computed tomography (CT) (1970s) and Magnetic resonance imaging (MRI) (1980s) have helped us to non-invasively peek into the heart at a remarkable level of detail.

All of these advances have dramatically changed the course of cardiovascular disease management and by 1970 the mortality due to these diseases in high income countries has tipped and has been steadily declining (Fuster and Kelly, 2010, p52) ever since. Yet, the cardiovascular diseases remain the number one killer in the world (Nichols et al., 2012, p 10;Roger et al., 2011), causing 47% of all deaths in Europe.

We are at the dawn of the age where new cardiac image acquisition techniques, predictive *in silico* cardiac models (Lamata et al., 2014), realistic image simulations (Glatard et al., 2013; Prakosa et al., 2013; Alessandrini et al., 2015), real-time patient monitoring (Xia et al., 2013), and large-scale cardiac databases (Suinesiaputra et al., 2014a; Petersen et al., 2013; Bruder et al., 2013) become ubiquitous and have the chance to further improve cardiac health and our understanding.

Data within these databases are only as useful as the questions they can help to answer, the insights they can generate, and the decisions they enable to make. Large population clinical studies with treatment recommendations can be made, supporting evidence can be tailored for each patient individually. Treatment can be adjusted by looking at similar, previously treated patients, comparing their outcomes, and predicting what is likely to happen. New teaching tools can be developed using the data to create virtual patient case studies and surgery simulations on 3D-printed models (Bloice et al., 2013; Kim et al., 2008; Jacobs et al., 2008) and boost the education and practice of cardiologists.

1.2 Challenges of large cardiac data organisation

The opportunities for novel uses of large image databases are countless, however, the usage of these databases poses new challenges. Rich cardiac collections with relevant images (including many rare conditions) are scattered across thousands of Picture archiving and communication system (PACS) servers across many countries and hospitals. Data coming from these heterogeneous sources are not only massive, but often also quite unstructured and noisy.

1.2.1 Data aggregation from multi-centre studies

The biobanks and international consortia managing medical imaging databases, such as the UK biobank (Petersen et al., 2013), the Cardiac atlas project (CAP) (Fonseca et al., 2011) or the VISCERAL project (Langs et al., 2013), have solved many difficult problems in ethics of data sharing, medical image organisation and data distribution — in particular, when aggregating the data from multiple sources. The PACS together with the Digital Imaging and Communications in Medicine (DICOM) standards have been invaluable in these efforts. Studies coming from multiple centres often use specific nomenclature, follow different guidelines or utilise different acquisition protocols. In these cases, even these standards are not sufficient.

1.2.2 Data standardisation

The image collections on PACS servers can be queried by patient information (*e.g.* ID, name, birth date, sex, height, weight), image modality, study date and other DICOM tags, sometimes by study description, custom tags of the clinicians, associated measurements (*e.g.* arterial blood pressure and cardiac heart rate) and disease or procedure codes. See Fig. 1.1 for an example of such interface.

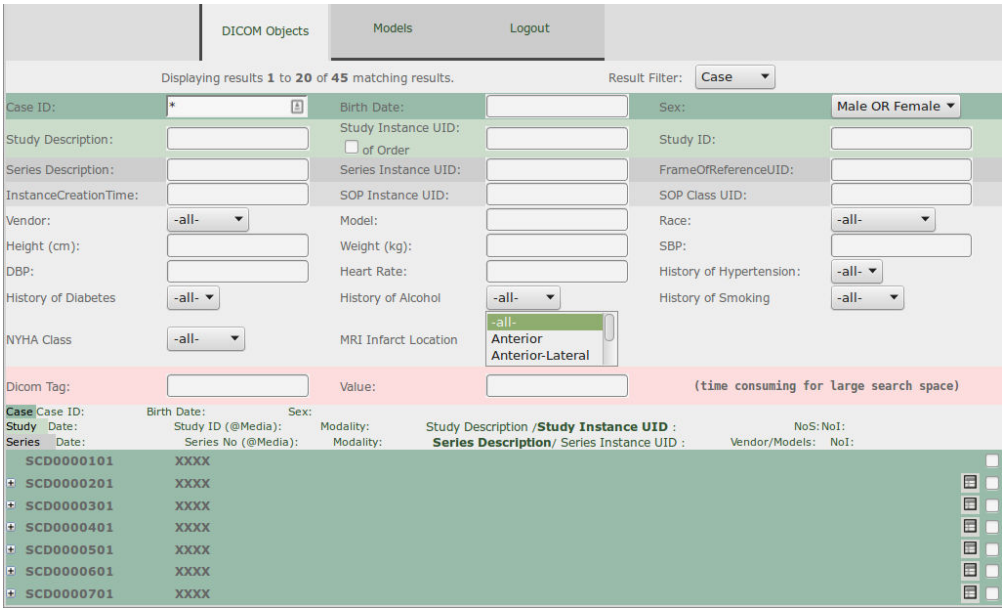


Figure 1.1: Cardiac atlas project web client interface.

There is no standard way to store some of the important image related information (*e.g.* the cardiac acquisition image plane information), where the naming depends on custom set-up of the viewing workstation and on the language practiced at the imaging centre. Even the standard DICOM tags often contain vendor specific nomenclature. For example, the same Cardiac magnetic resonance imaging (CMR) acquisition sequences are branded differently across the Magnetic resonance (MR) machine vendors (Siemens, 2010). While some implementation differences exist, these are not relevant for image interpretation, and the terminology could be significantly simplified (Friedrich et al., 2014). Parsing electronic publications with images is even a bigger challenge. These images are rarely in DICOM format and only the image content with textual description is available.

Such differences reduce our ability to effectively query and explore the databases for relevant images. The standardisation can be enforced by strictly following guidelines during image acquisition, and consistently using terminologies to encode the associated information such as the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) (Stearns et al., 2001). Care has to be taken to eliminate manual input errors. Images previously stored in the databases without the standardised information should be revisited for better accessibility.

1.2.3 Retrieving similar cases

Manually crawling through these growing databases to find similar previously treated patients (with supporting evidence) becomes very time consuming. Delivering archived images from PACS is, in practice, quite slow for such exploratory use. In addition, the cardiac imaging data stored in the databases are frequently 3D+t sequences, and important details can be easily missed during such visual inspection.

An alternative to this brute-force approach is to consistently describe the images with more compact representations. This prepares the cardiac image databases for future image retrieval. Though, it limits the search to the annotated data or to the cases known to the particular clinician. Most of the unannotated data therefore never gets used again and unused data means useless data.

1.2.4 Data annotation and consensus

Annotating these images simplifies their later reuse. However, together with the growth of the data, the demand for manual input becomes an increasing burden on the expert raters. One way to tackle this is to reduce the annotation task into very simpler questions that can be answered by a larger number of less experienced raters, for example via crowdsourcing.

As studied by Suinesiaputra et al. (2015), the variability of different radiologists (experts following the same guidelines) is not negligible. For example, in left ventricle (LV) segmentation, the papillary muscles (PMs) are myocardial tissue and therefore according to Schulz-Menger et al. (2013) should ideally be included in the myocardial mass and excluded from the left ventricular cavity (LVC) volume calculation. The corresponding reference values for volumes and masses (Maceira et al., 2006; Hudsmith†et al., 2005) should be used in this case. Some tools include the papillary muscles into the cavity volume instead. In this case a different set of reference values should be considered (Natori et al., 2006). The two reported measures can differ substantially. Ultimately, the PMs are part of the disease process (Harrigan et al., 2008) and deserve individual attention on their own.

The acquisition centres are equipped with different software tools and not all of these tools are equally capable. We still have a long way ahead to achieve reproducible extraction of image-based measures and consistent description of all relevant image information, especially given the constantly evolving guidelines.

1.2.5 The need for automated tools

For success in large scale analysis and use of the data, efficient ways of automatic clean-up and description of the cardiac data coming from several clinical centres with tools scalable to large data (Medrano-Gracia et al., 2015) are primordial. As we will see on the following example, manual design of such tools can rapidly become quite challenging.

1.3 A deceptively simple problem

Examine the following four CMR mid-ventricular short axis (SAX) slices obtained using the Steady state free precession (SSFP) acquisition sequence shown in Fig. 1.2. They belong to four individuals with different pathologies. One of them is an image of a healthy heart, another belongs to a patient after a myocardial infarction in the lateral wall, the third to a patient with a severely failing and non-compacting left ventricle and the last one shows a patient with idiopathic pericardial effusion. Can you correctly tell which one is which?

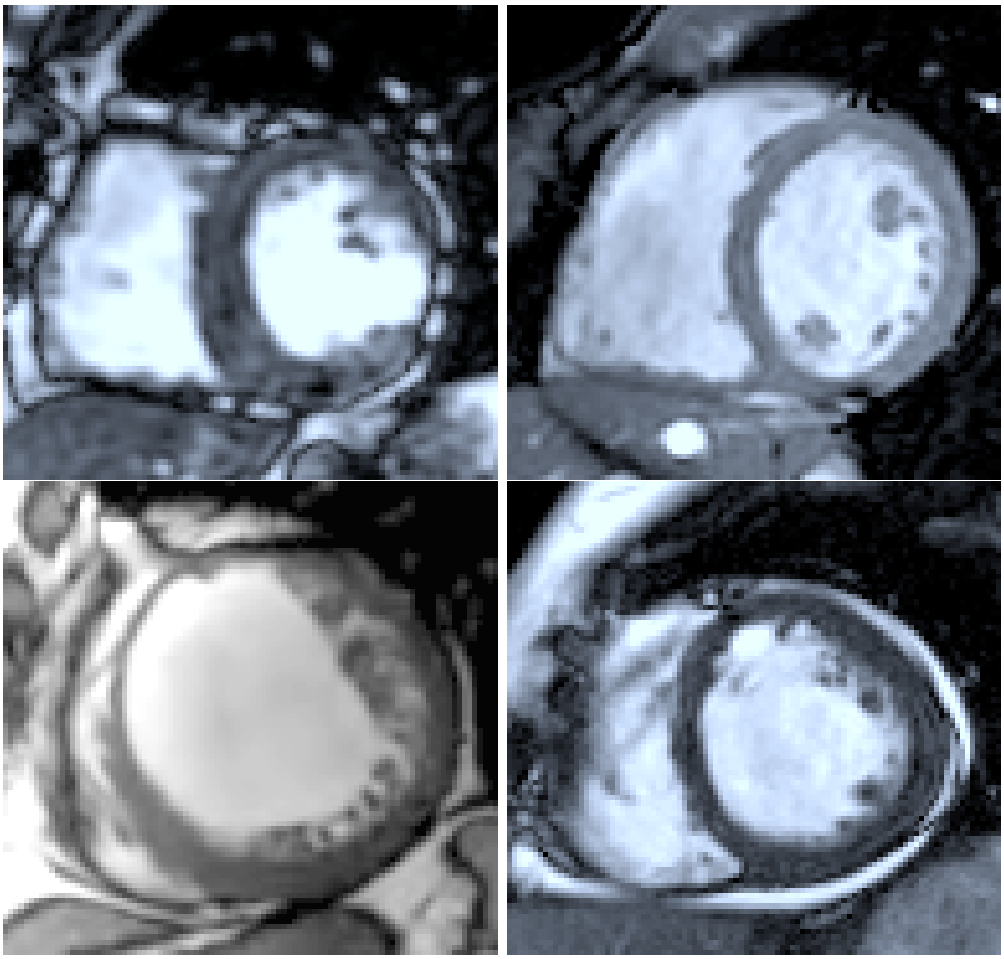


Figure 1.2: Four cardiac pathologies on MRI: heart with pericardial effusion, post lateral wall myocardial infarction heart, left ventricular non-compaction and a healthy heart. Can you identify them?[1]

This task of pathology identification is seemingly effortless for a person experienced in interpretation of cardiac images. Intuitively, we could recognise the post-myocardial infarction heart by a marked thinning of the lateral wall due to a transmural infarction and subsequent myocardial necrosis. One might also note sternal wire artefacts from a prior surgery. The failing non-compacting heart man-

ifests itself with massive dilation, prominent trabeculations in the left ventricular cavity, and significant reduction in myocardial contractility (best seen on cinematic image sequences). The pericardial effusion can be seen as a bright ring of liquid outside of the myocardium and swinging heart motion. And finally, the healthy heart looks "normal."

1.3.1 Automating the task

Only when we try to write a program to *mimic this reasoning on a computer* we can start to fully appreciate the true complexity of the visual tasks performed by the brain. The simplicity of relevant information extraction from the images is very deceptive. Intuitive concepts like the myocardial thinning, the cavity dilation, low contractility, bright ring, or swinging motion are concepts unknown to a machine. Not to mention the more global problem to automatically tell that all these images are short axis slices coming from a SSFP MR acquisition sequence.

One of the possibilities to extract this information by a computer is to start writing a set of rules. Myocardial thinning measurement can be measured as the length of the shortest line across the myocardium, counting pixels between two edges separating the white blob (the blood pool) and the grey (except for the effusion case) outer surroundings of the heart. Dilation is linked to the number of voxels within the ventricular blood pool and the cavity diameter. Both of these measures can be computed from segmentation of the left ventricular myocardium. The contractility can be estimated from displacement of the pixels, *e.g.*, via image registration. The subtle changes we might want to recognise are easily overshadowed by acquisition differences, *e.g.*, images coming from different MR acquisition machines, acquisition artefacts or differences in image orientation and heart position in the images. The images have no longer similar resolutions and image quality, tissue intensities on CMR between different machines do not match, acquisition artefacts are present or vendor specific variations of similar acquisition protocols are used. We soon discover that the set of rules to encode the relevant information and extract features to describe the cardiac images is endless.

1.3.2 The machine learning approach

The machine learning approach is quite different. Instead of manually hardcoding the rules, we specify a learning model and let the learning algorithm automatically figure out a set of rules by looking at the data, *i.e.*, to train the model. In the supervised learning setting, a set of examples together with desired outputs (*e.g.* images and their voxel-wise segmentations) is shown the training algorithm. The algorithm then picks rules that best map the inputs to the desired outputs. It is important that the learnt model generalises, *i.e.*, can reliably predict outputs for previously unseen images while ignoring irrelevant acquisition differences.

[1]Top left: Lateral infarction with thinning, top right: healthy, bottom left: left ventricular non-compaction, bottom right: pericardial effusion

Although good prediction is desirable, it is common to use "less than perfect" machine learning systems in a loop, and improve the models over time, when more data arrives. Also when guidelines change, these algorithms can be retrained and the images can be reparsed. Incorrect predictions can be fixed and added to the new training set and the model can then be retrained.

Machine learning in medical imaging has become remarkably important. This is partly due to the algorithmic improvements but mainly thanks to the increased availability of large quantities of data. While there are many machine learning algorithms, there is not (yet) a perfect one dealing with all the tasks at hand. One that is working for both large and small datasets.

Throughout this thesis we will use mainly three families of supervised machine learning (ML) algorithms: Linear regression models (the Support vector machines (SVMs) (Cortes and Vapnik, 1995) and ridge regression (Golub et al., 1979)), the Decision forests (DFs) (Ho, 1995; Amit and Geman, 1997; Breiman, 1999), and the Convolutional neural network (CNN) (Fukushima, 1980; LeCun et al., 1989).

1.4 Research questions of this thesis

This thesis aims to answer the following global question: **"How can we simplify the use of CMR image databases for cardiologists and researchers using machine learning?"** To help us answer this question, we addressed some of the main challenges introduced in Section 1.2.

1.4.1 How can we clean up and standardise DICOM tags for easier filtering and grouping of image series?

One of the first problems we face in cardiac imaging when dealing with large multi-vendor databases is the lack of standardisation in used notation in acquisition protocols (Friedrich et al., 2014) or naming of cardiac acquisition planes. Especially the knowledge of cardiac planes is essential for grouping the images into series and choosing the right image processing pipeline.

Chapter 2 presents our two methods for fixing noisy DICOM metadata with information estimated directly from image content. Our first method to recognise CMR acquisition planes uses classification forests applied on image miniatures. We show how cheaply generated new images can help to improve the recognition.

We then show how we modify a state of the art technique in a large scale visual object recognition, based on CNNs, to a much smaller cardiac imaging dataset. Our second method recognises short axis and 2-, 3- and 4-chamber long axis views with very promising recognition performance.

In **Appendix B** we show how the CNN-based features can be reused to regress cardiac point distribution models for inter-patient image alignment.

1.4.2 Can we teach the computer to understand cardiac anatomy and to segment cardiac structures from MR images?

Once we can describe cardiac images based on their views and merge them into spatio-temporal 3D+t series we can move on to teach the computer the basics of cardiac anatomy, *i.e.*, how to segment the cardiac images. Successful segmentation is essential to index cardiac images based on standard volumetric measures such as systolic and diastolic volume, ejection fraction, and myocardial mass.

In **Chapter 3** we extend the previous work on semantic segmentation using classification forests (Shotton et al., 2008; Geremia et al., 2011). We show how our modified algorithm learns to segment left ventricles from 3D+t MR short axis SSFP sequences without imposing any shape prior. Our decision forest classifier is trained in a layered fashion, and we propose new spatio-temporal features to classify the 3D+t sequences. We show that avoiding to hard-code the segmentation problem helps us to easily adapt this technique to segment other cardiac structures, the left atria — the black box of the heart, both from CMR and CT. We contributed these algorithms to two comparison studies for fair evaluation.

In **Appendix A** we propose a segmentation method exploiting unlabelled data in a semi-supervised setting to learn how to segment from sparse annotations.

1.4.3 How can we collect data needed by the computer for training of the machine learning algorithms and learn how to describe the hearts with semantically meaningful attributes?

Most of the practical machine learning problems are currently still solved in a fully supervised manner. It is therefore essential to acquire the ground-truth. **Chapter 4** deals with label collection for machine learning algorithms. We design a web-based tool for crowd-sourcing of cardiac attributes and use it to collect pairwise image annotations. We describe the cardiac shapes with their spectral signatures and use a linear predictor based on SVM classifier to learn ordering of the images based on their attribute values. Our preliminary results suggest that in addition to volumetric measurements obtainable from cardiac segmentations, the hearts could be described by cardiac attributes.

1.4.4 Can we automatically retrieve similar hearts?

The image similarity depends on the clinical question to be answered. Queries we might want to ask the retrieval system can be quite variable. **Chapter 5** builds on the Neighbourhood approximating forest (NAF) of Konukoglu et al. (2013) and presents our pipeline to learn shape, appearance and motion similarities between cardiac images and how we use them to structure the spatio-temporal cardiac datasets. We show how hearts with similar properties (similar ejection fraction) can be extracted from the database. In (Bleton et al., 2015), we then used a similar technique to localise cardiac infarcts from dynamic shapes only (no contrast agent needed).

1.5 Manuscript organisation

The presented thesis is organised around our published work and our work in preparation for submission. The manuscript also roughly progresses from global towards fine-grained description of the cardiac images. Each chapter in this thesis attempts to answer one of the objectives and to bring content-based retrieval of images from large-scale CMR databases closer to reality.

First, we train a system to fix image tags that are not captured by DICOM directly from image content. In Chapter 2, we show how to automatically recognise cardiac planes of acquisition. In Chapter 3, we propose a flexible automatic segmentation technique that learns to segment cardiac structures from spatio-temporal image data, using simple voxel-wise ground-truth as input, that could be used for automatic measurements. In Chapter 4, we suggest a way of collecting annotations necessary for training of automatic algorithms, and to describe the cardiac images with sets of semantic attributes. Finally, in Chapter 5, we propose an algorithm to structure the datasets and find similar cases with respect to different clinical criteria. Chapter 6 concludes the thesis with perspectives and future work. In the appendices, we illustrate how unlabelled data can be used for guided image segmentation (Appendix A), how to estimate cardiac landmarks for image alignment (Appendix B), or how to enhance pericardial effusion for image retrieval (Appendix C).

1.6 List of publications

Journal articles

- **J. Margeta**, A. Criminisi, R. Cabrera Lozoya, D. C. Lee, and N. Ayache, *"Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition"*, Computer methods in biomechanics and biomedical engineering: Imaging & visualisation, 2015.

- C. Tobon-Gomez, A. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, **J. Margeta**, Z. Sandoval, B. Stender, Y. Zheng, M. A. Zuluaga, J. Betancur, N. Ayache, M. A. Chikh, J.-L. Dillenseger, M. Kelm, S. Mahmoudi, S. Ourselin, A. Schlaefer, T. Schaeffter, R. Razavi, and K. Rhode, *"Benchmark for Algorithms Segmenting the Left Atrium From 3D CT and MRI Datasets"*, IEEE Transactions on Medical Imaging, vol. 34, no. 7, pages 1460 1473, 2015.

- A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M. P. Jolly, A. H. Kadish, D. C. Lee, **J. Margeta**, S. K. Warfield, and A. A. Young, *"A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images"*, Medical Image Analysis, vol. 18, no. 1, pages 50 62, 2014.

Peer reviewed conference and workshop papers

- **J. Margeta**, A. Criminisi, D. C. Lee, and N. Ayache, *"Recognizing cardiac magnetic resonance acquisition planes"*, in Conference on Medical Image Understanding and Analysis (MIUA 2014), 2014. Oral podium presentation.

- **J. Margeta**, K. S. McLeod, A. Criminisi, and N. Ayache, *"Decision forests for segmentation of the left atrium from 3D MRI"*, in International Workshop on Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, Held in conjunction with MICCAI 2013, Beijing, Lecture Notes in Computer Science, vol. 8830, pages 49 56, O. Camara, T. Mansi, M. Pop, K. Rhode, M. Sermesant, and A. Young, Eds., Springer Berlin / Heidelberg, 2014. Oral podium presentation.

- **J. Margeta**, E. Geremia, A. Criminisi, and N. Ayache, *"Layered Spatio-temporal Forests for Left Ventricle Segmentation from 4D Cardiac MRI Data"*, in International Workshop on Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, Held in conjunction with MICCAI 2011, Toronto, Lecture Notes in Computer Science, vol. 7085, pages 109 119, O. Camara, E. Konukoglu, M. Pop, K. Rhode, M. Sermesant, and A. Young, Eds., Springer Berlin / Heidelberg, 2012, Oral podium presentation.

- H. Bleton, **J. Margeta**, H. Lombaert, H. Delingette, and N. Ayache, *"Myocardial Infarct Localisation using Neighbourhood Approximation Forests"*, in International Workshop on Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, Held in conjunction with MICCAI 2015, Munich, O. Camara, T. Mansi, M. Pop, K. Rhode, M. Sermesant, and A. Young, Eds., 2015. Oral podium presentation.

- R. C. Lozoya, **J. Margeta**, L. Le Folgoc, Y. Komatsu, B. Berte, J. Relan, H. Cochet, M. Haïssaguerre, P. Jaïs, N. Ayache, and M. Sermesant, *"Confidence-based Training for Clinical Data Uncertainty in Image-based Prediction of Cardiac Ablation Targets"*, in International Workshop on Medical Computer Vision: Algorithms for Big Data, Held in conjunction with MICCAI 2014, Boston, Lecture Notes in Computer Science, vol. 8848, pages 148 159, B. Menze, G. Langs, A. Montillo, M. Kelm, H. Müller, S. Zhang, W. Cai, and D. Metaxas, Eds., Springer Berlin / Heidelberg, 2014. Oral podium presentation.

- R. C. Lozoya, **J. Margeta**, L. Le Folgoc, Y. Komatsu, B. Berte, J. S. Relan, H. Cochet, M. Haïssaguerre, P. Jaïs, N. Ayache, and M. Sermesant, *"Local late gadolinium enhancement features to identify the electrophysiological substrate of post-infarction ventricular tachycardia: a machine learning approach"*, in Journal of Cardiovascular Magnetic Resonance, vol. 17, no. Suppl 1, poster 234, 2015. Poster presentation.

In preparation

- **J. Margeta**, H. Lombaert, D. C. Lee, A. Criminisi, and N. Ayache, *"Learning to retrieve semantically similar hearts."*
- **J. Margeta**, E. Konukoglu, D. C. Lee, A. Criminisi, and N. Ayache, *"Crowdsourcing cardiac attributes"*,

Chapter 2

Learning how to recognise cardiac acquisition planes

Contents

Based on our published work (Margeta et al., 2014) on the use of decision forests for cardiac view recognition and the convolutional neural network approach (Margeta et al., 2015c) to further improve the performance.

Chapter overview

When dealing with large multi-centre and multi-vendor databases, inconsistent notations are a limiting factor for automated analysis. Cardiac MR acquisition planes are a particularly good example of a notation standardisation failure. Without knowing which cardiac plane we deal with, further use of the data without manual intervention is limited. In this chapter, we propose two supervised machine learning techniques to automatically retrieve missing or noisy cardiac acquisition plane information from Magnetic resonance imaging (MRI) and to predict the five most common cardiac views (or acquisition planes). We show that cardiac acquisitions are roughly aligned with the heart in the image center and use this to learn cardiac acquisition plane predictors from 2D images.

In our first method we train a classification forest on image miniatures. Dataset augmentation with a set of label preserving transformations is a cheap way that helps us to improve classification accuracy without neither acquiring nor annotating extra data. We further improve the forest-based cardiac view recogniser's performance by fine-tuning a deep Convolutional neural network (CNN) originally trained on a large image recognition dataset (ImageNet LSVRC 2012) and transfer the learnt feature representations to cardiac view recognition.

We compare these approaches with predictions using off the shelf CNN image features, and with CNNs learnt from scratch. We show that fine-tuning is a viable approach to adapt parameters of large convolutional networks for smaller problems. We validate this algorithm on two different cardiac studies with 200 patients and 15 healthy volunteers respectively. The latter comes from an open access cardiac dataset which simplifies direct comparison of similar techniques in the future. We show that there is value in fine-tuning a model trained for natural images to transfer it to medical images. The presented approaches are quite generic and can be applied to any image recognition task. Our best approach achieves an average F1 score of 97.66% and significantly improves the state of the art in image-based cardiac view recognition. It avoids any extra annotations and automatically learns the appropriate feature representation.

This is an important building block to organise and filter large collections of cardiac data prior to further analysis. It allows us to merge studies from multiple centers, to enable smarter image filtering, to select the most appropriate image processing algorithm, to enhance visualisation of cardiac datasets in content-based image retrieval, and to perform quality control.

2.1 Brief introduction to cardiac data munging

The rise of large cardiac imaging studies has opened us the door to better understanding and management of cardiac diseases. When handling data from various sources, *inconsistent, missing, or non-standard information* is unavoidable. The Digital Imaging and Communications in Medicine (DICOM) standard has solved many common problems in handling, archival, and exchange of information in med-

ical imaging by adding metadata to images and defining communication protocols. Nevertheless, a lot of the metadata crucial for filtering cases for studies is not standardised and still remains site and vendor specific.

Prior to any analysis, the data must be cleaned up and put into the same format. This process is often called *data munging* or *data wrangling*. Cardiac MRI acquisition plane information is a particularly important piece of information to be wrangled.

2.2 Cardiac acquisition planes

Instead of commonly used body planes (coronal, axial and sagittal) the CMR images are acquired along several oblique directions aligned with the structures of the heart. Imaging in these standard cardiac planes ensures efficient coverage of relevant cardiac territories (while minimising the acquisition time) and enables comparisons across modalities, thus enhancing patient care and cardiovascular research. The optimal cardiac planes depend on global positioning of the heart in the thorax. This is more vertical in young individuals and more diaphragmatic in elderly and obese.

An excellent introduction to the standard CMR acquisition planes can be found in Taylor and Bogaert (2012). These planes are often categorized into two groups — the short and the long axis planes (see Figures 2.1 and 2.2 for a visual overview). In this chapter, we learn to predict the five most commonly used cardiac planes acquired with Steady state free precession (SSFP) acquisition sequences to evaluate the left heart. These are the *short axis*, *2-*, *3-* and *4- chamber* and *left ventricular outflow tract* views. These five labels are the targets for our learning algorithm.

2.2.1 The need for automatic plane recognition

Why is it important to have an automatic way of recognising this information? Automatic recognition of this metadata is essential to appropriately select image processing algorithms, to group related slices into volumetric image stacks, to enable filtering of cases for a clinical study based on presence of particular views, to help with interpretation and visualisation by showing the most relevant acquisition planes, and in content-based image retrieval for automatic description generation. Although this orientation information is sometimes encoded within two DICOM image tags: *Series Description (0008,103E)* and *Protocol Name (0018,1030)*, it is not standardised, operator errors are frequently present, or this information is completely missing. In general, the DICOM tags are often too noisy for accurate image categorization (Gueld et al., 2002). Searching through large databases to manually cherrypick relevant views from the image collections is therefore very tedious. The main challenge for an image-content-based automated cardiac plane recognition method is the variability of the thoracic cavity appearance. Different parts of organs can be visible even in the same acquisition planes between different patients.

2.2.2 Short axis acquisition planes

Short axis slices (Figure 2.1) are oriented perpendicular to LV long axis.These are acquired regularly spaced from the cardiac base to the apex of the heart, often as a cine 3D+t stack. These views are excellent for reproducible volumetric measurements or radial cardiac motion analysis, but their use is limited in atrio-ventricular interplay or valvular disease study. The American Heart Association (AHA) nomenclature (Cerqueira et al., 2002) divides the heart into approximately three thirds: *basal*, *mid cavity* and *apical* slices (See Figure B.1 for more details).

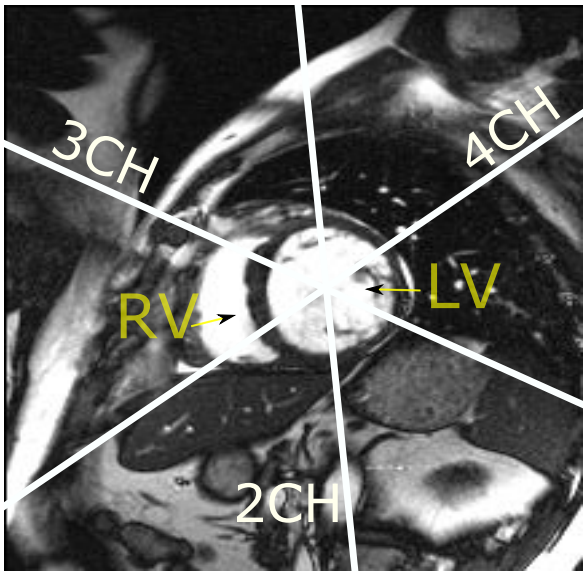
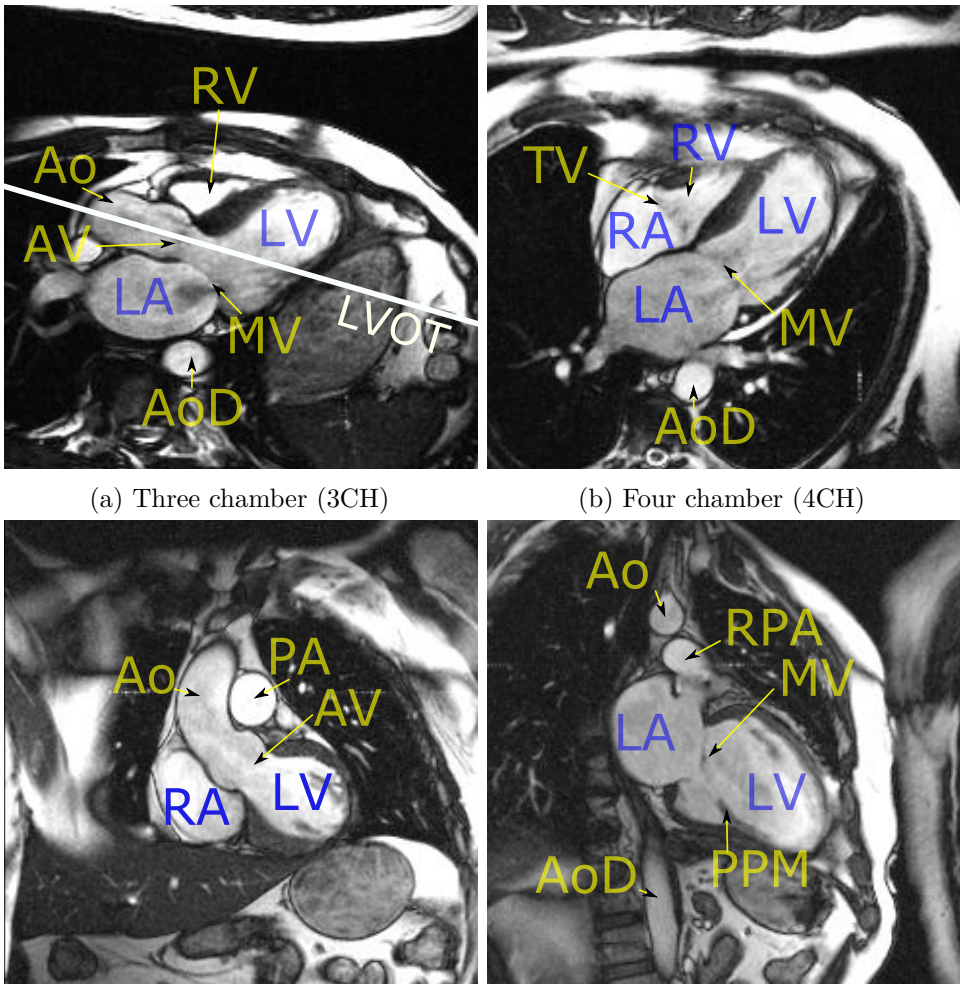


Figure 2.1: Example of a basal short axis view and mutual orientation of the long axis planes. Both left (LV) and right (RV) ventricle can be seen. The long axis planes are radially distributed around the myocardium to ensure the optimal coverage of the heart.

2.2.3 Left ventricular long axis acquisition planes

The long axis slices are usually acquired as 2D static images or cine 2D+t stacks. The 2-chamber, 3-chamber, and 4-chamber views (Figures 2.2a, 2.2b and 2.2d) are used to visualise different regions of the left ventricle (LV), mitral valve (MV) apparatus, aortic root and left atrium (LA). The 3-chamber (Fig. 2.2a) and the Left ventricular outflow tract (LVOT) (Fig. 2.2c) views provide visualisation of the aortic root from two orthogonal planes and help to detect any obstructions of the outflow tract and/or aortic valve (AV) regurgitation. The 4-chamber view (Fig. 2.2b) visualises both atrio-ventricular valves, all four cardiac chambers and their interplay.



(a) Three chamber (3CH)

(b) Four chamber (4CH)

(c) Left ventricular outflow tract (LVOT)

(d) Two chamber (2CH)

Figure 2.2: Examples of the main left ventricular long axis cardiac MR views. The main cardiac cardiovascular territories and structures can be visible such as: left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), mitral valve (MV), tricuspid valve (TV), aortic valve (AV), aorta (Ao), descending aorta (AoD) or posterior papillary muscle (PPM). Note the dark regurgitant jet into the left atrium (LA) adjacent to the mitral valve (MV) in Fig. 2.2a

2.3 Methods

2.3.1 Previous work

The previous work on cardiac view recognition has been concentrated mainly on real-time recognition of cardiac planes for echography (Otey et al., 2006; Park et al., 2007; Beymer et al., 2008). In addition to our work, there exists some work on MR (Zhou et al., 2012; Shaker et al., 2014). The common methods are based on dynamic active shape models (Beymer et al., 2008), require to train part detectors (Park et al., 2007) or landmark detectors (Zhou et al., 2012). Therefore, any new view will require these extra annotations to be made. Otey et al. (2006) avoid this limitation by training an ultrasound cardiac view classifier using gradient based image features. The most recently proposed work on cardiac view recognition from MR (Shaker et al., 2014) uses autoencoders. These learn image representations in an unsupervised fashion (the goal is to reconstruct images from a lower dimensional representation) and use this representation to distinguish between two cardiac views.

The state of the art in image recognition has been heavily influenced by the seminal works of Krizhevsky et al. (2012) and Cireşan et al. (2012) using Convolutional neural network (CNN). Krizhevsky et al. (2012) trained a large (60 million parameters) CNN on a massive dataset consisting of 1.2 million images and 1000 classes (Russakovsky et al., 2014). They employed two major improvements: *Rectified linear unit nonlinearity* to improve convergence, and *Dropout* (Hinton et al., 2012) to reduce overfitting.

Training a large network from scratch without a large number of samples still remains a challenging problem. A trained CNN can be adapted to a new domain by reusing already trained hidden layers of the network, though. It has been shown, *e.g.*, by Sharif et al. (2014) that the classification layer of the neural net can be stripped, and the hidden layers can serve as excellent image descriptors for a variety of computer vision tasks (such as for photography style recognition by Karayev et al. (2014)). Alternatively, the prediction layer model can be replaced by a new one and the network parameters can be fine-tuned through backpropagation.

2.3.2 Overview of our methods

A ground truth target label (2CH, 3CH, 4CH, LVOT or SAX) was assigned to each image in our training set by an expert in cardiac imaging. We use these labels in the training phase as a target to train. In the testing phase, we predict the cardiac views from the images and use the ground-truth only to evaluate our view recognition methods.

In this chapter, we compare the three groups of methods for automatic cardiac acquisition plane recognition. The first one is based on DICOM-derived orientation information. The algorithms in the other two families completely ignore the DICOM tags and learn to recognise cardiac views directly from image intensities.

In Section 2.4, we first present the recognition method using DICOM-derived features (the image plane orientation vectors, similar to Zhou et al. (2012)). Here,

we train a decision forest classifier using these 3-dimensional feature vectors.

The latter two approaches learn to recognise cardiac views from image content without using any DICOM meta-information. In Section 2.5, we present our classification forest-based method (Margeta et al., 2014) using pixels from image miniatures as features. We then introduce the third path for cardiac view recognition, using CNNs, as described in Section 2.6. In this section, we consider all commonly used approaches (training a network from scratch, reusing a hidden layer features from a network trained on another problem, and fine-tuning of a pretrained network) for using a CNN in cardiac view recognition.

To increase the number of the training samples (for image content-based algorithms) we augment the dataset with small label preserving transformations such as image translations, rotations, and scale changes. See Section 2.5.4 for more details.

In Section 2.7, we compare all of these approaches. We show how the CNN-based approaches outperform the previously introduced forest-based method and achieve very good perfomance. Finally, in Section 2.8, we present and discuss our results.

2.4 Using DICOM orientationtag

Plane normal + Forest

Zhou et al. (2012) showed that where the *DICOM orientation (0020,0037)* tag is present we can use it to predict the cardiac image acquisition plane (see Figure 2.3). This tag is not defined as a cardiac view but as two 3-dimensional vectors defining orientation of the imaging plane with respect to the MR scanner coordinate frame.

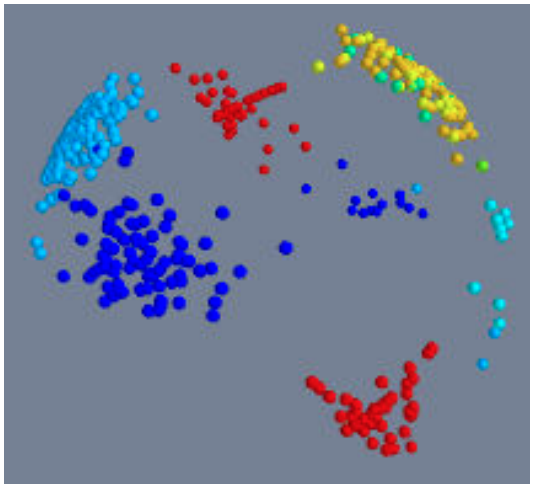


Figure 2.3: Tips of DICOM plane normals for different cardiac views. In our dataset, distinct clusters can be observed (best viewed in colour). Nevertheless, the separation might not be the case for a more diverse collection of images. Moreover, as we cannot always rely on the presence of this tag an image-content-based recogniser is necessary.

It is straightforward to compute the 3-dimensional normal vectors of this plane as a cross-product of these two vectors specified in the tag. We then feed these three-dimensional feature vectors into any classifier, in our case a classification forest (see Section 2.5.1 for more details on classification forests). This method is shown in the results section as **Plane normal + forest**.

2.4.1 From DICOM metadata towards image content

This method uses feature vectors computed from the DICOM orientation tag and cannot be used in the absence of this tag. This happens for example in DICOM tag removal after an incorrectly configured anonymisation procedure, when parsing images from clinical journals or when using image formats other than DICOM. In these cases we have to rely on recognition methods using exclusively the image content.

In the next two sections, we present two such methods. One that is based on classification forests and image miniatures (Margeta et al., 2014) and the other one is using CNNs. We learn to predict the cardiac views from $2D$ image slices individually, rather than using $2D+t$, $3D$ or $3D+t$ volumes. This decision makes our methods more flexible and applicable also to view recognition scenarios when only $2D$ images are present, *e.g.*, when parsing clinical journals or images from electronic publications.

2.5 View recognition from image miniatures

Miniatures + forest

First, we propose an automatic cardiac view recognition pipeline (see Fig. 2.4) that learns to recognise the acquisition planes directly from CMR images by combining image miniatures with classification forests.

2.5.1 Decision forest classifier

Decision forest classifier or classification forest (Ho, 1995; Amit and Geman, 1997; Breiman, 1999) is an ensemble machine learning method that constructs a set of binary decision trees with split decisions optimised for classification. This method is computationally efficient and allows automatic selection of relevant features for the prediction.

The decision forest framework itself is also quite flexible (Criminisi et al., 2011b; Pauly, 2012) and has already been used to solve a number of problems in medical imaging. For example, for image segmentation (Geremia et al., 2011), organ detection (Criminisi and Shotton, 2011; Pauly et al., 2011), manifold learning (Gray et al., 2011), or shape representation (Swee and Grbić, 2014).

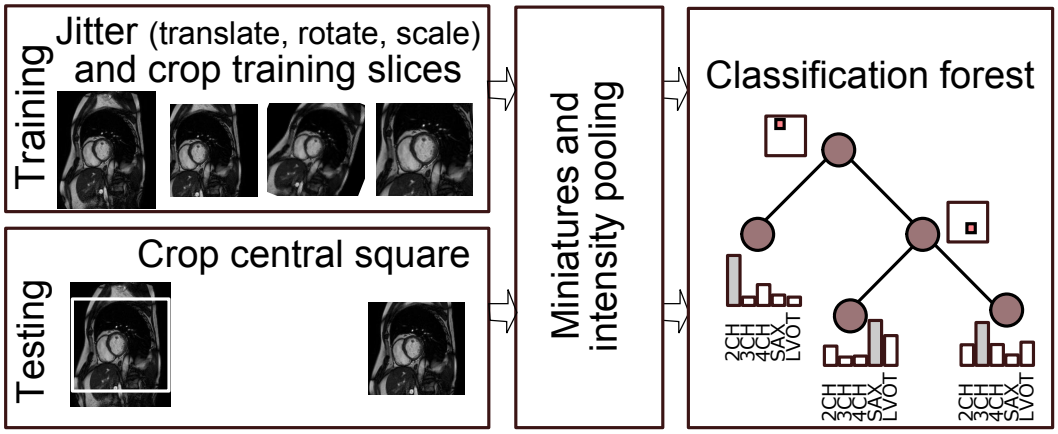


Figure 2.4: Discriminative pixels from image miniatures are chosen from a random pool as features for a classification forest. We jitter the training dataset to improve robustness to differences in the acquisitions without the need for extra data.

Training phase

During the training phase, the tree structure is optimised with a divide and conquer strategy on the collection of data points X by recursively splitting them into the left and right branches. This splitting is done such that points with different labels get separated while the same label points are grouped together, *i.e.*, the label purity in the branches increases. See Figure 2.5 for an illustration of this process.

At each node of the tree a feature from a randomly drawn subset of all features and a threshold value are chosen such that class impurity I in both branches is minimised. We weight samples from the under-represented views more (inversely proportionally to dataset view distribution) and normalise them to sum to one at each node.

$$I(X, \theta) = w_{left} H(X_{left}) + w_{right} H(X_{right}) \tag{2.1}$$

H is weighted entropy and X_{left} and X_{right} are point subsets falling to either the left or the right branch, based on the tested feature value and threshold. w_{left}, w_{right} are sums of sample weights at each branch and w_k is the sum of weights for a particular class k in the branch.

$$H(X) = -\sum_{k=1}^{K} w_k \log(w_k) \tag{2.2}$$

Only a random subset of features (*e.g.* components of the 3D plane orientation vector or a set of pixel intensity values at different fixed locations of the images) is tested at each node of each tree and a simple threshold on this value is used to divide the data points into the left or the right partition. This helps to make the trees in the forest different from each other which leads to better generalisation.

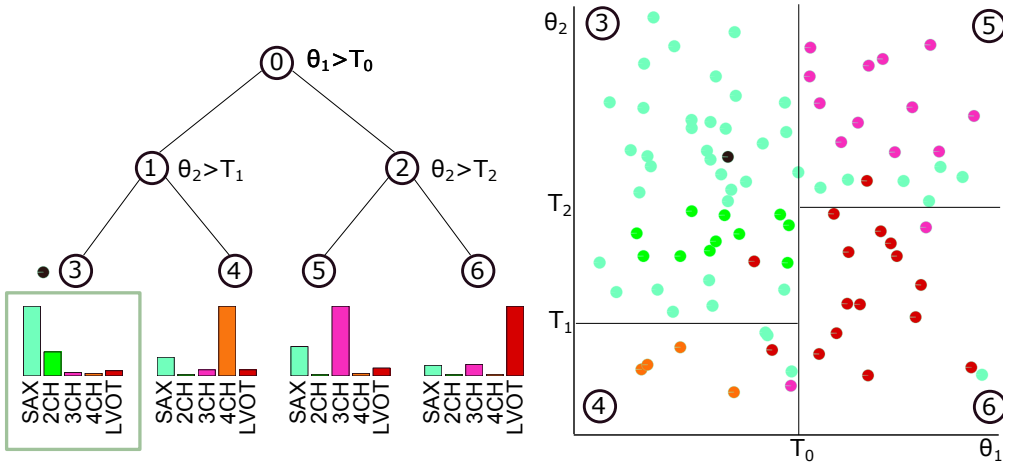


Figure 2.5: We illustrate a 2D feature space and a single tree from the classification forest. At the training phase, the feature space (for example constructed by sampling image miniature intensities at random locations) is recursively partitioned (horizontal and vertical lines cut through the feature space) to recognise cardiac planes of acquisition. Class distributions at the leaves are stored for the test time. At the test time, the tested images are passed through the decisions of each tree until they reach the final set of leaves (one per tree). Class with maximal average probability across the forest is chosen as the prediction.

When classifying a new image, features chosen at the training stage are extracted and the image is passed through the decisions of the forest (fixed in the training phase) to reach a set of leaves. Class distributions of the reached leaves are averaged across the forest and the most probable label is selected as the image view. For excellent in-depth discussions on decision forests, in particular applied to medical imaging, see (Criminisi et al., 2011b; Pauly, 2012).

2.5.2 Alignment of radiological images

The radiological images are mostly acquired with the object of interest in the image center and some rough alignment of structures can be expected (see Figure 2.6).

Note the large bright cavity in the center (3CH, 4CH), dark lung pixels just above the cavity (SAX), or black air pixels on the left and right side (2CH, SAX). Image intensity samples at fixed positions (even without registration) provide strong cues about the position of different tissues.

2.5.3 Pooled image miniatures as features

It has been shown by Torralba et al. (2008) that significantly down-sampled image miniatures can be used for image recognition. In our case, we extract the central square from each image, resample it to 128×128 pixels (with linear interpolation), and linearly rescale to an intensity range between 0 and 255. We subsample the

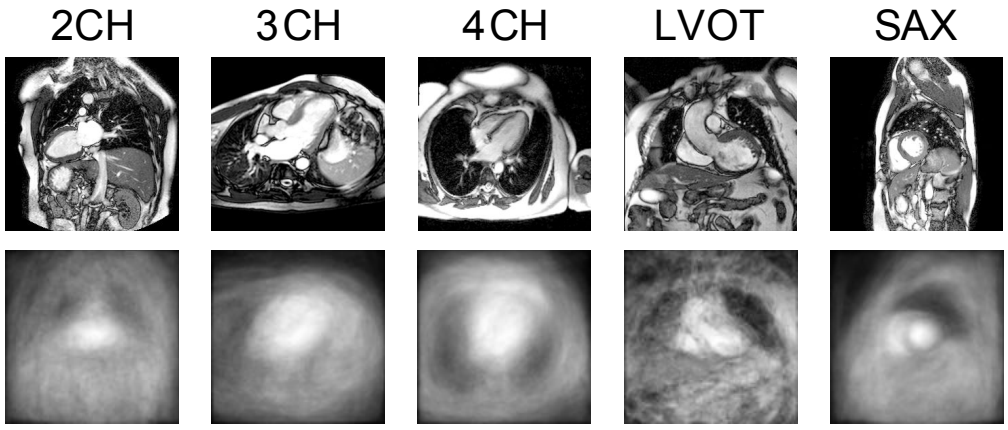


Figure 2.6: Example of each cardiac view used in this work (above) and corresponding central square region mean intensities across the whole dataset (below).

cropped centers to two fixed sizes (20×20 and 40×40 pixels). In addition, we divide the image into non-overlapping 4×4 tiles and for each of these tiles compute the intensity minima and maxima (see Figure 2.7).

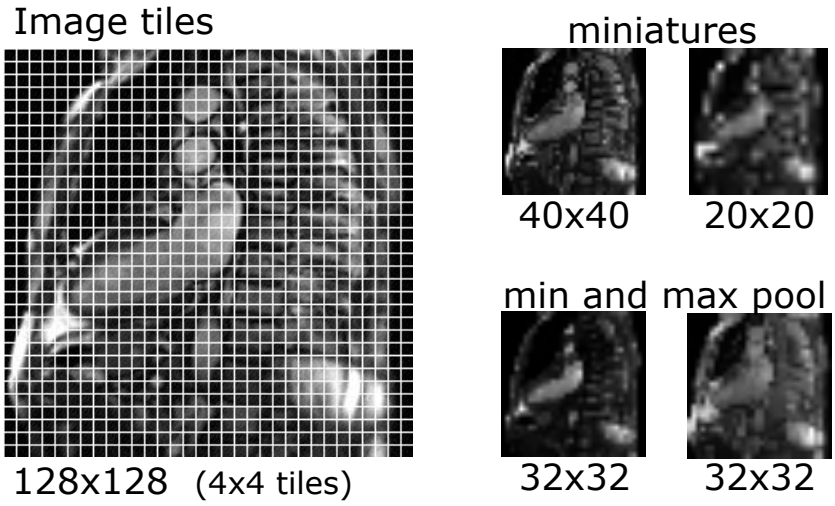


Figure 2.7: Image miniatures features: downsampled images and tile intensity minima and maxima are used.

This creates a set of pooled image miniatures (32×32 pixels each). The pooling adds some invariance to small image translations and rotations (whose effect is within the tile size). The pixel values at random positions of these miniature channels are then used directly as features.

In total 64 random locations across all four miniatures channels are tested at each node of each tree when training the forest. The location and threshold value combination that best (Eq. (2.1)) partition the data-points are then selected and stored and the data points are correspondingly divided into the left and right

branches. We recursively continue dividing the dataset until not less than 2 points are left in each leaf or no gain is obtained by further splitting. We trained 160 trees in total using Scikit-learn (Pedregosa et al., 2011).

This method is shown in the evaluation as **Miniatures + forest**.

2.5.4 Augmenting the dataset with geometric jittering

While the object of interest is in general placed at the image center, some differences between various imaging centres and positioning of the heart on the image remain (see Fig. 2.12). The proposed miniature features are not fully invariant to these changes *per se*. To account for larger differences in acquisition we augment the training set (artificially increase its size) with extra images created by transforming the originals. In other words, we artificially generate new images from the originals by geometric transformations. It makes sense to perform appearance transformations as well (*e.g.* intensity alterations done by Wu et al. (2015) or adding synthetic bias fields). Only care must be taken not to modify the target label. The advantage of data augmentation is that very realistic images can be obtained without extra acquisition or labelling cost. The downside is that excessive augmentation makes the images look more alike and there is a greater risk of overfitting to the training set.

For our purpose, we augment the dataset on a regular grid of transformation parameters. These were translations (all shifts in x and y between -10 and 10 pixels for a 5×5 grid), but also scale changes ($1 - 1.4$ zoom factor with 8 steps while keeping the same image size) and in-plane rotations around the image centre (angles between -10 and 10 degrees with 20 steps). The augmented images were resampled with linear interpolation. Note that the extra expense of dataset augmentation is present mainly at the training time as more data points are used. The test time remains almost unaffected, however, now a deeper forest could be learnt. As we will see later in the results, the benefit of dataset augmentation for this forest-based method is clear, yielding a solid 12.14% gain in the F1 score ($F1 = 2(precision.recall)/(precision + recall)$). Results using this augmented dataset are presented in the evaluation as **Augmented miniatures + forest**.

2.5.5 Forest parameter selection

We first trained and tested this forest-based algorithm on a subset of our dataset consisting of 960 image slices from 100 cardiac patients (SAX: 540, 4CH: 140, 2CH: 112, 3CH: 107, LVOT: 9) coming from the DETERMINE study (Kadish et al., 2009) and obtained via the Cardiac atlas project infrastructure (Fonseca et al., 2011) (www.cardiacatlas.org). Through the augmentation we obtained 51894 training images in total.

We ran a 25-fold cross validation by dividing the dataset on patient identifiers to prevent biasing our results due to repeated view acquisitions and other acquisition similarities. This means that images from the same patient (despite being from

different views) and therefore also from the same acquisition session never appeared in the training and testing set at the same time. We observed that, there is not much gain in classification accuracy beyond 80 trees in the forest (see Figure 2.8). As increasing the number of trees usually does not hurt the performance, we used 160 trees in our experiments, to stay safely in the saturated area.

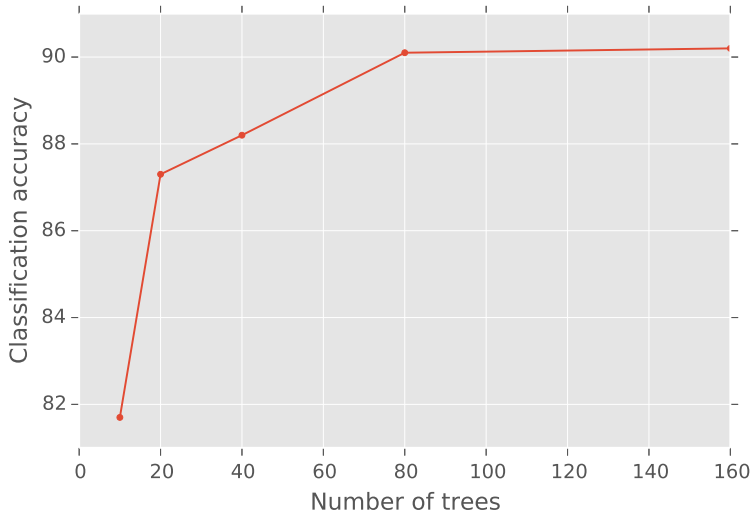


Figure 2.8: We selected the optimal number of trees via cross validation. Increasing the number of trees from 10 to 160 improves the prediction accuracy, however at the expense of increased training and classification computational cost. Beyond 80 trees, the forest reaches a plateau without further benefit.

2.6 Convolutional neural networks for view recognition

A larger training dataset could capture even more of these variations in appearance than just pure augmentation. Our forest-based method initially performed quite well (Margeta et al., 2014) but there was still room for amelioration.

The current image recognition works use the ImageNet dataset consisting of 1.2 million images (Russakovsky et al., 2014) to train high capacity models predicting labels over 1000 categories. We will show that with good initialisation we can use these high capacity models based on CNNs also to predict cardiac planes of acquisition, using our much smaller dataset to train, and improve on the performance of the forest-based approach.

CNNs are another powerful machine learning tool. They were originally inspired by a simple biological model of vision. The CNNs have been around for several decades, first introduced by Fukushima (1980) and later successfully applied to handwritten digit recognition (LeCun et al., 1989). For a detailed history of the neural networks and the CNNs see Schmidhuber (2015).

Recently, there has been a renewed surge in interest in the CNNs, allowed by rapid pace of improvements and understanding of their theoretical foundations. Their most recent applications significantly outperform previous approaches in many difficult machine learning problems and in some cases reach near human level performance — *e.g.*, in handwritten digit recognition (Cireşan et al., 2010), traffic sign recognition (Cireşan et al., 2012), face identification (Taigman et al., 2014), general image recognition (Ioffe and Szegedy, 2015; He et al., 2015), and speech recognition (Hannun et al., 2014). However, not much has changed in the original principles of the CNNs since first introduced to image processing.

We have just got better at optimisation (Bottou, 2010; Bengio, 2012; Dauphin et al., 2014), large training datasets became available (Russakovsky et al., 2014; Lin et al., 2014), we have greater processing power available, found good ways to initialise the networks (Glorot and Bengio, 2010; He et al., 2015; Ioffe and Szegedy, 2015), and overcame problems of diminishing gradients and with overfitting (Hinton et al., 2012; Ioffe and Szegedy, 2015).

This renewed interest, especially in image recognition, is mainly driven by the seminal works of Ciresan et al. (2011) and Krizhevsky et al. (2012), and is fuelled by the large ImageNet dataset (Deng et al., 2009) for image recognition and better hardware. The ImageNet Large scale visual recognition competition (Russakovsky et al., 2014) has become the battlefield in machine learning since. Many significant improvements have followed. Fortunately, we start to see similar efforts in medical imaging now.

2.6.1 Layers of the Convolutional Neural networks

In essence, the CNN is still just a hierarchical bank of convolutional filters combined with some nonlinearity in between, and whose parameters are optimised with gradient descent. It is composed of several layers, most frequently sequential stacked where outputs of the lower layers are fed into inputs of the upper ones. There are several layer types being currently used.

Convolutional layers

Convolution is a frequently used operation to enhance various aspects of images such as gradient orientation or a particular texture. It is the main element in CNNs. The convolutional layers map the input multichannel images into new multichannel images by convolving them with a trainable filter bank (see Fig. 2.9).

The output multichannel images from the convolutional layer are then transformed with a non-linear activation function σ.

Non-linear transformation - activation function

Most of the layers in the network are linear. Without any nonlinearity in between their combination can be reduced into a single linear operation. In the last few years, the Rectified linear unit (ReLU) nonlinearity (Krizhevsky et al., 2012; Malik

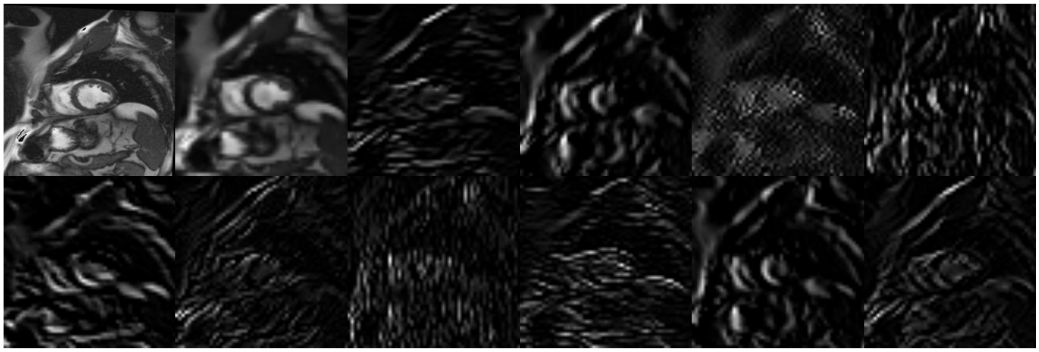


Figure 2.9: Example SAX image (top left) and channels obtained by convolving the source image with filters from the first convolutional layer of our convolutional network CardioViewNet detailed in Section 2.6.5. The first layer mainly picks up low level information such as edges at various orientations.

and Perona, 1990) has become quite popular. It simply clips negative input values to 0:

$$\sigma(x) = \max(0, x) \tag{2.3}$$

This nonlinearity was shown to dramatically speed up training of the network.

Local response aggregation via max-pooling

Max-pooling is another important operation in the CNNs. It can be seen as a form of nonlinear down-sampling. The role of the max-pooling operation is similar to its role in the previously described forest-based method with miniatures in Section 2.5. Max-pooling aggregates local responses within rectangular neighbourhoods and for each neighbourhood outputs the maximum response value. This adds some invariance to small image transformations. The output of the max-pooling operation creates a new set of image channels which are then fed into another layer of convolutions and nonlinearities.

Fully connected layers

However, the final layers of the network are usually not convolutional. We still need to transform the multi-channel images into class predictions. The image channels are therefore fed into the fully connected (fc) layers. The fully connected layers often reduce the data dimensionality and in the final layer produce the class activation vector (one element for each target class).

The multi-channel input images are first reshaped into flat vectors x_i^m (i is datapoint index and m is index of the layer). The fully connected layers then transform this vector into a vector of new length via matrix multiplication. This layer is parametrised by a matrix W^m and a bias vector b^m:

$$x_i^{m+1} = W^m x_i^m + b^m \tag{2.4}$$

Soft-max to estimate cardiac view probabilities

To obtain the target label probabilities $y_i \in \mathbb{R}^K$ (K is the number of classes), the output vector of the final fully-connected layer $a_i = x_i^{\text{last}}$ is transformed by the soft-max function.

$$y_i[k] = \frac{e^{a_i[k]}}{\sum_{k=1}^{K} e^{a_i[k]}} \tag{2.5}$$

To predict the view of a never seen image, it has to be passed through the layers of the CNN until the soft-max output is reached. The most likely view is then chosen. But first, the parameters of the network must be optimised.

Loss function for classification

With decision forests, a surrogate loss is greedily optimised at every node instead of a global loss for the classifier. The CNNs on the other hand are an end-to-end approach, *i.e.*, they learn optimal feature representation directly from raw image intensity values by minimizing some global loss $L(\theta_t)$. For classification we use the cross-entropy loss:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} z_i^k \ln y_i^k \tag{2.6}$$

Here, i indexes over the data-points in the training batch and k indexes over the views. y_i is the vector with predicted probabilities for each view (depending on parameters of the network θ_t and the input image). z_i is the ground-truth label indicator vector. *E.g.* for an image labelled as LVOT where the complete set of views is (2CH, 3CH, 4CH, LVOT, SAX), its ground-truth vector z_i is encoded as $(0, 0, 0, 1, 0)$.

Note that in a binary case ($K = 2$) this becomes the better known logistic regression loss:

$$L = -\frac{1}{N} \sum_{i}^{N} z_i \ln y_i + (1 - z_i) \ln(1 - y_i) \tag{2.7}$$

It is common, to also add a term regularising the network parameters into the loss function. This is in our case a simple L_2 penalty on the parameters of the network ($L_d = 0.5\|\theta\|_2^2$) which we add to the total loss weighted by parameter β, also called weight decay.

2.6.2 Training CNNs with Stochastic gradient descent

If the cost function is differentiable, we should be able to jiggle parameters of the network θ_t to minimize this cost. The parameters of the network such as weights of the convolutional filter kernels are then optimised through backpropagation in the direction of the steepest descent.

The batch Stochastic gradient descent (SGD) (Bottou, 2010, 2012) algorithm has been quite successful to optimise large neural networks. The principle is very similar to the gradient descent algorithm except that the descent direction is estimated from a smaller random batch only (instead of using the whole training dataset at the same time). This not only dramatically speeds up the optimisation, it also reduces the chances of getting trapped in local minima. The batch SGD is shown in Algorithm 1.

Algorithm 1: Batch Stochastic gradient descent (SGD) algorithm

while *stopping criteria not satisfied* **do**
  Take a small random batch of examples from different classes;
  Compute average descent direction with respect to these images;
  Update parameters with the update rule;
end

The stochastic gradient descent updates parameters of the network θ at each iteration using the following update rule:

$$V_{t+1} = \mu V_t - \alpha_t \nabla L_{\theta_t} \tag{2.8}$$

$$\theta_{t+1} = \theta_t + V_{t+1} \tag{2.9}$$

Momentum μ helps to regularize the locally estimated steepest descent direction of the cost function ∇L_{θ_t} with the previous weight update V_t which results in more stable and faster convergence. For zero momentum and when using all training images in the batch, this equals to the standard gradient descent algorithm with learning rate α.

As the network learns the optimal parameters through training in mini-batches, the average descent direction for the mini-batch is used. For generalisation and convergence, care has to be taken to avoid too many images from the same subject or the same class in one batch.

The batch training allows us to use massive (possibly infinite) quantities of training data or data generated on the fly by data augmentation. While it is technically possible to do similar data generation at each node for a decision forest, the forest structure is fixed in training and is usually not revisited afterwards.The SGD makes it simpler to continuously keep updating all of the network parameters when new data become available.

2.6.3 Network architecture

The state of the art in image recognition has been heavily influenced by the previously mentioned seminal work of Krizhevsky et al. (2012) who trained a large (60 million parameters) CNN on a massive dataset consisting of 1.2 million images and 1000 target classes. They employed two major improvements to the previously used CNNs: Rectified linear unit nonlinearity to improve convergence, and Dropout (Hinton et al., 2012) to reduce the effect of overfitting.

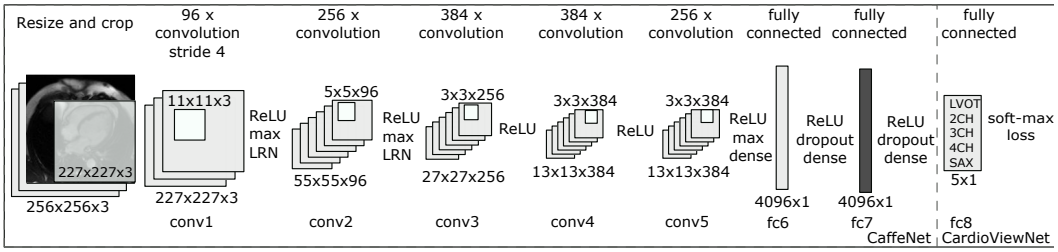


Figure 2.10: Our CardioViewNet is based on CaffeNet network structure and is adapted for cardiac view recognition. We initialised the network weights from a pretrained CaffeNet (Jia et al., 2014). We then replaced the last 1000-dimensional logistic regression layer (previous fc8) with a 5-dimensional for 5 cardiac views. Then we fine-tuned the network with our data. We also extracted features from the last 4096-dimensional fully connected layer fc7 (in dark gray) from both CaffeNet and our fine-tuned CardioViewNet and used them with a decision forest and a linear SVM classifier. We achieve the best performance with our fine-tuned network directly.

We therefore use this widely adopted neural network architecture (see Figure 2.10) as implemented in the Caffe deep learning framework (Jia et al., 2014) under the bvlc_reference_caffenet acronym (CaffeNet in short). The CaffeNet implementation differs from Krizhevsky's AlexNet in the order of Local response normalisation (LRN) and max-pooling operations. Not only AlexNet/CaffeNet is the winning entry of the ImageNet LSVRC 2012 competition, the weights and definitions of this network are publicly available thus improve reproducibility of this work and reduce computational time needed for the training.

2.6.4 Reusing CNN features tuned for visual recognition

The CNNs trained on the larger datasets are often used for feature extraction. Similarly to the work of Karayev et al. (2014) for photography style recognition or Sharif et al. (2014) for a variety of image recognition tasks, we use CNN features from the fully connected layer of the network pretrained on the ImageNet LSVRC (Russakovsky et al., 2014) dataset. This time, to recognise cardiac acquisition planes. The fully connected layer (in our case **fc7** — see Figure 2.10) helps us to describe the cardiac image slices with 4096-dimensional feature vectors.

Before putting the cardiac images through the network, simple preprocessing is done. The cardiac images are resized to 256×256 squares regardless of their input dimensions. Since the CaffeNet pretrained with ImageNet takes Red-Green-Blue (RGB) images as input, we simply replicate the 2D cardiac slices into each colour channel. We compute a pixel-wise mean image for the whole cardiac dataset and subtract it from all training and testing images prior to entering the CNN. This centers image intensities around zero and serves as a very simple intensity normalisation. As we found later, the pixel-wise mean image across all classes

is almost uniform in the central region of interest and a scalar value could be subtracted instead. The central ($227 \times 227 \times 3$) crop of this image is then fed forward through the network.

We then use the extracted CaffeNet fc7 features with a linear SVM classifier (Cortes and Vapnik, 1995) to predict the cardiac views. We ran cross-validation on the training subset folds to maximise the prediction precision. This helped us to choose the penalty parameter C of the SVM classifier from standard set of values [0.1, 1, 10, 100, 1000] as C = 1. We report results of this method as **CaffeNet fc7 + SVM**.

Similarly, we trained a classification forest (with 64 tested features per node and 1000 trees) using these CNN features (instead of image miniatures) and report these results as **CaffeNet fc7 + forest**.

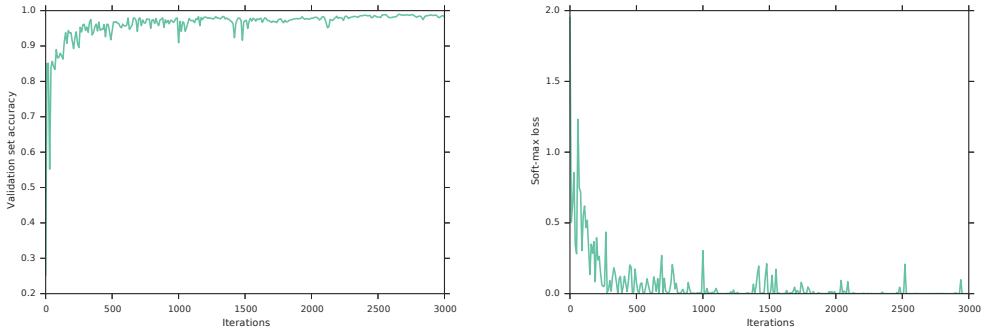
These features were adapted to a general object recognition task and come from a CNN that never saw a cardiac MR image to optimise its weights. As we will show in Table 2.1, this already performs quite well for the cardiac view recognition, in fact, much better than our method using classification forests with image miniatures. In the following, we will show how we can further improve performance by adapting the CNN weights to our target domain.

2.6.5 CardioViewNet architecture and parameter fine-tuning

In practice, many examples are often needed to train a large capacity neural network. However, by starting from the weights of a pretrained network we can just fine-tune the network parameters with new data and adapt it to the new target domain. Here, we use the pretrained CaffeNet (Jia et al., 2014) and replace the last 1000-class multinomial regression layer with a 5-class one (see Figure 2.10).

The net is fine-tuned with stochastic gradient descent. We use higher learning rate α for parameters in the newly added layer (10^{-2}) and smaller (10^{-3}) in the rest of the network — we want to mainly optimise the newly added layer and preserve the more general low level preprocessing ones. We choose momentum μ of 0.9 in the SGD optimiser, and a small weight decay β (10^{-4}). The batch size used in each iteration was 32 images and the step size is kept constant for the whole training. As done in the previous section, the images were resized to $256 \times 256 \times 3$ and the mean image was subtracted from them prior to the training.

At each iteration, a batch of 32 random (not necessarily central) $227 \times 227 \times 3$ crops is extracted from the resized cardiac slices and is fed forward through the network. Compared to the implementation of the forest-based method where all augmented images were precomputed, the translations are cheaply generated on the fly at each iteration. Already after 3000 iterations the prediction error on the validation dataset reaches a plateau and further improvements are only marginal (see Figure 2.11). We stop the optimisation at 8000 iterations and pick this model in our experiments as it yields the best performance. To reduce overfitting, we use the Dropout strategy (Hinton et al., 2012) in the fully connected layers *fc6* and *fc7* with probability of dropping output of a neuron to be 0.5.



(a) Validation set accuracy. (b) Soft-max loss.

Figure 2.11: Fine-tuning our CardioViewNet model, the prediction accuracy and classification loss rapidly converge to the best performance.

The fine-tuning is quite efficient and 8000 SGD iterations take approximately 4 hours on a single NVIDIA Tesla M2050 Graphical processing unit (GPU). Results of this method are presented as **CardioViewNet**.

We also show results for prediction of an SVM classifier using fc7 features extracted from the fine-tuned network as **CardioViewNet fc7 + SVM**. In other words, we replace the 1000-class multinomial regression layer of the fine-tuned network by a linear SVM classifier. Results using a classification forest instead are listed as **CardioViewNet fc7 + forest**. The possibility to replace the final classifier is important for quick retraining with additional views without extra fine-tuning. In addition to the training set augmentation, we perform oversampling at the test time, *i.e.*, average predictions of ten 227×227 image crops: the central crop and the four 227×227 corner aligned crops and their horizontally flipped versions (vertically flipped images are rare). We report these results as **CardioViewNet + oversample**. We will see that this improves performance on an independent dataset.

2.6.6 Training the network from scratch

Good initialisation of the network is important and the CaffeNet trained on the ImageNet dataset helps us with that. The initial motivation behind the fine-tuning was that there were few images in our dataset to train the whole network from scratch. While the number of images is certainly smaller than in the ImageNet dataset, our target domain is also much simpler. We are predicting only 5 classes whose appearance variability is lower than the one across the ImageNet classes (*e.g.* variability of felines in different poses and appearance all labelled as a cat).

To test whether there is any value in the fine-tuning instead of learning the network parameters from scratch, we train from the ground up two networks. First, a simpler LeNet-5 network (shown as **LeNet-5 from scratch**) previously designed for handwritten digit recognition (LeCun et al., 1998), with the last layer adapted to a 5 class target (similarly to what was done for the CardioViewNet). Again, for

better reproducibility, we use the network definition from the Caffe library. The second network architecture is the CardioViewNet (**CardioViewNet from scratch**). We found the choice of the learning rate (10^{-3} for CardioViewNet and 10^{-5} for LeNet-5, both using batch sizes of 32) and good random initialisation to be crucial to ensure convergence.

We chose the learning rate by trial and error strategy of Bengio (2012), *i.e.*, we reduce the learning rate until the network starts to reduce the loss without diverging. We initialised the weights with the procedure described by He et al. (2015) that is well suited for networks with the ReLU nonlinearity.

2.7 Validation

We trained and validated all the presented methods on a dataset of slices from 200 patients (2CH: 235, 3CH: 225, 4CH: 280, LVOT: 12, SAX: 2516) from a multi-center study on post myocardial infarction hearts **DETERMINE** (Kadish et al. (2009)) from SSFP acquisition sequence (see Fig. 2.12 for illustration). The LVOT views are severely under-represented and served us only as a test case for learning from very few examples. They are not taken into the account when computing the mean scores in the results as it would make unrealistic variation between the methods for fair comparison.

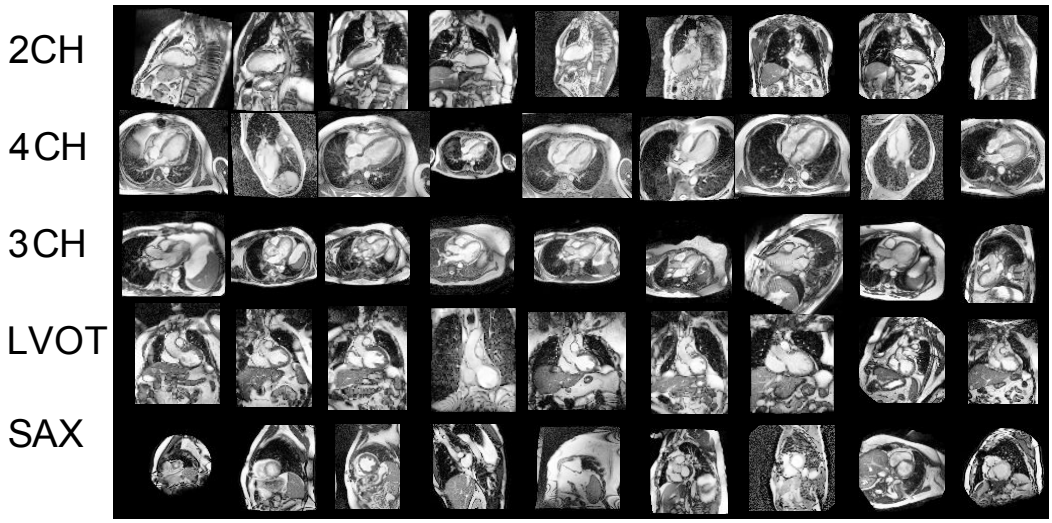


Figure 2.12: Typical examples of the training images from the DETERMINE dataset for different views. Note the acquisition and patient differences. In addition, the short axis slices cover the heart from the apex to the base with quite different appearances.

We ran a randomized 10-fold cross-validation by taking a random subset of 90% of the patients (rather than image slices) for training and using the remaining 10% for validation. The patient splits guarantee that repeated acquisitions from the same patient, that are occasionally present in the dataset, never appear in both the

training and the validation set together and do not bias our results. Classification accuracy is not a good measure for imbalanced datasets as performance on the dominant class (*i.e.* short axis) can obfuscate the true performance. Therefore, in this chapter we report means and standard deviations of average (averaging is done across the classes) precisions, recalls and F1 scores. In the context of content-based retrieval, these measures can be interpreted as following: The precision (also known as positive predictive value or false positive rate is defined as TP / (TP + FP)) is the fraction of relevant images (having the same view as the query view) out of all returned images. The recall (also known as sensitivity or true positive rate is defined as TP / (TP + FN)) is the probability of retrieving a relevant image out of all existing relevant images. TP is the number of true positives, FP the number of true negatives, and FN the number of false negatives. The F1 score is the harmonic mean of the precision and the recall.

To study the robustness of the presented algorithms against the dataset bias, we did the training exclusively on images from the DETERMINE study (Kadish et al. (2009)) and tested them on a completely independent dataset — the **STACOM motion tracking challenge** (Tobon-Gomez et al., 2013a) (KCL in short) containing SSFP slices from 15 patients (2CH: 15, 4CH: 15, SAX: 207). This allows us to evaluate performance on the 3 cardiac views present. The KCL dataset consists of images from healthy volunteers. The images are in general of higher and more uniform quality and with more consistently chosen regions of interest. We invite the interested readers to look at this open access dataset through the Cardiac Atlas Project website (Fonseca et al., 2011).

2.8 Results and discussion

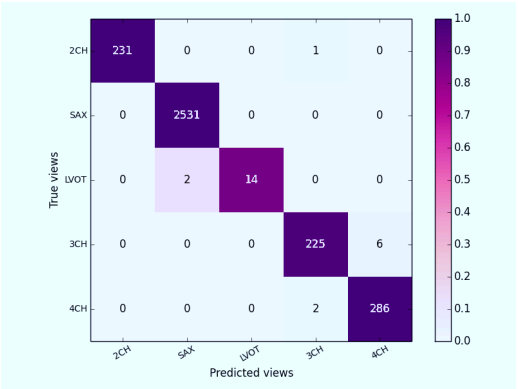
Here, we present prediction results for the DICOM-based and image-content-based methods. The mean average F1 scores, precisions and recalls are summarised in Table 2.1 and total confusion matrices for the two best methods from each family (DICOM-based and Image-based) are shown in Figure 2.13.

We confirm findings from the previous work that cardiac views can be predicted from image plane normals (Zhou et al., 2012) but require presence of the relevant DICOM tag. The DETERMINE dataset turned out to be more challenging for the miniature-based method using classification forests. It is clear, however, that data augmentation helps to boost recognition performance. On the contrary, the performance of the CNN features from CaffeNet on the cardiac view recognition problem is quite remarkable. These features were not trained for cardiac MR images, yet they perform better than most methods with handcrafted features. They most likely encode local texture statistics helping the prediction. Adding some texture channels to the miniature method could therefore further improve its performance.

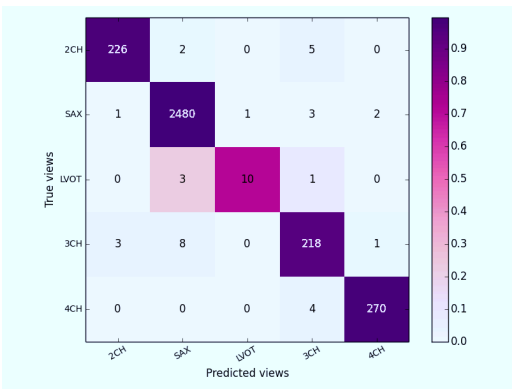
The quality of predictions using the fine-tuned CardioViewNet is almost on par with the approach using DICOM derived information and significantly outperforms the forest-based method using image miniatures (Margeta et al., 2014) while not

	DETERMINE			KCL		
	F1 score	precision	recall	F1 score	precision	recall
DICOM tag based prediction						
Plane normal + forest (2.4)	**99.14 ± 1.23**	**98.91 ± 1.13**	**99.20 ± 1.53**	**99.08 ± 0.46**	**98.76 ± 0.78**	**99.16 ± 0.32**
Image content-based prediction						
Miniatures + forest (2.5)	59.33 ± 4.15	62.13 ± 5.74	55.61 ± 3.80	39.36 ± 1.75	42.71 ± 4.63	37.98 ± 4.39
Augmented miniatures + forest (2.5.4)	71.46 ± 2.68	72.33 ± 2.77	68.01 ± 2.65	48.87 ± 2.02	54.74 ± 2.33	43.77 ± 1.98
CaffeNet fc7 + forest (2.6.4)	75.94 ± 4.50	94.03 ± 1.75	69.08 ± 4.53	88.09 ± 1.29	92.60 ± 1.63	86.86 ± 1.08
CaffeNet fc7 + SVM (2.6.4)	91.86 ± 4.33	92.48 ± 3.98	91.61 ± 4.71	86.72 ± 1.49	86.70 ± 2.19	87.30 ± 1.08
CardioViewNet fc7 + forest (2.6.5)	97.48 ± 2.34	98.28 ± 1.84	96.81 ± 3.03	93.43 ± 2.05	95.79 ± 3.10	91.67 ± 2.34
CardioViewNet fc7 + SVM (2.6.5)	97.39 ± 2.27	**98.37 ± 1.88**	96.65 ± 2.77	88.40 ± 1.84	**97.51 ± 2.02**	88.95 ± 4.44
CardioViewNet (2.6.5)	**97.66 ± 2.04**	97.82 ± 1.93	**97.62 ± 2.37**	91.01 ± 3.29	92.23 ± 3.80	90.57 ± 3.26
CardioViewNet oversample (2.6.5)	97.53 ± 2.06	97.98 ± 2.12	97.30 ± 2.30	**93.50 ± 3.12**	95.31 ± 5.17	**92.62 ± 2.30**
LeNet-5 from scratch (2.6.6)	69.59 ± 5.12	76.79 ± 5.40	67.89 ± 4.26	63.81 ± 4.88	72.03 ± 9.67	60.41 ± 4.53
CardioViewNet from scratch (2.6.6)	92.36 ± 3.51	92.63 ± 4.44	92.97 ± 2.79	79.72 ± 3.65	80.39 ± 5.39	81.65 ± 3.64

Table 2.1: Evaluation of the algorithms in the two groups of algorithms, we highlight in bold the best performance from each group. Note: References to relevant sections in the chapter with more details on each algorithm in the parentheses. We computed average of individual view F1 scores, precisions and recalls for each fold (except for the underrepresented LVOT) for the two datasets. Here, we display means and standard deviations of these average scores across all 10 folds. The prediction using classification forests on the DICOM orientation vector is the best performing method. However, from purely image-based methods, the fine-tuned convolutional network CardioViewNet outperforms the rest.



(a) Predictions from DICOM-derived image normals (Plane normal + forest).



(b) Predictions from image content with a fine-tuned CNN (CardioViewNet).

Figure 2.13: Sum of the confusion matrices over the 10 folds of our cross-validation on the DETERMINE dataset for the two best models (one using DICOM normal information and the best image-based predictor using our fine-tuned neural network).

requiring to train any extra landmark detectors as in Zhou et al. (2012). As features extracted from the CardioViewNet perform well even when used with external classifiers, they could be likely used to learn extra view recognisers without additional fine-tuning on these new views.

In Figure 2.14, we present examples of some of the least confident correct predictions using the fine-tuned CardioViewNet. It is important to note that the softmax output of the neural network not only returns the label but also some measure of confidence in the prediction. The incorrectly classified images (see Figure 2.15) often belong to views more difficult to recognise even for a human observer and the second best prediction is often the correct result.

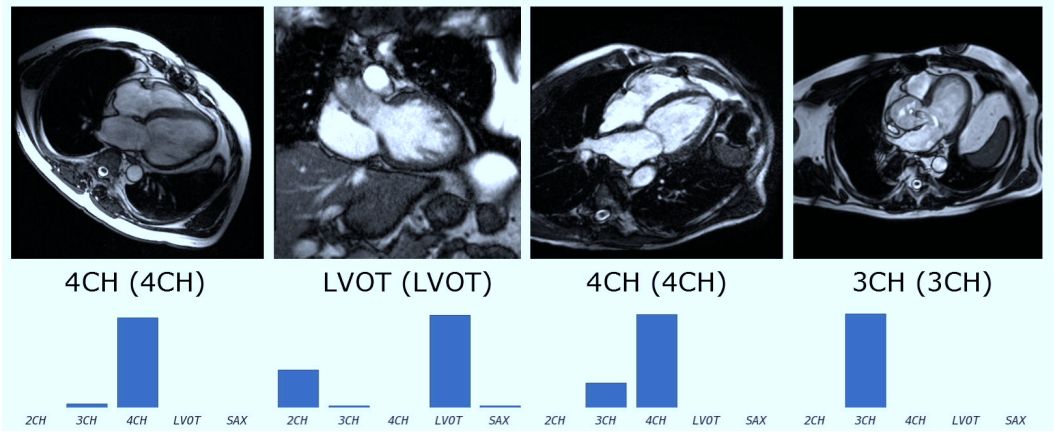


Figure 2.14: Examples for some of the least confident correct (the normal predictions are usually very peaky) predictions using CardioViewNet. Predicted and true (in parentheses) labels shown under the images. Below them are view-specific prediction confidences.

The predictions on the KCL dataset are naturally slightly worse as some differences between the studies still exist. We have observed (as shown in Table 2.1) that test time oversampling (averaging predictions of the central and corner patches and their horizontally flipped versions) improves the scores for this dataset while it does not improve the quality of predictions on the DETERMINE dataset. This might indicate that better thought dataset augmentation strategies, oversampling, or image normalisation might further improve the cross-dataset generalisation.

When training from scratch, the performance does not seem to improve beyond 10,000 iterations and we pause the backpropagation there. Although the performance is much lower than for the fine-tuned networks, the network learns to predict the cardiac views. For the LeNet-5 model trained from scratch, the performance closely follows the CaffeNet fc7 + SVM model. We did not observe any benefit of using forests (at least when using orthogonal splits) over the linear SVM classifier when using features from the convolutional nets, and the forests perform in general slightly worse. We would likely improve their performance using oblique splits (Menze et al., 2011; Criminisi et al., 2011b). In our case, the use of the

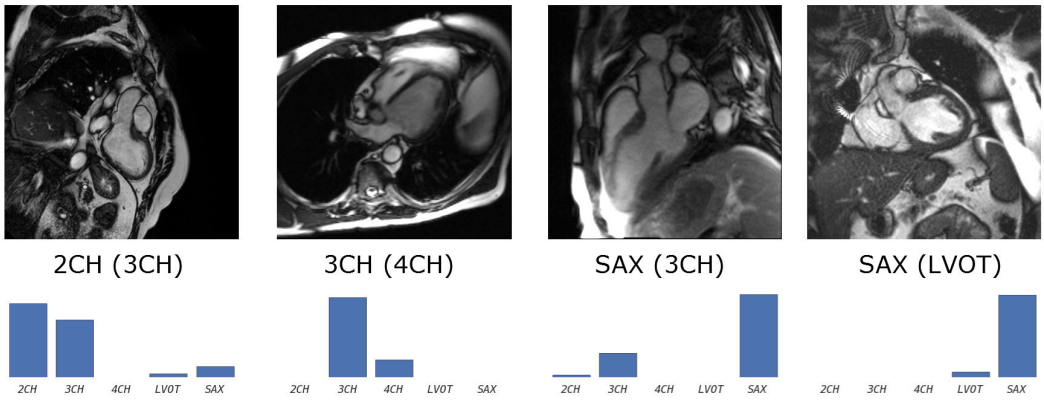


Figure 2.15: Example misclassifications using CardioViewNet. Predicted and true label (in parentheses) are indicated under the images and below them are view-specific prediction confidences. The failures typically happen with less typical acquisitions and truly ambiguous images. The misclassified 2CH image indeed looks like a 3CH image except that the left atrium is missing and the ventricle is acquired with non-standard oblique orientations. Similarly for the 4CH view, the right atrium is missing and one can already see parts of the outflow tract branching out of the left ventricle typical for a 3CH view. The 3CH view is captured with a detailed region of interest not very common in the dataset. Extra data augmentation could probably help to fix this case. Finally, the LVOT views are severely under-represented and can be confused with basal short axis views. Note that for all of these cases, the correct prediction is in the top two most likely views.

CardioViewNet trained from scratch is a better choice than using the generic *fc7* features from CaffeNet. However, the training is significantly more time-consuming. The fine-tuning approach is the clear winner among the image-based methods in terms of performance.

2.9 Conclusion and perspectives

Convolutional neural networks and features extracted from them seem to work remarkably well for medical images. As large datasets to train complex models are often not available, retargeting the domain of a previously trained model by fine-tuning can help. This speeds up the learning process and achieves better predictions. Even models trained for general object recognition might be a great baseline to start. In our case, doing network surgery by replacing the final layer and fine-tuning the pretrained model allowed us to make significant progress in cardiac view recognition from image content without handcrafting the features or training with extra annotations.

This also allowed us to gain performance over models learnt from scratch. However, even the performance of models learnt from scratch is very encouraging for further exploration. More recent and much deeper network architectures (*e.g.* VGG (Simonyan and Zisserman, 2014) or GoogLeNet (Szegedy et al., 2014; Ioffe and Szegedy, 2015)) have achieved significant improvements in image recognition performance over AlexNet, and would likely help to further improve performance of the view recognition.

Features extracted from our network should be useful as descriptors for new views (*e.g.* pathology specific views such as those used in congenital heart diseases) and acquisition sequences other than SSFP, but also to recognise the acquisition sequences themselves. Several recommendations (Friedrich et al., 2014; Kramer et al., 2013) have been recently proposed to simplify and remove vendor specific naming of acquisition sequences. Such efforts are crucial for improved communication in cardiology. The methods presented in this chapter could help us to learn the mapping from image content to these standardised nomenclatures and further clean up inconsistent and missing metadata for better organisation and search within cardiac databases. As we will see later in this thesis, the features from our fine-tuned network can also be used to predict positions of cardiac landmarks (Appendix B).

These are important additions to the arsenal of tools for handling noisy metadata in our datasets and are already helping us to organise collections of cardiac images. In the future, our method can be used for semantic image retrieval and parsing of medical literature. Recognition at an image level (image classification) can be also extremely useful for the costly problem of quality control (as done, *e.g.*, by Criminisi et al. (2011a) to verify contrast enhanced images). For example, “I have asked my radiographer to take all standard views of the heart. Did he do it correctly?”

Chapter 3

Segmenting cardiac images with classification forests

Contents

Based on our paper on left ventricle segmentation (Margeta et al., 2012) whose results we contributed to the left ventricle collation study by Suinesiaputra et al. (2014b). Our left atrial segmentation work was published in (Margeta et al., 2013) and was contributed to the benchmark study led by Tobon-Gomez et al. (2015).

Chapter overview

The most commonly used clinical indices for cardiac disease quantification are based on geometrical measurements of structures in the images, such as cavity volumes, myocardial masses or wall thicknesses. It would be quite useful to have them stored in the cardiac databases alongside the images and use them to select patient cohorts for clinical studies. These measurements are often computed from segmentations of these cardiac structures. The segmentation also forms the basis for understanding the cardiac anatomy by a computer.

In this chapter we present a flexible machine learning based method for CMR image segmentation of two of the most important cardiac structures. First, we learn how to segment the left ventricles from dynamic 3D+t SAX images directly from voxel-wise ground-truth by training classification forests with spatio-temporal context rich features. We then show that this method can be easily adapted to other segmentation problems such as the left atria.

3.1 Segmentation of the left ventricle

3.1.1 Measurements in cardiac magnetic resonance imaging

The left ventricle plays a fundamental role in circulation of oxygenated blood to the body. To assess its function in clinical practice, the guidelines (Kramer et al., 2013; Fratz et al., 2013; Friedrich et al., 2012) suggest to measure, calculate and report several measures. Many of these are based on ventricular volume and mass measurements at reference cardiac phases. These can be then compared with reference values (Schulz-Menger et al., 2013; Maceira et al., 2006; Hudsmith†et al., 2005; Chuang et al., 2014; Buechel et al., 2009) for CMR.

Compared to CT, CMR imaging offers superior temporal resolution, excellent soft tissue contrast, no ionising radiation, and a vast flexibility in image acquisition characteristics. On the other side, MRI scans often yield significantly lower resolution in the plane orthogonal to the plane of acquisition, the images can suffer from magnetic field inhomogeneities and respiration artefacts can manifest as slice shifts. Moreover, the lack of standard units (compared to the Hounsfield scale in CT) makes it more challenging to directly apply most of the intensity based segmentation techniques.

Still, to calculate the quantities, accurate delineations of the myocardium and the cavity borders are often necessary. Such manual delineations are still labour intensive (2 minutes per cardiac phase), especially when applied retrospectively to the previously acquired data. The inter-rater variability of manual segmentations (and derived measures) is also quite noticeable (Suinesiaputra et al., 2015) and can be likely improved by finding a consensus and training of the raters.

Moreover, the differences in measures between inclusion or exclusion of papillary muscles and trabeculations can be also significant (Papavassiliu et al., 2005; Winter et al., 2008; Chuang et al., 2011; Han et al., 2008).

To obtain consistent measurements, to be able to resegment the images when the guidelines change (or segment all variations), and to save clinicians' time, automatic segmentation techniques are desirable.

3.1.2 Previous work

There have been numerous examples of semi-automatic or automatic LV segmentation algorithms to this date. See Petitjean et al. (2011) for a thorough review. Here, we mention only a selection of the notable automatic ones.

Shape based segmentation

Two beautiful frameworks for automatic whole heart cardiac image segmentation were proposed by Zheng et al. (2008); Ecabert et al. (2008). Machine learning based detection of the heart is followed by an iterative refinement of the segmentation regularised with probabilistic shape models.

These methods offer an excellent performance and regularisation of the segmentations. In addition, they naturally complete out of image shapes, *e.g.*, in the apical or atrial regions and divide the image into the compartments even when no image gradient between them is present. As they are often based on Principal Component Analysis (PCA), they perform slighly worse on data that cannot be well represented by the linear combination of the bases and do not work at all when the topology significantly changes (such as for congenital heart diseases). In these cases, additional models have to be built specifically for each topology.

Atlas-based segmentation

The atlas-based segmentation techniques (Shi et al., 2011) use non-rigid image registration to deform and position a previously segmented image (or images for a multi-atlas segmentation) to better match the target image. The labels are then propagated from the atlas to the target image. These methods rely on robust non-rigid registration.

The inter-patient non-rigid registration is in general a difficult problem due to the extra-cardiac structures which misguide the registration metric. Masking of the images (with cardiac segmentation) might be needed (Ou et al., 2012). To avoid the chicken and egg problem (segmenting the heart in order to register in order to segment) Shi et al. (2011), proposed to register the images and use only the information contained within the cardiac region of interest. This region is detected with a standard trainable object detector (Viola and Jones, 2004).

Voxel-wise semantic segmentation

Instead of shape-based or atlas-based methods, the cardiac segmentation can be posed as a semantic image segmentation problem (Shotton et al., 2008). Here, each pixel is assigned a class label (myocardium or background). In the work of Shotton

et al. (2008), classification forests and simple pairwise difference/sum features were used. Lempitsky et al. (2009) modified this algorithm for LV myocardium segmentation from 3D ultrasound sequences in near real time and extracted Haar-like 3D box features. Later, Geremia et al. (2011) extended this method for the segmentation of multiple sclerosis lesions from multi-channel MRI.

3.1.3 Overview of our method

In this chapter, we propose a fully automated voxel-wise segmentation method for 3D+t cardiac images inspired by the above techniques. We avoid the need for robust registration to an atlas (Shi et al., 2011), to build a statistical model (Zheng et al., 2008; Ecabert et al., 2008; Lu et al., 2011), or design a highly specialised cardiac segmentation algorithm.

Instead, the left ventricle segmentation problem is defined as a binary classification of image voxels into myocardium and background. We make no assumptions on shape, appearance, or motion (except for periodicity and temporal ordering) or knowledge about the cardiac phase of the images in the sequence. There is no constraint on the topology *per se* either. We define a set of simple feature families and leave the learning algorithm to automatically figure out which features are relevant for solving the segmentation problem providing only a pixel-wise segmentation ground-truth.

In principle, segmenting any cardiac pathology and tissue can be learnt this way once it is contained in the voxel-wise ground-truth and similar voxels are present within the training dataset. This includes for example the controversial papillary muscles and trabeculations, but also pericardial effusion or surrounding fat.

The previously used decision forest-based segmentation algorithms (Lempitsky et al., 2009; Geremia et al., 2011) rely on features that work best when image intensities and orientations are very similar. To tackle the highly variable dataset, we propose a layered learning approach, where the output of each layer serves a different purpose. The first layer is used to prepare the data (normalised image intensities and pose) for a more semantically meaningful and accurate segmentation task in the second layer. Once the images are intensity- and pose-normalised, we add coordinate features encoding spatial information and then train a second forest layer (Section 3.1.7). This helps the trees to automatically build their own latent shape representations. When results of one learning stage are used as semantic input to improve a second learning stage, this is often referred to as the auto-context (Tu and Bai, 2010).

Our main contribution is a method to segment left ventricular myocardia from cine MR SAX 3D+t sequences with decision forests. We show, how our method can be used for robust inter-patient rigid alignment (Section 3.1.6) and a possible way to tackle the MR intensity standardisation problem (Section 3.1.6). We also introduce temporal dimension into the currently used 3D random features (Section 3.1.5).

Dataset for LV segmentation

STACOM 2011 LV segmentation challenge data (Fonseca et al., 2011) were divided into two sets. Training set (96 3D+t short axis (SA) volumes with manually delineated myocardia at each cardiac phase) and validation sets (5×20 3D+t SA volumes). In this work, we use ground-truth that excludes the papillary muscles and trabeculations from the myocardium. This dataset clearly shows the anatomical variability of heart shape and appearance and some of the main issues of CMR mentioned above.

3.1.4 Layered spatio-temporal decision forests

We train a classification forest as described in the previous chapter (see Section 2.5.1). Instead of representing one 2D slice as a point in the feature space, we consider each voxel as one point in this space.

The goal of the forest is to divide this space and predict label for every voxel as being either the myocardium or the background based on its feature representation θ and context. The binary decision on each feature is $\tau_l < \theta < \tau_h$. At each node, local features and a randomly sampled subset of context-rich features are considered for feature selection. We detail these features in Section 3.1.5.

Bagging strategy for training from 4D datasets

In our approach, we serially train two layers of decision forests, each with the aim to learn how to segment, but using slightly modified training data and features. Training with the complete 3D+t dataset in memory is not always feasible. In Figure 3.1 and in Algorithm 2 we present our bagging strategy to reduce the memory footprint when using larger training datasets (*e.g.* when using the 4D image sequences).

This strategy is repeated for each tree:

Algorithm 2: Our bagging strategy for training with 4D data

for *Every tree in the forest* **do**
  Select a random subset of k 4D volumes from the whole training set;
  Randomly choose a reference 3D frame I^c for each selected 4D volume;
  Select two frames I^{c-o}, I^{c+o} with a fixed offset o on both sides from the reference cardiac image I^c (with periodic temporal wrapping);
  Train the tree using the set of k triplets (I^c, I^{c-o}, I^{c+o});
end

To fit into the memory constraints, the size of the subset for each tree was set to $k = 15$, and only one fixed offset $o = 4$ is used. The value of o was chosen such that the motion between the selected frames is significant even when more stable cardiac phases (end systole or end diastole) are selected as the reference frame and that almost a half of the cardiac cycle could be covered. Note that for $k = 1$ and when no offset images I^{c-o}, I^{c+o} are used, the algorithm becomes similar to

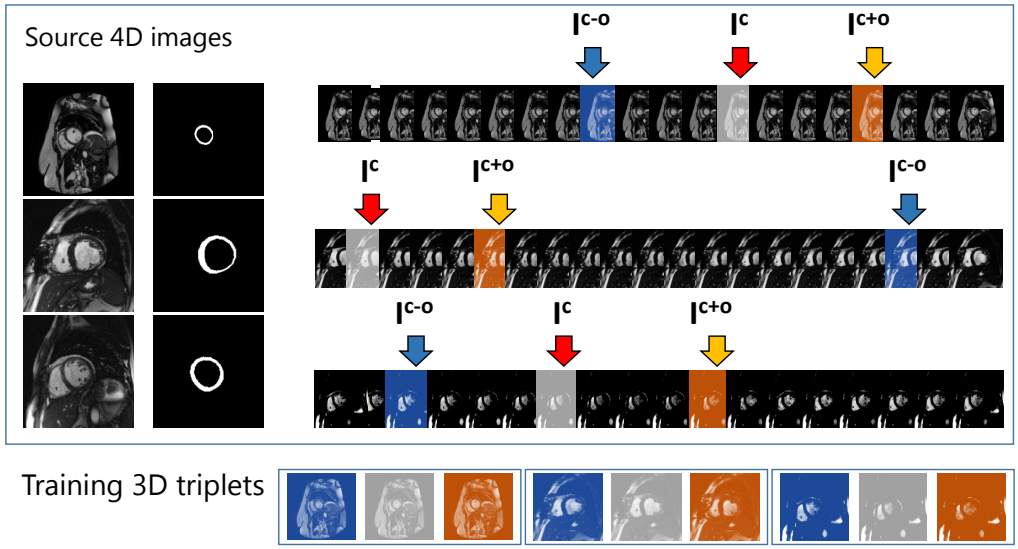


Figure 3.1: Our strategy to train a segmentation forest from 3D+t cardiac data. For each tree we randomly select a subset of training images (3 in this example). For each image we pick a random reference frame and two frames offset by a constant number of frames (while respecting the periodicity of the sequence). We then train the forest on these triplets.

the later introduced Atlas encoding by randomized forests (Zikic et al., 2013). In addition, instead of randomly selecting the reference images for each tree, guided bagging (*e.g.* by training each tree on similar images only) could help to improve the segmentation quality as shown by Lombaert et al. (2014).

3.1.5 Features for left ventricle segmentation

We use several feature families to generate the random feature pool operating on the triplets of frames. These can be seen in Figs. 3.2 to 3.6.

Local features

Proposed in Geremia et al. (2011) as an average of intensities in the vicinity of the tested voxel to deal with noise in magnetic resonance imaging:

$$\theta^{loc}_{I^c}(x) = \theta^{loc}_{I^c}([x,y,z]) = \sum_{x'=x-1}^{x'\leq x+1} \sum_{y'=y-1}^{y'\leq y+1} \sum_{z'=z-1}^{z'\leq z+1} I^c([x',y',z']) \tag{3.1}$$

Although these features are not intensity invariant, they can still quite well reject some highly improbable intensities (*e.g.* bright blood-like voxels will likely never belong to the myocardium).

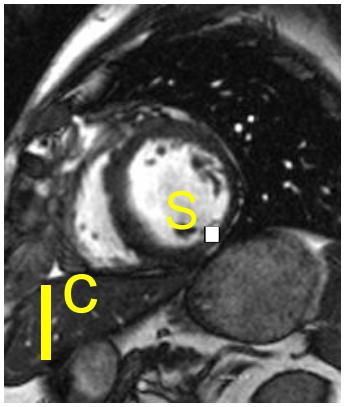


Figure 3.2: Local features ($3 \times 3 \times 3$ box average S around the source voxel in the current frame I^c) Geremia et al. (2011).

Context rich features

Defined also in Geremia et al. (2011), for multichannel MR acquisitions as a difference between the local source image intensity I^S and box averages of remote regions in image I^R:

$$\theta^{CR}_{I^S,I^R}(x) = I^S(x) - \frac{1}{Vol(R_1)} \sum_{x' \in R_1} I^R(x') - \frac{1}{Vol(R_2)} \sum_{x' \in R_2} I^R(x') \qquad (3.2)$$

The 3D regions R_1 and R_2 are randomly sampled in a large neighbourhood around the tested voxel. These capture strong contrast changes and long-range intensity relationships. In our case we define context-rich features as $\theta^{CR}_{I^c,I^c}(x)$.

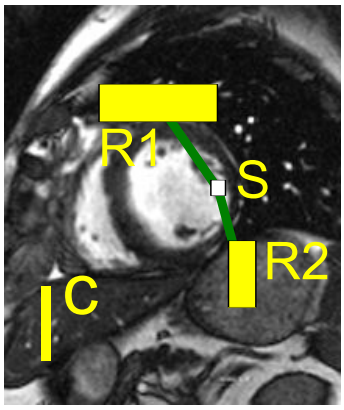


Figure 3.3: Context rich features (Geremia et al., 2011) measuring the difference between source box average S and the sum of remote region averages $R1$ and $R2$.

Voxel coordinates

Finally, as in (Lempitsky et al., 2009), we can insert absolute voxel coordinates: $\theta^X_C(x) = x_x$, $\theta^Y_C(x) = x_y$, $\theta^Z_C(x) = x_z$ into the feature pool. However, not until these

coordinates have a strong anatomical meaning. This happens later, in the second forest layer when we reorient the images into the standard pose.

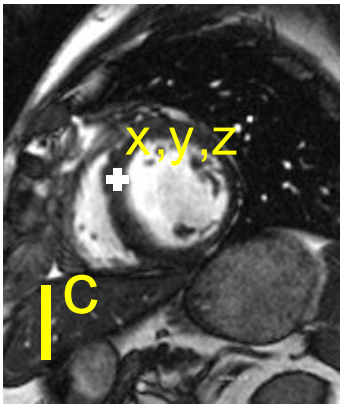


Figure 3.4: Components x,y,z of voxel coordinates as features (Lempitsky et al., 2009).

Spatio-temporal context rich features

The domain of the moving heart can be coarsely extracted by just thresholding the temporal intensity ranges of the image (Section 5.4.2.1). We propose to exploit the wealth of information in time and extend the previous context-rich features into the temporal domain by comparing the "current" 3D frame I^c and another frame offset from c by $\pm o$. The temporal context-rich features can be defined as $\theta^{TCR1}_{I^c} = \theta^{CR}_{I^c,I^{c+o}}(x)$ and $\theta^{TCR1}_{I^c} = \theta^{CR}_{I^c,I^{c-o}}(x)$.

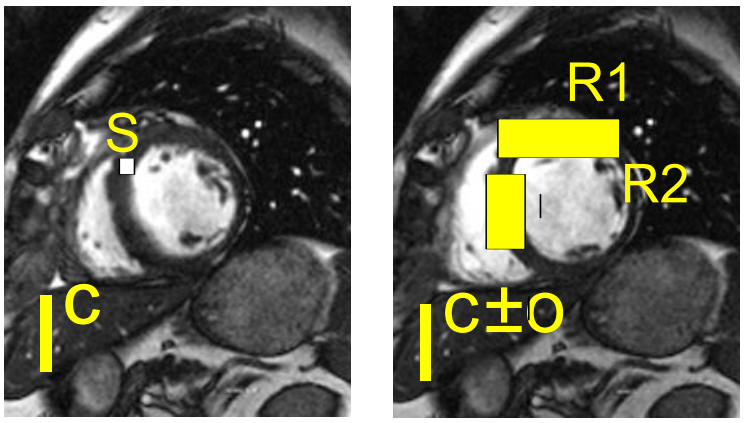


Figure 3.5: Spatio-temporal context rich features with one of the offset frames as the source image and the other as remote.

Similarly, we measure the differences between the symmetrically offset frames contained in the triplet as $\theta^{TCR2}_{I^c}(x) = \theta^{CR}_{I^{c+o},I^{c-o}}(x)$ and $\theta^{TCR2}_{I^c}(x) = \theta^{CR}_{I^{c-o},I^{c+o}}(x)$. These spatio-temporal features can be seen as an approximation to temporal differ-

entiation around the centre frame. Note that we use both $+o$ and $-o$ to keep some symmetry of the remote region distribution.

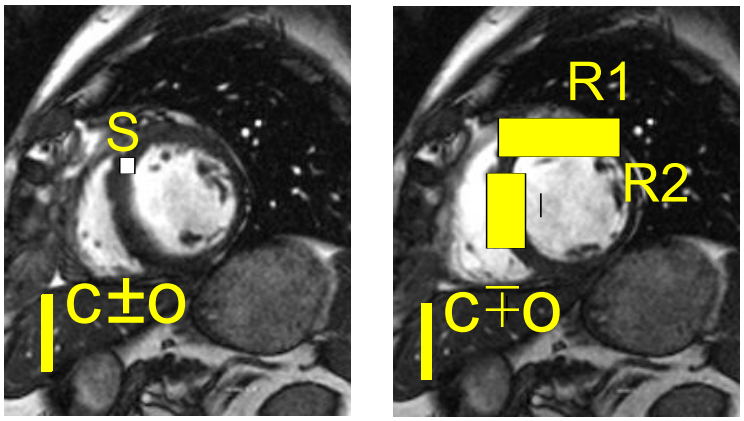


Figure 3.6: Spatio-temporal context rich features with one of the offset frames as the source image and the other as remote.

Data preprocessing

To use fast evaluation of previously defined features based on integral images (Viola and Jones, 2004; Crow, 1984), it is necessary to have consistent spacing. Therefore, all the volumes were resampled to one of the most common voxel size in the dataset $(1.56, 1.56, 7.42)$ milimetres and temporal sequence length (20 frames).

Intensity ranges of the images were all linearly rescaled to a fixed range. Similarly to Nyúl and Udupa (1999), we clip intensities beyond the 99.8 percentile as they usually do not convey much useful information.

3.1.6 First layer: Decision forests for image intensity standardisation and pose normalisation

We train the first layer of the forests using the above mentioned training subset selection strategy. The training is done directly using the images after intensity rescaling, *i.e.*, images are brought into the same intensity range but have their original poses. Although short axis scans are often acquired close to a position where the ventricular ring is centred, slice orientation is chosen manually during the acquisition, and precise alignment cannot be guaranteed. Therefore, we skip the usage of absolute voxel coordinate features at this step.

Several authors (*e.g.* Shi et al. (2011); Pavani et al. (2010)) have proposed to use Haar-like features to detect the heart and crop the cardiac region of interest. Images can be then registered using only information in the cropped volumes. This reduces the influence of background structures on the registration and improves its success rate. However, an extraction of the cropped region will not be necessary to perform a robust registration in our case. We train the first layer of the forests

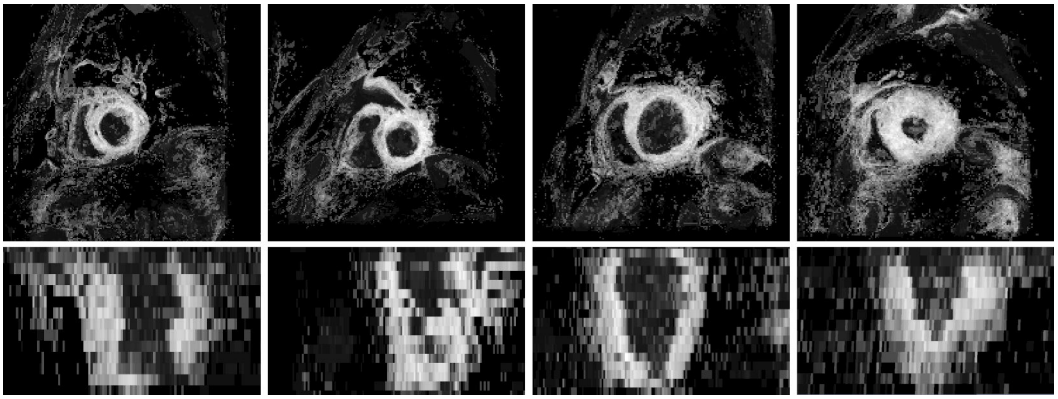


Figure 3.7: Short (top) and long (bottom) axis views of the posterior probabilities after the first layer. Brighter value means higher probability.

on a rather general scenario, to end up with at least a very rough classification performance (see Fig. 3.7). The performance of this first layer may not be sufficient for image segmentation. We show in the next two sections, how these rough posterior probability maps (of a tissue belonging to the myocardium) are already good enough for ventricle detection, intensity standardisation and alignment of the ventricles onto a reference pose without any prior knowledge of the data apart from the ground-truth.

Intensity standardisation

MR intensity value differences of the same tissue are significant not only between scanners and acquisition protocols (Shah et al., 2010) but also for the same follow-up patients (Nyúl and Udupa, 1999). Therefore, good intensity standardisation is crucial for any intensity based segmentation algorithm. The variance in median intensities of the myocardia between different cases in the training set is quite large (see Figure 3.8). There is no unique mode and the distribution is fairly spread across the whole intensity range $(0, 65535)$. Median myocardial intensities span range $(1954, 36430)$, with standard deviation of 5956 and inter-quantile range 7663. This is a serious problem for any intensity-based segmentation method.

Many of the intensity standardisation algorithms (Bergeest and Jäger, 2008) used today are based on Nyúl and Udupa (1999); Nyúl et al. (2000) and the observation that most of the MR intensity histograms are often either unimodal or bimodal for which the second mode corresponds to the foreground object. The alignment of histogram-based landmarks (*e.g.* modes, percentiles or statistics of homogeneous connected regions) is done by rescaling image intensities with a piecewise linear mapping. These methods do work well for brain images where the white matter is clearly the most dominant tissue. In CMR, the largest homogeneous regions belong most of the time to the lungs, the liver or the blood pool rather than the myocardium.

However, from the rough image first layer classification we already obtain some

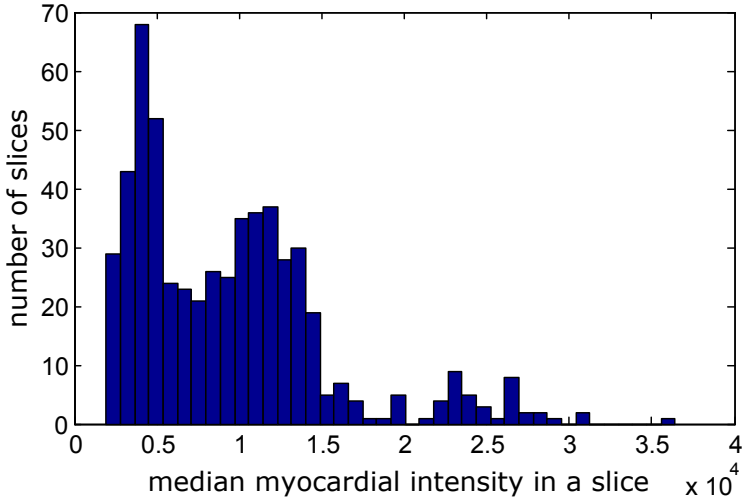


Figure 3.8: Histogram of median intensities of the myocardium across the dataset.

information about the strength of the belief in the foreground and background object. We propose to remap the source image intensities by a piece-wise linear function, such that the weighted median intensities (as median is more robust to outliers than the mean) M^c_{source} of the images are transformed to a reference intensity value M_{ref}. The weighted median is defined as follows:

$$M^c_{source} = \arg\min_{\mu} \sum_{x \in I^c} w(x)|I^c(x) - \mu| \tag{3.3}$$

Where x is the voxel iterator and $w(x)$ are the weights (first layer posterior probabilities of voxel x belonging to the myocardium). We avoid sorting all volume intensities by approximating the weighted median in one iteration over the volume voxels by using the weighted version of the P^2 algorithm (Jain and Chlamtac, 1985; Egloff, 2005). This algorithm dynamically approximates the cumulative probability density function with a piece-wise quadratic polynomial by adjusting positions of just five markers as the weighted samples are streamed in.

Pose normalisation

In the approach of Lempitsky et al. (2009), absolute voxel coordinates are used as features directly. The use of the image coordinates cannot be justified without first aligning the images onto a reference pose. In addition, the features we use for classification are certainly not rotation invariant. Therefore, registering the volumes to roughly the same orientation and position not only helps the classification but also allows us to use those voxel coordinates as features.

In general, the interpatient cardiac registration is a difficult problem due to the high variability of the thoracic cage. One way to do this is by first training a commodity computer vision detector (Viola and Jones, 2004) trained for the hearts (Shi et al., 2011) and then apply a locally affine registration method within

this detected region followed by a non-rigid alignment. A robust learning based linear inter-patient organ registration was proposed by Konukoglu et al. (2011). Here, each organ is represented with a smooth probability map fit to the bounding boxes obtained as a result of a regression forest. Then, registration of these probability maps is performed. This representation of probabilities is however rather limiting as the boxes are axis-aligned and disregard the orientation that we would like to correct for.

We propose to rigidly align the myocardium enhanced first layer posterior probability maps instead. For this step we use a fast and robust rigid block matching registration technique (Ourselin et al., 2000). The reference we used is chosen semi-randomly from images where the apex was at least partially present and closed. An even more automatic way to construct the reference such as an algorithm similar to Hoogendoorn et al. (2010) or a generative technique like described in Iglesias et al. (2011) could be used.

To reduce the computational time, only probability maps of frames from the middle of the sequence are used to estimate the intensity and pose transformations. The same transformations are then applied for all the frames and ground-truths in the sequence which will be needed to train the second layer.

3.1.7 Second layer: Learning to segment with the shape

Retraining with pose and intensity-normalised images

Once the images are registered to a reference volume and intensities are standardised, we retrain a new classification forest. The voxel coordinates start to encode spatial relationships with respect to the reference coordinate frame and the coordinate features can be now included in training of the second decision forest layer. Moreover, if the intensity standardisation step succeeds, the intensities have more tissue specific meaning (at least for the myocardium).

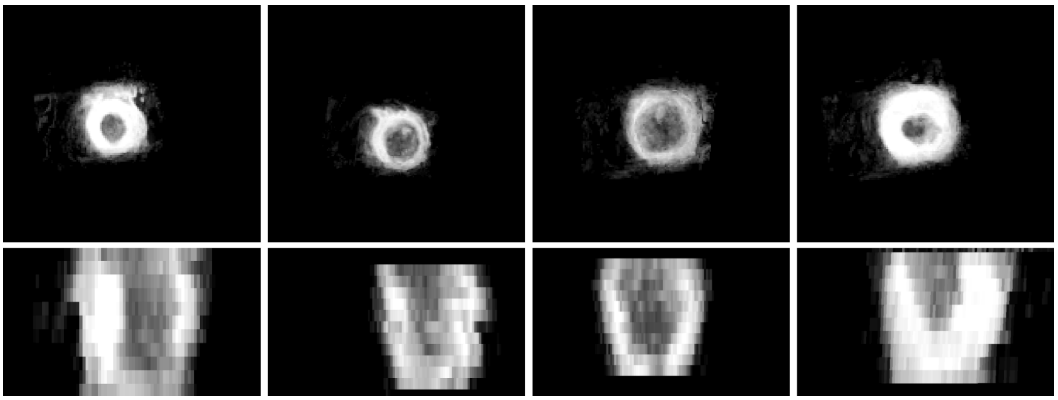


Figure 3.9: Mid-cavity short (top) and artificial long (bottom) axis views on the posterior probabilities after the second layer. The predictions are clearly more confident and most of the remote misclassified voxels are now properly removed.

Thanks to the incorporation of coordinate-based features, the trees can completely automatically learn their own latent representation of the possible set of shapes, regularize the classification, and help to remove objects far away from the ventricles (see Fig. 3.9). This step strongly relies on the success of the previous registration step. Currently, only one reference image is used, which is chosen randomly. Registration to multiple targets could improve robustness and alleviate this problem.

Transforming the volumes back

After the classification is done in the reference space, the posterior probability maps can be transformed back to the original reference frame and resampled accordingly. This shows the advantage of a soft classification technique like the classification forests. The final binary segmentation masks (see Fig. 3.10) are obtained by thresholding the floating point posterior probability maps after the transformation, thus avoiding the possible interpolation artefacts.

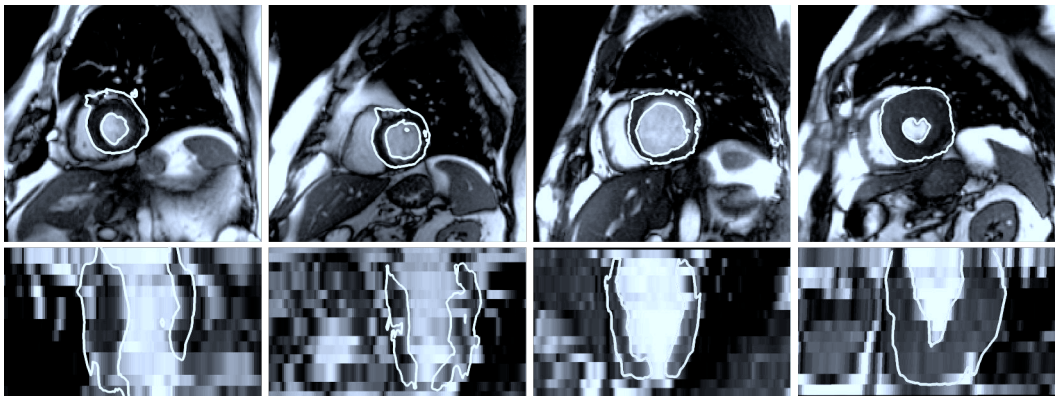


Figure 3.10: Short (top) and long (bottom) axis views on the second layer segmentation results (isocontour of the probability maps at 0.5).

3.1.8 Validation

The forest parameters for the first layer were fixed as follows: 20 trees with depth 20 each. For tree, 15 triplets of frames were randomly selected from different volumes of the training set containing 96 volumes in total. For the second layer 27 trees were trained, each with maximal depth 20. The threshold value to obtain the final segmentation was chosen empirically as 0.5.

Common training and validation datasets allowed fair comparison of our method with others (Suinesiaputra et al., 2014b). Most of the methods in the comparison benchmark were based on manual segmentation in one frame and propagation of the segmentation contours with registration (AU (Li et al., 2010), AO (Fahmy et al., 2012), DS). The only fully automatic techniques in this benchmark were SCR (Jolly et al., 2012) and ours (INR).

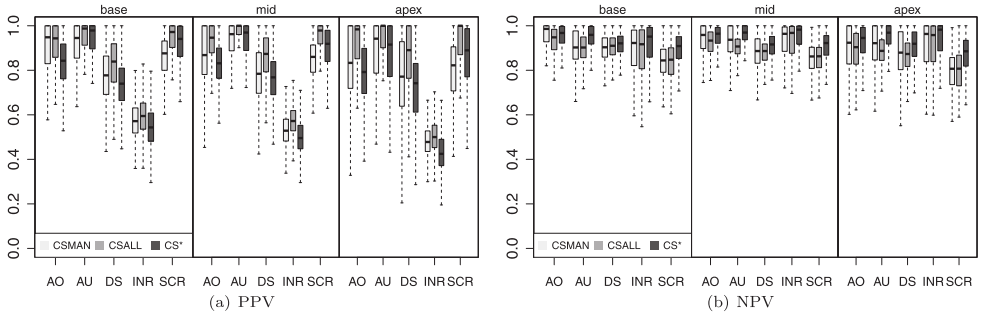
	sensitivity	specificity	PPV	NPV	Jaccard
CSMAN	0.88 (0.17)	0.53 (0.18)	0.54 (0.09)	0.91 (0.10)	0.49 (0.10)
CSALL	0.86 (0.20)	0.60 (0.16)	0.56 (0.11)	0.90 (0.11)	0.51 (0.13)
CS*	0.89 (0.17)	0.56 (0.15)	0.50 (0.10)	0.93 (0.09)	0.43 (0.10)

Table 3.1: Performance of our algorithm as compared to CSMAN, CSALL and CS* consensuses by Suinesiaputra et al. (2014b). All values[1] are expressed as "mean (standard deviation)"

From these methods, consensus segmentations were established using the STAPLE algorithm (Warfield et al., 2004) to which the individual segmentation algorithms could be compared. Three different consensuses were computed to which our method was compared (**CSMAN** - using only manual segmentations methods: AU, AO, DS; **CSALL** - using all five methods: AU, AO, DS, SCR, INR; **CS*** - using all methods but ours: AU, AO, DS, SCR).

3.1.9 Results and discussion

The following results were obtained after evaluating our segmentations on 95 previously unseen test volumes. See Table 3.1.



(a) PPV (b) NPV

Figure 3.11: Comparison of PPV and NPV values from each rater between CSMAN and CSALL consensus (Suinesiaputra et al., 2014b). Only SCR and our method (INR) were fully automatic.

The SCR algorithm first segments the blood pool as the largest moving component of the image. The myocardium is segmented in polar representation of the image using Dijkstra's shortest path algorithm (Dijkstra, 1959). A clever use of an externally trained landmark detector on the long axis images lets them cut off

[1]Positive predictive value (PPV) is defined as $TP/(TP+FP)$ and Negative predictive value (NPV) as $TN/(TN+FN)$.

Measure	CSMAN	CSALL	CS*
EDV (ml)	-55.94 (40.78)	-73.32 (47.74)	-55.65 (41.03)
ESV (ml)	-38.43 (31.55)	-48.11 (33.21)	-39.76 (27.74)
EF (%)	-1.33 (24.10)	-2.45 (22.42)	1.18 (22.01)
LV mass (g)	124.35 (36.08)	90.23 (32.52)	127.80 (34.32)

Table 3.2: Clinical LV function differences between our segmentations from the CSMAN, CSALL and CS* consensus. All values are expressed in "mean (standard deviation)" (Suinesiaputra et al., 2014b). Our method clearly oversegments the myocardium by misclassifying the papillary muscles and trabeculations as myocardial tissue. This is reflected as large positive bias in LV mass and negative in ESV and EDV. The EF is not affected that much by these biases.

voxels beyond the mitral valve plane. Segmentations of individual frames are then propagated with non-rigid registration, similarly to the other methods.

In the following discussion we use the CSMAN consensus results as the reference since it is based on manually initialised segmentation techniques only. In most of the cases, our algorithm was able to correctly identify the left ventricle myocardium (with mean sensitivity of 0.88, specificity of 0.53 and Jaccard index of 0.49). Sensitivity (also called true positive rate) measures the proportion of myocardial voxels that are properly segmented as such. Specificity (also called true negative rate) is the proportion of background voxels that are properly segmented. Higher sensitivity and lower specificity indicate that our method oversegments. Indeed, the trabeculations and papillary muscles are often misclassified and included into the myocardial mass (contrary to the ground-truth). The effect of this can be also seen on the volumetric and mass measures extracted (Table 3.2).

Our method was clearly not the best performer between automatic segmentation methods for left ventricles in the benchmark. On the other hand, the layered forest algorithm did not need any explicitly defined segmentation rules and problem specific assumptions (*e.g.* circularity of the myocardium for polar transformation). It also did not use extra annotation into the training set (*e.g.* manual segmentation of one frame in the sequence) nor used an external dataset to train a robust landmark detector.

It should be also noted that our classification is run independently for each voxel. No smoothness, connectivity nor temporal consistency constraints are enforced to demonstrate the performance of the pure machine learning approach. Therefore, isolated segmentation islets or holes in the resulting binary segmentation can occur as a result of misclassification. Thanks to the coordinate features, most of the voxels far from the myocardium are usually discarded and also the solution becomes more regular as a result of the latent cardiac shape model built by the forests. In the soft classification, the holes are represented as a drop in the segmentation confidence

but rarely fall to zero. This information could be considered in a subsequent post-processing step (*e.g.* the graph cut algorithm (Malik, 2000)), to further improve the final segmentation.

3.1.10 Volumetric measure calculation

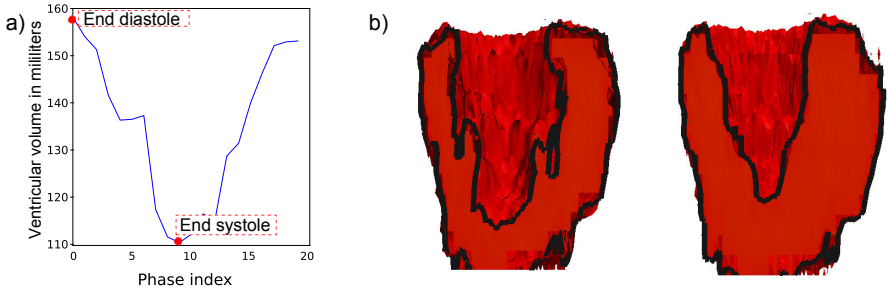


Figure 3.12: (a) Automatically calculated volume curve from patient DET0026701 during a single cardiac cycle with detected end systole (ES) and end diastole (ED) frames at the volume maximum and minimum respectively. (b) Long axis crosssection through the binarised segmentations at ED and ES.

Finally, using a curvature-based iterative hole filling algorithm (Krishnan et al., 2009) we fill up the binarised myocardial segmentation, in order to automatically calculate volumetric and mass measurements (Table 3.2) and to detect the main cardiac phases as the volume curve extrema (see Fig. 3.12). End diastolic volume (EDV) is the maximal and end systolic volume (ESV) the minimal volume of voxels labelled as the cavity. Ejection fraction (EF) is computed as (EDV − ESV)/EDV. The LV mass is obtained from myocardial volumes at ED assuming constant tissue density of 1.05g/ml.

3.1.11 Conclusions

In this section, we presented a fully automatic machine-learning-based algorithm for left ventricle segmentation from 4D CMR with no explicit definition of specific segmentation rules, model creation, user interaction, nor post-processing. The algorithm learnt to automatically select the most discriminative features for the task using the ground-truth only. The only assumptions we made is that the motion of the object to be segmented is periodic for the construction of frame triplets and that the tissue intensity mapping between two different cases can be roughly approximated by a piece-wise linear function. We also introduced a machine-learning-based intensity standardisation method that allows to do tissue specific remapping of intensities and obtain a more CT-like behaviour.

Our method is currently not the most accurate method for myocardial CMR segmentation (see Fig. 3.11) (Suinesiaputra et al., 2014b). On the other hand, it is rather flexible and only very little prior knowledge was hardcoded into the algorithm. The recommendations of the Board of Trustees Taskforce of the Society for

Cardiac Magnetic Resonance (SCMR) suggest excluding the trabeculations and papilary muscles from the ventricular cavity and including them into the myocardial mass (Schulz-Menger et al., 2013). This is a challenging problem for the shape and registration-based methods. In addition to the papillary muscles and the trabeculations, there are other important cardiac fine-scale anatomical structures (Lamata et al., 2014) and findings of interest. These also include small blood vessels, or hyperintense scar regions in late enhanced MRI or pericardial effusion. In these cases, voxel-wise semantic segmentation techniques are likely a better solution. Once enough examples of voxels belonging to these small structures are annotated in the ground-truth, our method should be able to figure out how to discriminate them from other tissue, and segment these without much modification. This method could also be easily adapted to the segmentation of long axis views, usually 2D or 2D+t sequences. In this case we could just replace the spatio-temporal boxes with spatio-temporal rectangles.

We will now show that our algorithm is also amenable to segment another important but large cardiac structure — the left atrium (LA).

3.2 Left atrium segmentation

The LA plays an important role in facilitating uninterrupted circulation of oxygenated blood from pulmonary veins to the left ventricle and in cardiac electrophysiology. To quantify its function, simulate electrical wave propagation and determine the best location for ablation therapy, it is important to be able to first accurately segment the atrial contours. A common approach to segment the left atrium from 3D images is to use statistical shape models (Kutra et al., 2012; Ecabert et al., 2008). A levelset-based method with a heuristic region split and merge strategy was proposed by Karim et al. (2007). Finally, label fusion techniques (Depa et al., 2010) seem to yield accurate atrial segmentations but require to non-rigidly register the image to be segmented to every training image. All of these methods are specifically hand-crafted for atrial segmentation and thus treat the training set in a particular way or need a set of non-rigid registrations which can be computationally expensive.

Similarly to our method for LV segmentation, we use a fully automated voxel-wise segmentation technique based on classification forests. This time, the problem is formulated as a binary classification between atrial and background voxels. The advantage of these approaches is that very few assumptions are necessary and it is possible to learn how to segment directly from the image - label map pairs. For this method we do not use a robust registration method, build a statistical model, nor explicitly define the classification problem. We only require a larger number of training images with the atria carefully delineated and good blood pool contrast in the images.

3.2.1 Dataset

Compared to the LV segmentation, the dataset is much smaller in this case. The STACOM 2013 LA segmentation challenge dataset (Tobon-Gomez et al., 2013b) contains 30 CT and 30 MR images. Each of these sets was divided into a training set (10 3D volumes with voxel-wise LA segmentation maps) and a validation set (20 3D volumes with no delineation provided used for blind evaluation).

3.2.2 Preprocessing

The only preprocessing step is linear intensity rescaling of the 0 and 98.5 percentile image range as done in the first layer of the LV segmentation. This was chosen to cut off high intensity variation due to noise and imaging artefacts, similarly to Nyúl and Udupa (1999).

3.2.3 Training forests with boundary voxels

Training of decision forests is in general relatively fast as each tree can be learnt individually in parallel. To further cut down from the training time, and more importantly to better balance the background/atrium voxel proportion and to better learn how to classify voxels on the boundaries, we train only on some of the voxels from the annotated training set. We include all atrial voxels in the training set as positive examples. However, we subsample the negative example voxels. These are sampled only on a sparse regular grid. We also add all voxels in the immediate atrial neighbourhood (approximately 15 pixels thick, obtained by morphologically dilating the atrial mask).

3.2.4 Segmentation phase

During the segmentation phase, each voxel of the test image is passed through the forest to reach a set of leaves. The average class distributions of all reached leaves then represent the posterior probabilities of the voxel belonging to either the atrium or the background given its appearance in the feature space. This means that we obtain an atrial probability map for the whole volume. To obtain binary masks required for evaluation in the benchmark study of Tobon-Gomez et al. (2015), we simply threshold the atrial probability maps with a fixed threshold value. Afterwards, we apply simple morphological hole-filling and extract the largest connected component which we deliver as the final binary atrial segmentation.

3.2.5 Additional channels for left atrial segmentation

To describe the appearance of each voxel and discriminate between the atrium and background we generate a random feature pool operating on the 3D images from two feature families (local and context-rich) previously shown in Section 3.1.5 and in Geremia et al. (2011), but applied on three different image channels. These two feature families capture both local and remote information about the tested voxel.

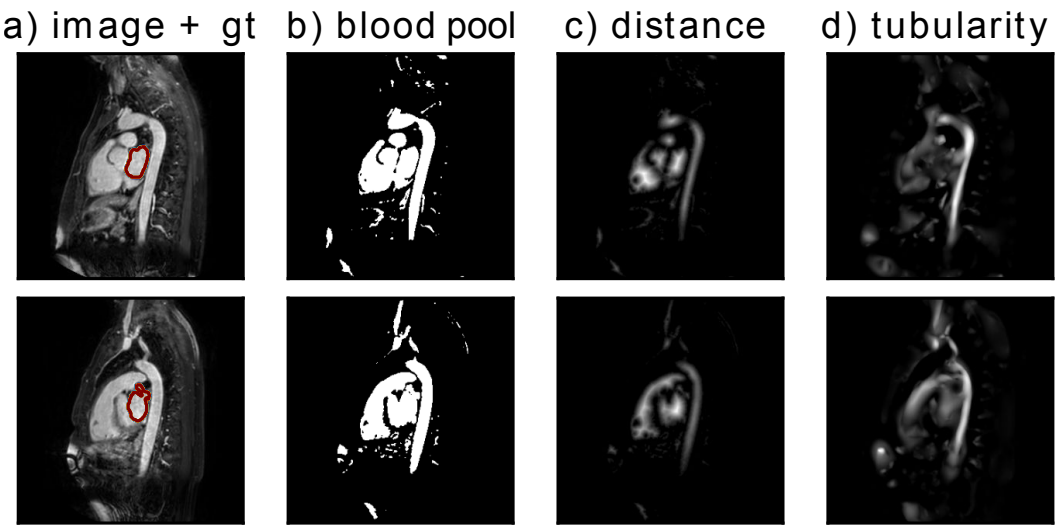


Figure 3.13: Image channels extracted from two example images. (a) Source image overlaid with the ground-truth segmentation. (b) Blood segmented with sequentially applied Otsu thresholding.(c) Distance to the blood contours.(d) Tubularity enhancing vascular structures (*e.g.* the strong signal in the aorta)

3.2.5.1 MR image intensity

Voxel intensity in combination with context rich features give wealth of information about its position. For example, regions of the bright atrium are located close to the darker spine or lungs, and next to other bright cavities. It is however much more difficult to discriminate between voxels on the border between the atrium and the ventricle as there is no clear intensity change (apart from occasional faint signal drop from the mitral valve). Therefore, we extract the local and the context rich features not only on image intensity (Fig. 3.13a), but also add two other channels (see Fig. 3.13c and Fig. 3.13d).

3.2.5.2 Distance to blood pool contours

The atrium is always contained within the bright blood pool in the image. Thanks to the high contrast between the blood pool and the rest of the image, all blood can be simply extracted by sequentially applying the thresholding algorithm of Otsu (1979). The first round divides image voxels into the black air and the brighter part of the thorax. The second round splits the brighter part of the thorax into the very bright blood and the rest (Fig. 3.13b).

We can observe that the atrium is mostly separated at blood pool narrowings (such as at the mitral plane or atrial septum). These can be located by measuring the distance to blood pool contours as distance minima (Fig. 3.13c). Local maxima are, on the other hand, located near centres of blobs in the image such as the atrium. Therefore, similarly to Karim et al. (2007), we exploit these properties and compute the Euclidean distance to the blood pool surface for each voxel in the image (voxels

out of the blood pool are assigned zero distance). Instead of manually defining region splitting and merging criteria we let the forest pick the most discriminative decisions from the above mentioned feature families.

3.2.5.3 Tubularity features

To further help in distinguishing the atrium from the other bright structures such as the aorta, we calculate the vesselness information for each voxel. This also adds context based on enhanced arteries present (*e.g.* the atrium is near the aorta — a large tubular structure). We use a multi-scale vesselness filter of Sato et al. (1997) enhancing all tubular objects ranging from 5 to 15 millimetres in diameter (see Fig. 3.13).

3.2.6 Validation

We trained a classification forest with previously described features on these three image channels. We chose the best parameters by running cross-validation on the training set. As the size of the training set is quite limited, we used a leave-one-out approach, *i.e.*, we trained on 9 images and tested on the remaining one (Fig. 3.14). The best settings were then applied to the validation data.

3.2.6.1 Effect of the parameters

Number of trees (5): More trees result in increased accuracy and slightly smoother segmentations but also increase training and classification time.

Number of tested features (200) and thresholds (20) at each node: Testing fewer features results in more randomness in the forest but also less efficient splits. On the other hand, higher numbers decrease generalisation strength as the trees look more alike.

Maximal tree depth (20): Deeper splits can better capture the structure, but can lead to overfitting.

Minimal number of points at leaves (8): Too small number of samples in a leaf could result in noisy segmentations. For example, a leaf containing a single incorrectly labelled training voxel could significantly influence the result.

Neighbourhood in which context rich features are sampled ($70 \times 70 \times 70$): Larger neighbourhoods can capture more context, however result in more frequent out of image bounds evaluation of the context rich features. We clip the boxes to the image extent (they evaluate to zero when completely outside of the images).

Binarising threshold (0.9): We use a higher threshold to keep only the most confident voxels and to reduce the effect of oversegmentation.

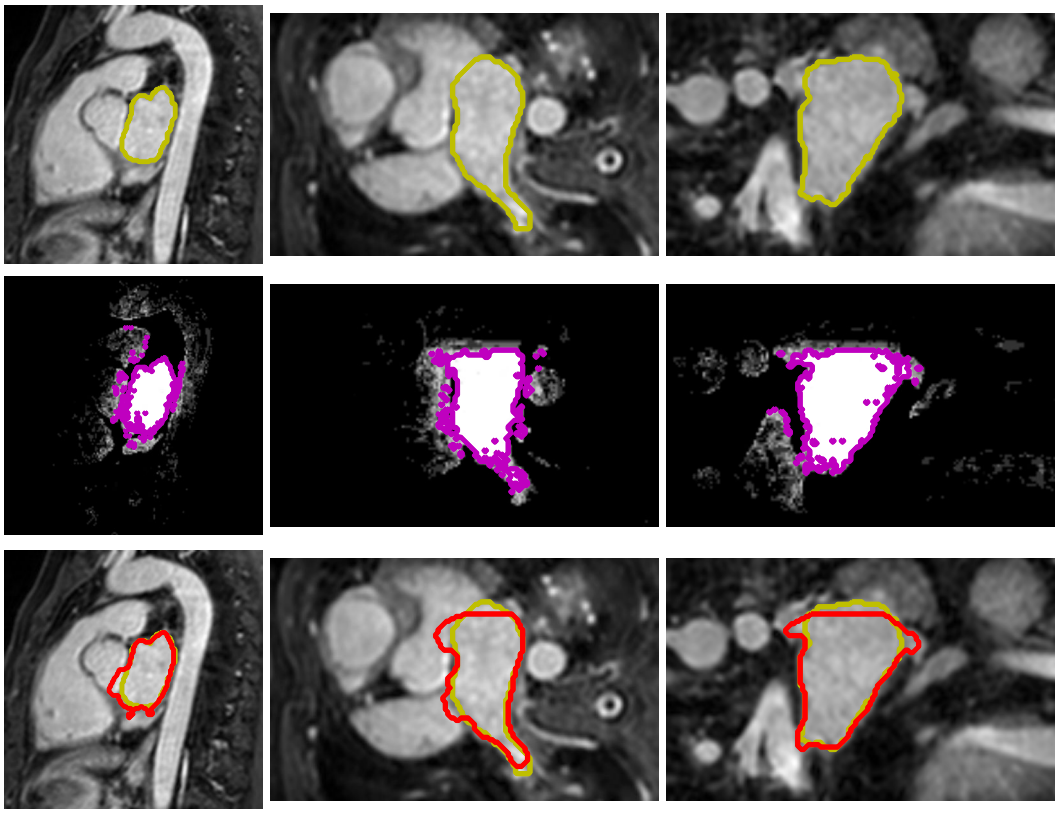


Figure 3.14: Coronal, sagittal and axial views of the segmentation results on one of the test cases during the leave-one-out cross-validation. Top row: source image with ground-truth, middle row: atrial probability map with contour of the probability map at 0.6 (brighter values means more confidence in the segmentation), bottom: source image with overlaid ground-truth (green) and final segmentation after hole filling (red).

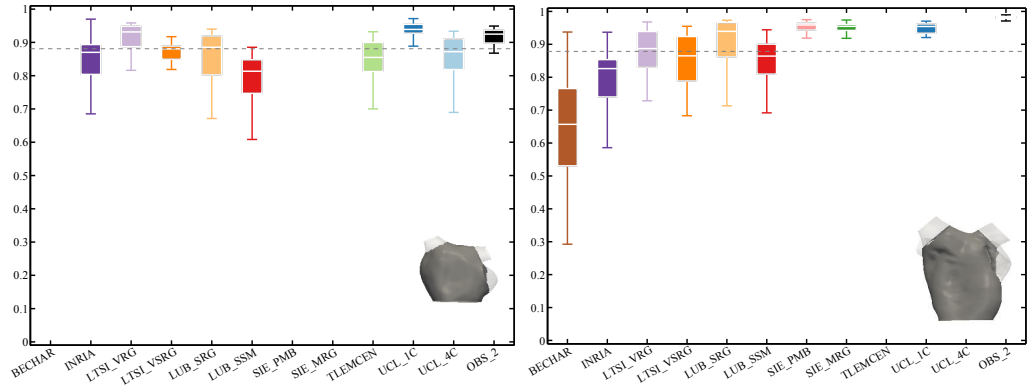
3.2.7 Results and discussion

After the selection of parameters above via cross-validation we obtained a dice coefficient 0.63 ± 0.14. This small forest of 5 trees took on average 2 hours to train and just around a minute to fully automatically segment a single atrial MR image on a 12 core Intel Xeon 3.3GHz CPU.

This algorithm performs reasonably well to extract the main part of the atrial volume, its shortcomings are mainly in the segmentation of the pulmonary veins which are often missed or misclassified. The other main drawbacks of our method is that that the segmentation contours do not necessarily adapt to the cavity contours in the images (see Fig. 3.14).

3.2.7.1 Benchmark results

Tobon-Gomez et al. (2015) developed an elegant benchmarking framework for comparison of the atrial segmentation for various parts of the LA. We submitted our segmentation results from the MR images and ran almost the identical pipeline to segment also the CT images. The comparison of all submitted methods on the whole validation set for the LA body can be seen in Fig. 3.15.



(a) MRI (b) CT

Figure 3.15: Boxplot of Dice coefficients for all methods in the benchmark (Tobon-Gomez et al., 2015). The dotted line represents the mean of medians of the benchmarked methods. Our method is listed as INRIA, OBS_2 is second rater.

3.2.8 Conclusions

We used our machine-learning-based image segmentation technique and extended it with a set of image features to better distinguish vascular structures from the rest of the blood pool. We then learnt to directly predict voxel labels (atrium / background) from the images without hand-tuning the segmentation pipeline. Our only assumptions were the strong blood pool contrast and the presence of tubular structures in the images.

Our segmentation algorithm is quite fast but lacks for correct classification of boundary voxels and small structures, *i.e.*, the pulmonary veins that were also marked as part of the atrium in the ground-truth. Our segmentations serve as an excellent atrium detector and initialisation for a refinement step in order to produce accurate segmentations. Such refinement will be needed for electro-physiological studies.

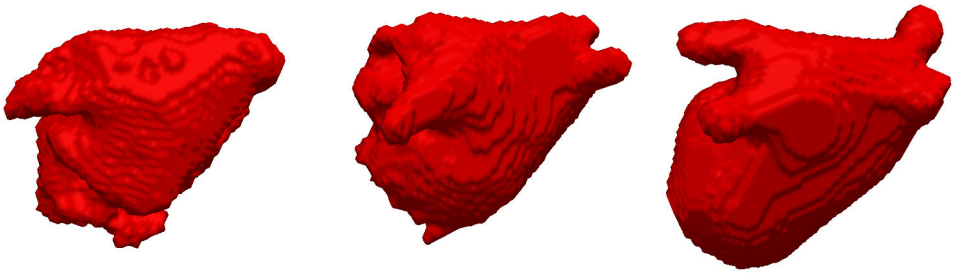


Figure 3.16: Qualitative display of segmentated atrial meshes from the validation dataset.

Due to the fact that the training set consists of only 10 images it does not cover many of the atrial shape variations present in the validation images. To capture some of this variability and avoid image registration to an atlas in the classification phase it might be possible to augment the training dataset, *e.g.*, by transforming the images with a set of smooth deformations.

3.3 Conclusions and perspectives

In this chapter, we have shown that segmentation can be learnt directly from voxel-wise labels using classification forests. We have also demonstrated how segmentation can be learnt for spatio-temporal images. Although in this work we apply this method only to MR and CT images, it is quite straightforward to apply it to other modalities and other structures of interest. There is very little prior information included and the method can automatically pick relevant features from the given set and improve with more input data.

Nevertheless, when less examples are given, resorting to extra (more hand-crafted) feature channels helps to improve the segmentations and allows to learn how to segment from fewer training images. Yet, a couple of improvements to the method are necessary to allow us to reliably index and query cardiac images using volumetric measurements.

3.3.1 Perspectives

Since this work, further studies on segmentation of cardiac structures and other organs using semantic voxel-wise machine learning classification techniques have

appeared. These include different sets of features or image channels to describe the voxels (Mahapatra, 2014; Wang et al., 2014), to use more than two layers (Keraudren et al., 2014), to more intelligently select subsets example training images for each tree (Lombaert et al., 2014), or to exploit also unlabelled data (Appendix A). Many ideas in these approaches are orthogonal to ours and could further help to improve the quality of our segmentations. In Lombaert et al. (2014) alone, the Dice coefficient increased from 66.6% up to 79% by smarter selection of training image subsets for each tree.

The use of classification forests for segmentation greatly simplifies the problem of learning how to segment directly from voxel-wise ground-truth without hardcoding the segmentation rules. Nevertheless, the algorithm still requires to define the feature families and design image channels that enhance relevant regions, and to improve the smoothness and contour matching. In addition, the tree structure is optimised and fixed in the training.

For growing databases, when new training examples are continuously added into the training set, the forest can be retrained from scratch or the leaf distributions can be updated accordingly. The more recent CNN-based segmentation approaches (*e.g.* Long et al. (2015)), automatically learn optimal image channels to solve the segmentation and can update the network parameters with SGD when new data become available. Finally, Kontschieder et al. (2015) showed a hybrid approach where the classification forests are made differentiable and trained within a CNN using backpropagation.

Chapter 4

Crowdsourcing semantic cardiac attributes

Contents

Based on our work in preparation for submission (Margeta et al., 2015b).

Chapter overview

In Chapter 2, we have shown how to automatically predict the CMR image view information using CNNs. We can easily retrieve, for example, all short axis image slices from the database. Automatic cardiac image segmentation methods (such as the ones presented in Chapter 3) allow us to index and query the cardiac databases using simple geometric measurements, (*e.g.* cavity volumes, myocardial mass, or wall thickness). These two tools are important steps towards automated content-based image retrieval of cardiac images.

Some images, however, cannot be directly described and indexed on such simple geometric measurements alone. Consider a slightly more complex query: "retrieve all short axis images of akinetic hearts with significant wall thinning, and filter out low quality images with artefacts at the same time".

We aim to achieve computerised description of cardiac images with a set of semantic *shape*, *motion* and *appearance* attributes. In this chapter, we focus on images of post-myocardial infarction hearts with mild to moderate left ventricular dysfunction. We develop a tool that will allow filtering of cardiac databases with such a set of meaningful attributes. We learn to order images based on attributes learnt from pairwise image comparison ground-truth. The ground-truth consists of two images and an indicator for which image of the two has lower or higher presence of the attribute. We designed a web interface allowing to collect such ground-truth via crowdsourcing.

4.1 Describing cardiac images

Finding a way to interactively query and explore large cardiac imaging collections, and to find relevant cases linked to Electronic health records (EHRs) with treatment history and outcomes, could enrich the process of diagnosis and therapy planning. Such resource could also serve as a rich learning tool and to generate new insights for disease discovery.

One possibility is to computationally describe the hearts in these database with some semantic concepts. As semantic we mean a description that is understandable by a human, *e.g.*, large left ventricular hypertrophy or small muscle contractility, in contrast to an arbitrary feature vector used by the machine, *e.g.*, a histogram of filter responses. Similarly to how a graphic designer can interactively find appropriate fonts described with a set of traits such as dramatic, playful, legible or thin (O' Donovan et al., 2014), a cardiologist could use a set of traits describing human hearts for exploration of an unknown dataset. Of course, most hearts in the databases are not labelled with these traits. We build a method to ultimately describe with them any heart.

4.1.1 Non-semantic description of the hearts

Some of the most successful examples of non-semantic description of medical images, in particular for Content based image retrieval (CBIR), are based on texture of the images. André et al. (2011b) used dense image gradient-based SIFT features (Lowe, 2004) combined with the Bag of words (BOW) approach (Sivic and Zisserman, 2003) to create a computerised representation of confocal endomicroscopy images. The BOW approach is widely popular. Foncubierta-Rodríguez et al. (2012) used the BOW representation to describe pulmonary images from wavelet and Riesz transform responses. The similarity is then derived from a similarity measure between histograms of these visual words.

These systems work particularly well for images where the *semantic gap* (the difference between the information captured with the extracted features and the expert similarity) is quite small. Except for some measures simple to obtain from image segmentation (such as cardiac volumes, muscle masses or myocardial wall thickness) the semantic gap in cardiac imaging can be a real challenge.

In cardiac imaging, Glatard et al. (2004) extracted texture-based features to pre-segment the images and to find similarly segmented cardiac slices. Eslami et al. (2012) used a binary cardiac segmentation overlap measure to query a database of already segmented images for the closest match and to guide their segmentation algorithm.

Numerous feature representations have been used to characterise cardiac shapes (Bernardis et al., 2012; Zhang et al., 2014b; Le Folgoc et al., 2013; Medrano-Gracia et al., 2014), and cardiac motion (Duchateau et al., 2011; Mantilla et al., 2013; McLeod et al., 2015), in order to discriminate between normal and abnormal cases. An interesting way to compare cardiac shape differences due to the remodelling was proposed by Medrano-Gracia et al. (2014) and by Zhang et al. (2014b). These atlas based methods look at differences in the principal modes of shape variation (the LV is represented with a personalised finite element model) between two different cardiac cohorts. Zhang et al. (2014b) compared shapes of infarcted hearts against asymptomatic volunteers. The left ventricle shapes at both systole and diastole were then described with up to 20 parameters, the first 20 modes best explaining the shape variation. This approach discovered statistically significant shape changes between the two cohorts that are not simple to capture otherwise.

A search for images with similar or greater value of the 5th mode of variation for example is not very intuitive though. Links of the modes of variation to semantic attributes of the heart (first mode linked to the size of the LV, second mode to the sphericity of the ventricle, and the third to the mitral valve orientation) were found *a posteriori*, by observation.

4.1.2 Semantic attributes

Sphericity has been previously linked to long term survival (Wong et al., 2015). Semantic descriptions like this one are commonly seen in the clinical journals, however, their computational definition is often challenging or ambiguous. Nevertheless, even simpler attributes commonly used to describe the hearts (such as the *hypertrophy*) are often not straightforward to compute automatically. For example, consider assigning the value of hypertrophy to be a number between 0 and 100, where 0 is a normal heart and 100 is an extreme case where the myocardium even in relaxed state leaves little space for the cavity.

André et al. (2011b) proposed an interesting way of retrieving similar sequences from endomicroscopic videos of the colon by first computing a Bag of words (BOW) representation from dense textural descriptions of the image. Then, they mapped the histograms of visual words to a semantically meaningful description (André et al., 2012) helping with interpretation of the images.

Quantifying concepts such as the *degree of remodelling* is quite challenging. Zhang et al. (2014b) used logistic regression to discriminate healthy from infarcted hearts from the modes of shape variation as a feature vector. Instead of the binary classification of the hearts, the learnt weights of the linear projector in the logistic regression could be used to obtain a measure of remodelling.

Learning a linear model that projects the extracted features to semantic attributes is similar to our work in this chapter. The attributes are very different from classses. Classes are discrete (*e.g.* normal, hypoplastic left heart syndrom, Tetralogy of Fallot heart) and they do not even have ordinal relation, attributes are continuous (more or less dilated, more or less spherical, more or less akinetic). In many ways, attributes are related to a regression type of task.

We aim to describe at least some of the aspects of post-myocardial infarction hearts and use the extracted values to filter the dataset. We learn to regress these attributes from pairwise comparisons using machine learning tools used in web ranking. First, let's have a look, which characteristics of the hearts are the most frequently present in the database (Kadish et al., 2009).

4.2 Attributes for failing post-myocardial infarction hearts

Following myocardial infarct, the hearts have several compensatory remodelling mechanisms in play. As a consequence of the myocardial infarction and the loss of viable tissue, myocardial cells are corrupted by the synthesis of interstitial fibrillar collagen. The affected myocardial wall becomes **thinner** (Figure 4.1) and **stiffer** (Mihl et al., 2008). The overall **kineticity**[1] of the heart becomes compromised.

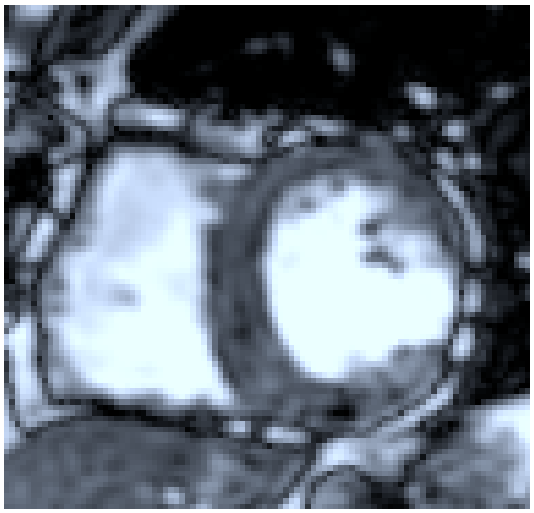


Figure 4.1: Thinning of the lateral wall due to transmural post-infarction necrosis.

To deal with pressure overload secondary to conditions such as aortic stenosis or

[1]The kineticity can be seen as a continuous extension of the discrete clinical labels: hyperkinetic, normokinetic, hypokinetic, akinetic, or dyskinetic

hypertension, the heart muscle usually responds with **hypertrophy** (Figure 4.2). The increase in the muscle mass, leads to temporarily improved contractile force of the myocardial cells. Sustained pressure on the cardiac cavity, however, leads to its **dilation** (Figure 4.3).

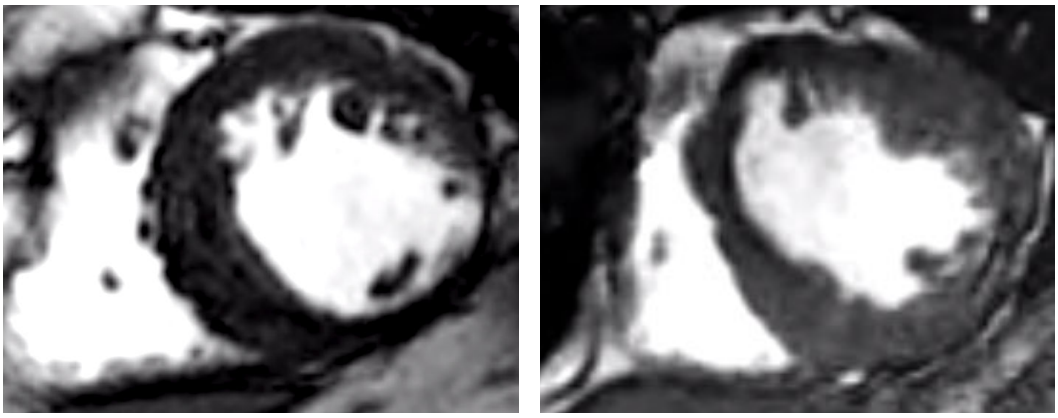


Figure 4.2: Hypertrophy in post-myocardial infarction hearts located opposed to the infarction location. Note also the myocardial wall thinning of the inferolateral wall on the left image.

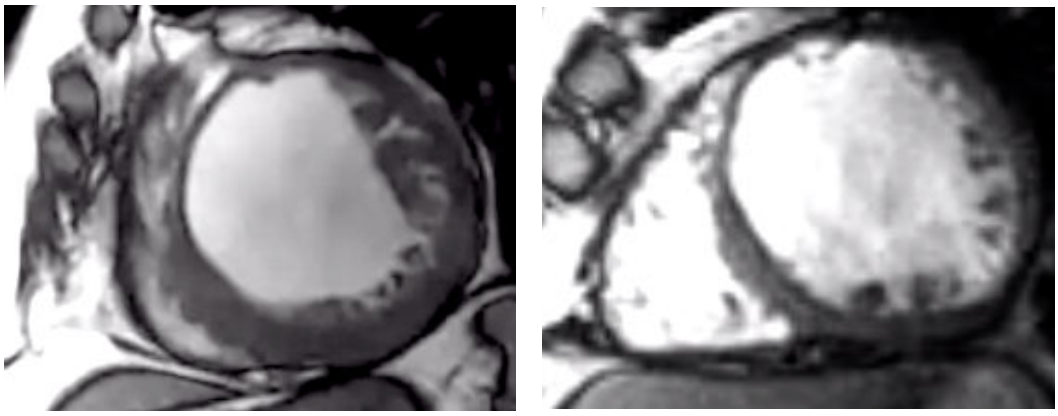


Figure 4.3: Examples of the left ventricular cavity dilation in post-myocardial infarction hearts. The left heart is also significantly trabeculated. Note also the anterior wall thinning in both images.

The changes in the cardiac shape due to the remodelling can also lead to valvular diseases such as **valve insufficiency or regurgitation** which further worsen the cardiac function. A binary tag is sufficient to describe these two findings.

The **pericardial effusion** (see Fig. 4.4) is a relatively common finding (8% of patients) in post-acute myocardial infarction patients (Sagristà-Sauleda et al., 2011). An excessive amount of the liquid can cause cardiac tamponade reducing cardiac function and is often associated with swinging motion of the heart. There are two reasonable options to describe the effusion: a binary flag indicating the

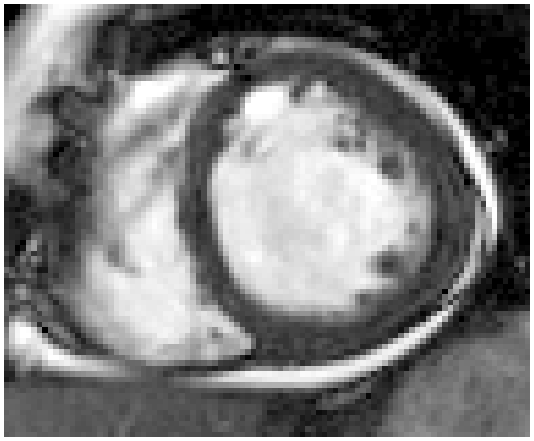


Figure 4.4: Pericardial effusion — fluid built up in the sac around the heart. This is often seen as a bright ring around the heart. The effusion can be easily confused with pericardial fat.

presence or absence of the effusion, or a real value quantifying the amount of the pericardial fluid.

4.2.1 Pairwise comparisons for image annotation

Most of the attributes are continuous. Unless these are derived from well defined measures, expert annotations of such values can differ from centre to centre. It is often much simpler to visualise two cardiac images side by side and let the expert choose which of the two hearts has a more dominant presence of the attribute, *e.g.*, "heart on the left has more hypertrophy than heart on the right" (see Fig. 4.5).

Learning continuous attributes (scores) from pairwise comparisons is a powerful technique to deal with the attribute value assignment from multiple experts (or less experienced annotators, such as students of cardiology). The effect of inter-expert variability will likely be smaller than if assigning an absolute score in isolation.

The consistency of pairwise annotations between different centres and different experts has been recently shown by Burggraaff et al. (2015). The experts viewed pairs of videos of multiple sclerosis patients doing standardised actions. For the video pairs, the experts indicated if the patients were performing equally or one of them had poorer performance. TrueSkill^{TM} (Herbrich et al., 2007; Dangauthier et al., 2007) algorithm, was then used to learn continuous scores (and confidences in them) from the pairwise comparisons for each video. This simple Bayesian model is used daily by millions of people playing Xbox online games for matching players of similar skills. Finding opponents with similar skills from as few matching trials as possible is important for fun and engaging games, and faster and more accurate skill estimation. In the case of pairwise medical sequence annotation, proposing matching videos is important for faster (less burden on the annotators) and more accurate clinical score/attribute predictions.

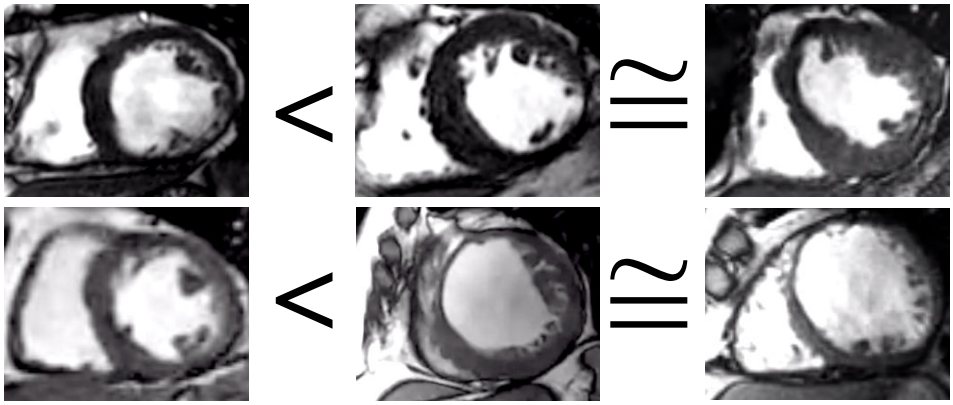


Figure 4.5: Relative comparison of hypertrophy between two hearts. Any two hearts can be assigned a strict inequality or approximate equality. The pairwise comparison is a much simpler task than to assign an absolute degree of hypertrophy, Top row: The amount of hypertrophy correlates mainly with the mass of the muscle. Bottom row: The dilation is mainly determined by the ventricular cavity volume. Together with the dilation one can see thinning of the myocardium due to post-myocardial infarction tissue loss.

In our work, instead of asking for more manual input, we use pairwise comparisons as ground-truth and learn how to predict the attributes directly from image features. This allows us to automatically mimic the pairwise image comparison and ordering by a computer.

4.2.2 Selected attributes for shape, motion and appearance

In this chapter, we chose (based on their presence in the dataset) the following attributes to describe the cardiac *shape*: **hypertrophy**, **dilation** and **thinning**. We describe the cardiac *motion* with the **kineticity** attribute. Moreover, we use the same approach to estimate image ranking based on the *image appearance* — **image quality** — to allow removal of low quality images.

These attributes are not meant to be a disease classification tool. They serve as mid-level concepts to meaningfully describe the cardiac images with interactive CBIR in mind. Image of any heart, regardless of the underlying disease, can be described by all of the attributes simultaneously, *e.g.*, a particular failing heart with a transmural infarction can be described by these attributes as a very dilated and hypokinetic heart with significant thinning. For more severe changes in the anatomy, such as the hypoplastic left heart syndrome hearts, an additional attribute could be devised for the underdeveloped left ventricle.

4.3 Crowdsourcing in medical imaging

Learning to rank (establish order) and to filter the images based on the attributes using supervised machine learning algorithms needs the ground-truth annotations. Crowdsourcing is a powerful way to engage a large number of people to provide data. In machine learning and computer vision it is a well established technique to generate ground-truth for large databases. Only recently, we have seen its adoption in medical imaging (Foncubierta-Rodríguez and Müller, 2012; Maier-Hein et al., 2014). A notable example is the work of André et al. (2012) building a "Smart atlas" as a case lookup software for practitioners and students of colonoscopy. The authors collect binary image attributes and ground-truth for image similarity from pairwise expert annotations ranging from very dissimilar, rather dissimilar, rather similar to very similar. From the discrete annotations, they then estimate the continuous semantic attributes.

4.3.1 A ground-truth collection web application

Collecting the ground-truth is a tedious task, but crowdsourcing helps to speed up the process and reduce the burden on individual annotators. In this work, we design a cloud-hosted web application to collect pairwise cardiac image comparisons through crowdsourcing (See Fig. 4.6). The advantage is that the clinicians, students of medicine, or other cardiac imaging knowledgeable crowd can rapidly label the data on any platform that supports HTML5 and ECMAScript, *i.e.*, most modern web browsers. Thus allowing data annotation even on the go, *e.g.*, on a mobile phone while being stuck in public transport.

We randomly pick an attribute a_m and two images belonging to two cases I_i and I_j. We then present these short axis 3D+t cine stacks for pairwise comparisons. The annotators (or workers in the crowdsourcing terminology) can play the cardiac video sequence, change the slice position, or look into the associated patient information (sex, age, infarction location), if that helps them with the decision. Then, they can decide in which out of the two images is the selected attribute more dominant and add the image pair into the ordered set O_m. However, if the amount of presence/absence of the attribute is approximately equal in both images, the pair is added to the similarity set S_m instead by clicking on the "About the same" button. After the choice is made, another random attribute together with a left and a right pair of cases are shown to the annotator.

4.4 Learning attributes from pairwise comparisons

In this work, we derive the raw computerised representations of the images from myocardial segmentations, their spectral signatures, and texture features. Clearly, the semantic gap between these representations and the semantic attributes is quite large. We use the obtained pairwise annotations to learn functions that are mapping the extracted features to the attribute values.

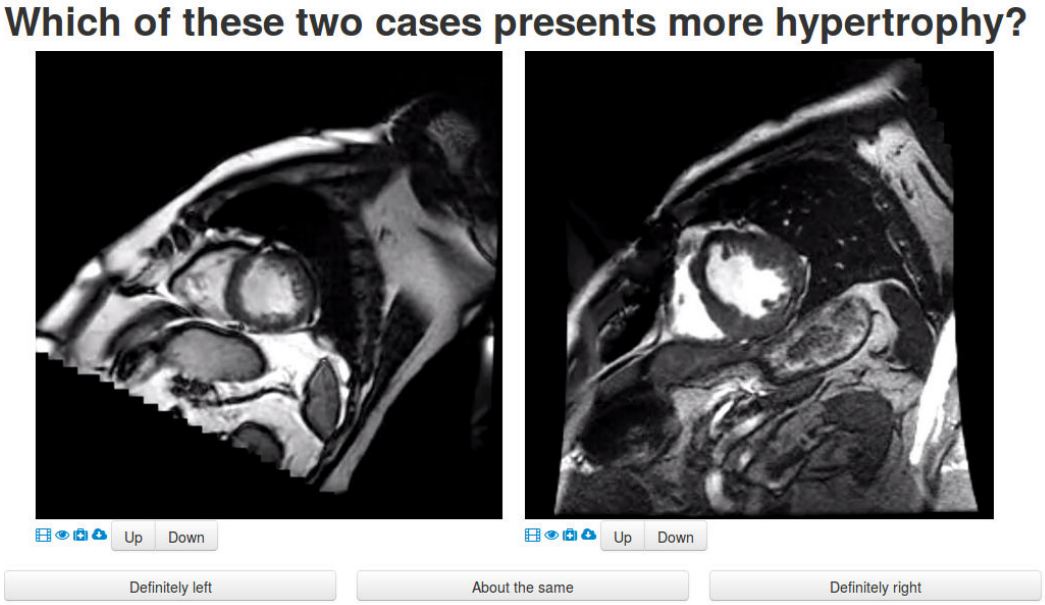


Figure 4.6: Comparison interface where video sequences are displayed and the expert is allowed to browse through the slices. The annotator can then choose the most suitable response to the posed question. We believe that in this case the image on the right is more likely to be chosen.

We borrow the ideas from Parikh and Grauman (2011) and RankSVM (Joachims, 2002) where ordering of instances is posed as a max-margin classification problem, commonly seen as the linear SVM classifier. Ranking functions that map feature measurements to continuous attributes are efficiently learnt from sparse pairwise image comparisons. These comparisons are represented in the ground-truth with only two discrete options (smaller/greater than or similar) and therefore lessen the impact of expert variability. Pedregosa and Gramfort (2012) and Huang et al. (2013) were some of the first to learn from pairwise comparisons in medical imaging. We apply a similar concept to cardiology. Contrary to Parikh and Grauman (2011) where the attributes are learned to describe each class (group of images), we aim to describe individually each heart, *i.e.*, treat each heart as a separate class.

The ranking functions will help us to automatically reproduce the image comparisons and allow us to describe the cardiac images with an attribute vector. We will later show that even with a small number of collected samples we observe a clearly positive trend in the prediction of pairwise image ordering.

We describe each image I^i with a set of M scalar attributes $A^i = \{a_m^i\}$. This description then leads to a natural way of comparing the images, *e.g.*, one heart is more hypertrophied than the other, or one image is of a poorer image quality than another one. However, only a set of features $\theta^{\mathbf{i}}$ can be directly measured and not the attribute value itself a_m^i. Our goal is to learn for each attribute a ranking function $r^m(\theta^{\mathbf{i}})$ that best satisfies the two expert provided pairwise constraints:

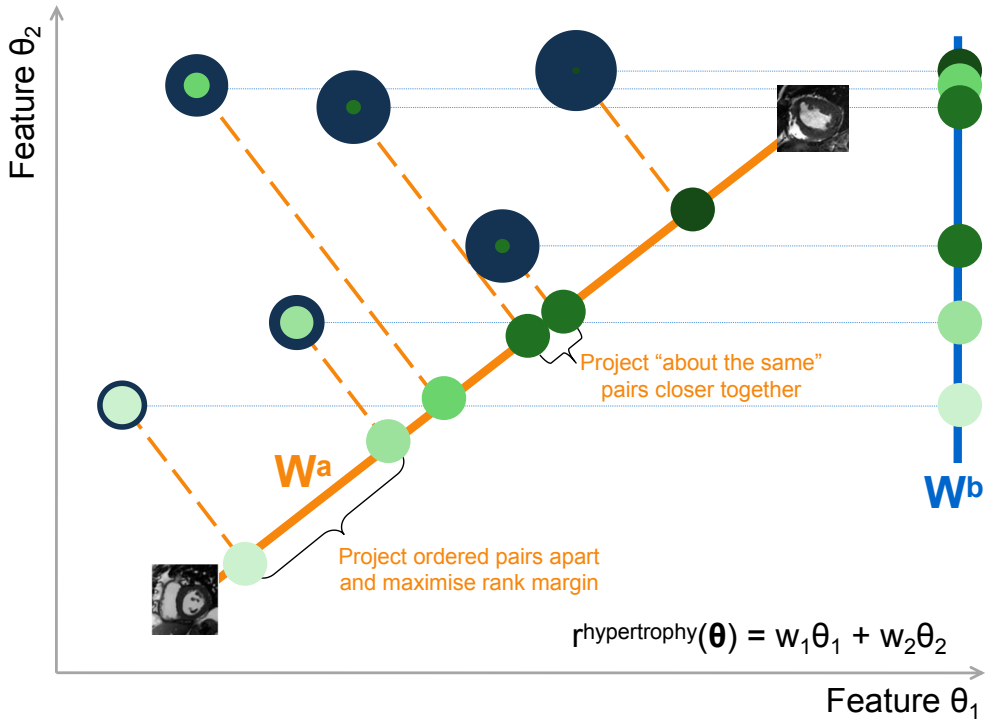


Figure 4.7: Ranking function estimation in a synthetic 2-dimensional feature space. Circles represent different hearts, their thicknesses the corresponding degrees of hypertrophy. A computationally efficient way to find the optimal projection $\mathbf{W}$ from pairwise comparisons that maps the feature vector θ to the attribute values can be derived from an SVM-like formulation. The goal is to map similar hearts closer together and the ordered pairs more spread apart. In this example, $\mathbf{W_a}$ (orange) is a clearly superior projector over $\mathbf{W_b}$ (blue) in preserving the ordering with respect to the hypertrophy attribute.

Attribute value ordering constraint when $a_m^i > a_m^j$

$$\forall (i,j) \in O_m : r^m(\theta^{\mathbf{i}}) > r^m(\theta^{\mathbf{j}}) \tag{4.1}$$

Attribute similarity constraint when $a_m^i \simeq a_m^j$

$$\forall (i,j) \in S_m : r^m(\theta^{\mathbf{i}}) \simeq r^m(\theta^{\mathbf{j}}) \tag{4.2}$$

As in Parikh and Grauman (2011), we use a linear ranking function $r^m(\theta^{\mathbf{i}}) = \mathbf{W_m}^T\theta^{\mathbf{i}}$ that is not only rapid to optimise but also guarantees a monotonic mapping between the features and the estimated attribute value. Learning of the ranking function is then approximated by solving the following convex optimisation problem (Eq. (4.3)), similar to the formulation of the linear SVM classifier (Cortes and Vapnik, 1995) (see Fig. 4.7).

$$
\begin{aligned}
W_m = \text{argmin} \quad & \left(\frac{1}{2} \|\mathbf{W_m}\|_2^2 + C_O \sum {\xi_{\mathbf{ij}}}^2 + C_S \sum {\gamma_{\mathbf{ij}}}^2 \right) \\
\text{subject to} \quad & \mathbf{W_m^T} \left(\theta^{\mathbf{i}} - \theta^{\mathbf{j}} \right) \geq 1 - \xi_{\mathbf{ij}}; \forall \left(i, j \right) \in O_m \\
& \left| \mathbf{W_m^T} \left(\theta^{\mathbf{i}} - \theta^{\mathbf{j}} \right) \right| \leq \gamma_{\mathbf{ij}}; \forall \left(i, j \right) \in S_m \\
& \xi_{\mathbf{ij}} \geq 0; \gamma_{\mathbf{ij}} \geq 0
\end{aligned}
\tag{4.3}
$$

Here, C_O and C_S are the trade-offs between margin maximisation between the ranked points and satisfying the pairwise ordering constraints (C_O) and pairwise similarity (C_S) constraints provided by the experts. The slack variables $\xi_{\mathbf{ij}}$ and $\gamma_{\mathbf{ij}}$ are penalties for incorrect ranking.

4.5 Spectral description of cardiac shapes

If we consider the heart as a binary object on a closed and bounded domain $\Omega \subset R^d$, it can be represented with its spectral signature. In essence, the spectral signature is representing the hearts' modes of vibration, similar to how a drum membrane would vibrate when hit by a stick. This has been used to globally describe the intrinsic geometry of the heart shape and to discriminate between healthy hearts and repaired tetralogy of Fallot patients (Bernardis et al., 2012; Konukoglu et al., 2012), and similarly by Lombaert et al. (2012) to establish better correspondences for inter-patient cardiac image registration. The main advantage of this method over the traditionally used active shape methods is that no shape correspondences are required. The description is also invariant to rigid transformations.

The spectral signature is determined by the spectrum of the Laplace operator Δ_Ω defined on a domain Ω as shown in Eq. (4.4).

$$
\Delta_\Omega f \triangleq \sum_{i=1} \frac{\partial^2}{\partial x_i^2} f, \forall \mathbf{x} \in \Omega \tag{4.4}
$$

Eigenvalues λ_i and eigenfunctions f_i of the operator Δ_Ω are solutions of the Helmholtz equation with Dirichlet type boundary conditions, $\partial\Omega$ is the object's boundary and $\mathbf{x}$ are the image coordinates.

$$
\Delta f = \lambda f \ \forall \mathbf{x} \in \Omega, \forall \mathbf{x} \in \partial\Omega, f\left(\mathbf{x}\right) = 0 \tag{4.5}
$$

There are infinitely many pairs (λ_i, f_i) that satisfy this equation. The spectral shape description is formed by ordering N smallest eigenvalues such that $0 < \lambda_1 < \cdots < \lambda_N$.

In our case, we describe mid-cavity slices of each heart by the 100 smallest eigenvalues of the spectrum; separately for the cavity and the myocardium at the end-diastolic and end-systolic frame. We concatenate these vectors and call it $\theta_{\mathbf{s100}}$ in the evaluation. We also compute temporal volume curves of the cavity blood pool and the mass of the myocardium. We concatenate these measurements with the spectral feature vector and obtain the final shape vector $\theta_{\mathbf{s100,vm}}$. Please note that

we use manual segmentations in this work to discount the effect of bad automatic segmentation. There is a large body of work on automatic or semi-automatic cardiac image segmentation (see Section 3.1 for a brief overview). Any sufficiently accurate and regularised segmentation method can be used to extract the binary masks of the myocardium and the blood pool for the spectral representation.

4.6 Texture features to describe image quality

Images corrupted by the imaging artefacts might be required to be removed (Markonis et al., 2012) from the retrieval results. Twelve criteria were proposed by Klinke et al. (2013) to assess the CMR image quality based on the present artefacts. These include banding or field of view wrap-around artefacts (often seen as dark curved lines across the images), respiratory and cardiac motion ghosting (replication of thoracic structures), mistriggering (image blurring), metal artefacts (changes of magnetic field homogeneity, *e.g.*, around sternal wires), shimming artefacts (dark bands across the image caused by off-resonance) and flow related artefacts.

The spectral shape or volumetric features described above are not capturing any of this information. We assume that the main information about image quality can be inferred from the texture statistics of the images. To capture this information we use the first thirteen textural features defined by Haralick et al. (1973).Haralick features are a set of measures (such as contrast, entropy, or variance) on local grayscale co-occurrence matrices. These features make our texture feature vector $\theta_{\mathbf{h}}$.

4.7 Evaluation and results

The goal of relative attributes is to establish pairwise comparisons between any pair of images, and then be able to order them with respect to the attribute of choice.The absolute ground-truth values $a^{left,gt}$ and $a^{right,gt}$ and not known except for their relative ordering. As we do not have a global ordering nor the value of the attributes as ground-truth, we report average success rate of order prediction for image pairs simply as $S = P/(P+Q)$. Where P is the number of concordant pairs, *i.e.*, pairs where the the estimated pairwise comparison of the left and right images $r^{left} < r^{right}$ is in agreement with the ground-truth comparison $a^{left,gt} < a^{right,gt}$ for a particular attribute. Q is the number of discordant pairs, *i.e.*, pairs where the estimated and ground-truth comparisons differ from the prediction. We ignore ties (pairs labelled as "about the same") in the evaluation.

4.7.1 Data and preprocessing

The data used in this study consists of 96 post myocardial infarction acquisitions of short axis Steady State Free Precession (SSFP) 3D+t stack of varying resolution and image quality from the DETERMINE dataset (Kadish et al., 2009). We use the manual expert segmentations of the LV myocardium.

4.7.2 Evaluation

So far, we have collected 329 pairwise annotations from 3 in-house annotators for all of the attributes altogether. From the annotated pairs (and for each attribute individually) we randomly choose 80% for training and then use the remaining 20% to measure the success rate of the order prediction. This is repeated 5 times and averaged. The number of collected pairs for each attribute and the mean results after a 5-fold cross validation for different combinations of features is shown in Table 4.1 below.

	hypertrophy	**kineticity**	**dilation**	**thinning**	**im. quality**
# pairs	76	41	75	67	70
$\theta_{\mathbf{s100}}$	83.5	**74.5**	67.0	58.0	52.0
$\theta_{\mathbf{s100,vm}}$	**92.5**	71.5	**73.5**	**73.0**	52.5
$\theta_{\mathbf{h}}$	64.5	51.5	54.5	47.5	**72.0**

Table 4.1: Mean pairwise order success rates for different attributes trained with empirically chosen $C_O = 1$, $C_S = 0.2$ with different feature sets (s100 — first 100 smallest eigenvalues of the Laplace operator spectrum for the blood pool and myocardium at end systole and end diastole, vm - end-systolic and end-diastolic volumes of LV cavity and myocardial mass and h-Haralick features).

On average, our method estimates the correct pairwise orders for the majority of the pairs in the test set with success rate significantly higher than chance (50%). Spectral features, together with volumetric and mass measurement, seem to be a decent shape description to predict image ordering for the hypertrophy attribute. The success rate of 92.5% means that when querying the database for images with hypertrophy larger than some reference value, approximately 9 out of 10 retrieved images (on average) correctly contain more hypertrophied hearts than the reference. The performance of the spectral description drops for the thinning and dilation attributes where the thin myocardium poses a challenge for the graph construction. Using the spectral description together with volumes and myocardial masses improves the ranking estimate. As expected, using these features to predict the image quality attribute is not better than flipping a fair coin. The Haralick features at least partly capture the image quality attribute.

4.8 Conclusions and perspectives

We applied the approach of relative attributes to cardiac imaging, to automatically compare images of post myocardial infarction hearts. Relative attributes can help to bridge the gap between extracted image features and semantically meaningful

descriptions of the images. In our preliminary study, we obtained encouraging pairwise image ordering based on the estimated semantic attribute values. However, collecting a much larger training set and/or better features will be needed to capture the continuous attribute values, to extrapolate to extreme values of the attributes, and to completely sort the dataset.

In this study, the attributes were global, without any spatial information. This description is invariant to the underlying cardiac condition and by adding extra attributes it should be possible to extend this work to more complex pathologies. For a congenital heart disease, such as the Tetralogy of Fallot (TOF), relevant attributes describing the right ventricle, the pulmonary valve, and the ventricular septum should be added. Here, we would likely discover that in general the right ventricles are more dilated and hypertrophied than normal hearts. This approach also opens up possibilities for faster creation cohorts for studies using the attribute values, *e.g.*, select hearts with a hypertrophy larger than some reference case and with sufficiently high image quality. We also have to remind the reader that describing cardiac images without any motion information is not ideal. Adding motion could help us to introduce attributes such as asynchrony, stiffness, and contractility, or better estimate the kineticity of the myocardium. Also, although linear ranking functions are very fast to train and test, they might not create the most optimal mappings between the features and attributes.

4.8.1 Perspectives

We developed a crowdsourcing platform to start collecting ground-truth in cardiology. Once more annotations from several experts are collected, it might be possible to compare rankings of different experts and weigh their contributions in the cost function accordingly. The more certain decisions usually happen much faster. The time to decide could therefore be used as the decision strength.

Further improvements can be done to reduce the required annotators' time. Smarter choices of the image pair — attribute triplets could be done by presenting the most similar images (Tamuz and Belongie, 2011) and to avoid annotating the obvious cases. Currently, the annotators are free to ignore the transitive relations between the annotated images (if A is greater than B, and B is greater than C, A does not have to be annotated as being greater than C). Therefore, probabilistic modelling of the attributes and their comparisons (*e.g.*, as done with TrueSkill in Burggraaff et al. (2015); Herbrich et al. (2007)) should be explored. It would be useful to have a measure of confidence in the attribute and to decide, which cases need more input and which attributes are less certain. A gamified interface (von Ahn and Dabbish, 2004) should be designed to improve engagement and learning of the annotators.

Approaches like this one will help to reduce the semantic gap and enable natural language queries of the databases, *e.g.*, Retrieve images like this one, but with a more dilated ventricle and with higher image quality. They also allow us to retrieve the images using a set of sliders specifying the desired attribute ranges (Kovashka

et al., 2012). They could be also used directly to assess relevance of the images for image retrieval. The ground-truth collection application could then ask questions such as “If you had a patient with the current image, which of the images presented are more relevant for the diagnosis and prognosis?”

Chapter 5

Learning how to retrieve semantically similar hearts

Contents

Based on our work in preparation for submission (Margeta et al., 2015a) and our work presented at Statistical Atlases and Computational Modeling of the Heart (STACOM) workshop (Bleton et al., 2015).

Chapter overview

In this chapter, we propose a method for automated content-based retrieval of semantically similar hearts by learning how to approximate the similarity measure between the images. We build upon the Neighbourhood approximating forest (NAF) (Konukoglu et al., 2013) algorithm which we train to capture similarities between cardiac images and to allow efficient retrievals of hearts from the most similar patients based on clinical criteria. We illustrate its use on a database of post-myocardial infarction patients.

In Bleton et al. (2015), we already showed how cardiac neighbours can be used to locate infarcts from dynamic LV segmentations without injecting any contrast. Here, we combine spatio-temporal image-based features and the NAFs with ejection-fraction-derived similarity to find hearts with similar pumping function. No image segmentation is required.

5.1 Content based retrieval in medical imaging

An important step in the diagnosis process is to compare findings with the state of the art literature and collections of previously treated patients. This is particularly useful when found abnormalities are not known. Finding the most similar images is also useful to better estimate prognosis, to choose from the available treatments, to better predict their outcomes, and to discuss with patients possible impact on their quality of life. Retrieval of similar patients can also indicate cases that can be retrospectively studied as virtual patients. The CBIR systems aim to fill up these needs.

5.1.1 Visual information search behaviour in clinical practice

Markonis et al. (2012) conducted a survey on visual information search behaviour and requirements on a sample of 34 young radiologists. Despite the limited sample size and potentially biased selection, some insights on the use cases of such systems can be drawn.

5.1.1.1 Personal annotated collections

At the time of the survey, the radiologists mostly relied on text-based web search on Google, PubMed, Goldminer, eAnatomy, or Eurorad in order to find published literature with relevant findings. They often maintained local personal collections of images, annotated with keywords, and tagged interesting cases for later reuse in clinical practice and teaching.

Without these personal annotated collections, searching within large collections of image data can rapidly become tedious. Such a search involves slow retrieval of the data from PACS archives, and interpretation of previously unseen 3D+t images to find cases with relevant characteristics. The DICOM tags are also not very

reliable with error rates up to 15% (Gueld et al., 2002). The search within hospital databases is therefore only the third most popular option.

Restricting the search to the personal annotated collections instead of the much larger and richer multi-centre databases seriously hampers the possibility to transcend boundaries of the clinicians' own experiences, and to discover relevant and rare cases available elsewhere. Nonetheless, this strategy compensates for the limited ability to search within the PACS databases, and for the text-only search within the clinical literature.

5.1.1.2 Time before quitting

During clinical activities, the average time before quitting when searching through the databases was between 5 and 10 minutes. To diagnose unknown abnormalities, an efficient lookup of relevant cases in large hospital databases is necessary. Even shorter times are needed for exploratory search (Kules et al., 2009) where quick turnaround is essential.

5.1.1.3 Desired functionalities

Based on the survey, the most often sought functionalities of the CBIR systems are: search by pathology, modality, and search for "similar images". The surveyed radiologists judged the usefulness of the retrieved images based on their experience and comparison with the queried case, and by matching image properties (*e.g.* the image modality). They also found important to filter the images based on their quality.

5.1.2 Where are we now?

In Chapter 2, we showed how to automatically index the CMR images based on the acquisition plane information. This information is one of the most important properties of cardiac MR slices. With segmentation (Chapter 3), we describe the hearts with geometric indices. In Chapter 4, we addressed describing the images with their semantic description, and the image quality. This description simplifies query formulation and can help with exploratory search for similar images within the image collections.

Our next technical challenge is to facilitate querying of cardiac databases (perhaps less interactively) from image content for "similar" images. We therefore need to find a flexible way of estimating the similarity between any pair of images to approximate the clinicians' metrics (*e.g.* similarity in image-derived measures or pathology). In this chapter, we propose a technique that learns how to approximate such similarity from the examples.

5.2 Similarity for content-based retrieval

When are two cardiac images "similar"? The similarity metric depends on the clinical application of the CBIR system. A different similarity measure is needed to find images with equivalent findings than to find hearts with comparable prognosis. And yet another one to find similar hearts for initialisation of a segmentation algorithm.

As we will see, the measure of similarity between the images for retrieval is frequently defined from distances of some intermediate representations such as texture histograms or segmentation overlaps. This creates a discrepancy between what these measures capture and the desired clinical similarity (the *semantic gap*). In Section 4.1.1, we have already described a couple of methods to extract features from cardiac images.

5.2.1 Bag of visual words histogram similarity

One of the most common ways to describe and compare medical images for retrieval is through their tissue or texture statistics. *E.g.*, Depeursinge et al. (2012) used histograms of tissue volumes to retrieve similar lung images. Instead of these volumes requiring image segmentation, André et al. (2011a) used histograms of visual words to describe and retrieve images from confocal endo-microscopy. Both approaches define the similarity as a measure of distance between the histograms. This creates a semantic gap between the measured and the perceived similarity.

André et al. (2012) later managed to reduce this gap by learning the distance between images using the perceived clinical similarity as ground-truth (helping to discriminate healthy tissue from malignant tumours). In their work, clinicians annotated pairs of videos on a four point scale from very dissimilar to very similar. This is comparable to our estimation of semantic cardiac attributes from pairwise image annotations presented in Chapter 4.

5.2.2 Segmentation-based similarity

Segmentation overlap can describe the global similarity of shapes. Glatard et al. (2004) used a texture-based method to roughly segment the cardiac images. An overlap measure between these segmentations then served to find the most similar slices. Neighbouring slices and frames from the same 3D+t stack were correctly retrieved. However, retrieving neighbouring slices from the same patient does not aid with the decision-making process. In addition, even Euclidean distance would likely perform well in this scenario.

Segmentation overlap (measured with Dice coefficient) was also used by Eslami et al. (2012) to find the most similar hearts and to use them as atlases to guide their LV segmentation algorithm. Computing overlaps of fine structures (*e.g.* thin myocardium) can cause problems. Small misalignments of these structures can cause zero overlap and negatively affect the similarity measure.

In Bleton et al. (2015), we showed how spatio-temporal myocardial thickness profiles extracted from image segmentation, instead of the segmentation overlap,

can be used for fine-grained infarction localisation (without contrast agent).

5.2.3 Shape-based similarity

Statistical shape models are powerful tools to capture fine differences between various cohorts (Zhang et al., 2014b; Medrano-Gracia et al., 2014). The hearts are represented with the same model across the dataset and reduced to a lower-dimensional representation via PCA. The modes of variation then form a descriptor. Such representation can be used to compute distances between the cases, *e.g.*, as the Euclidean distance. Correspondences between the hearts must be first established to construct the shape models.

Bernardis et al. (2012) extracted a pose invariant spectral representation (see Section 4.5) from image segmentations to describe shapes of the hearts without the need for the correspondences. They used Weighted Spectral Distance (Konukoglu et al., 2012) to measure the shape dissimilarity.

5.2.4 Registration-based similarity

A different approach was introduced by Duchateau et al. (2011). They used image registration in a diffeomorphic framework to construct an atlas of septal motion. For each heart in the database a septal motion abnormality map was computed. This map highlights the septal flash pattern — a factor to predict cardiac resynchronisation therapy response. Euclidean distances between these abnormality motion maps then serve to compute pairwise similarities between the hearts and find a lower dimensional representation of the dataset. This method characterises the septal motion and disregards the shape differences.

To capture shape similarity between different pathologies, Ye et al. (2014) proposed to characterise cardiac ventricles via a deformation metric. Similarly to Duchateau et al. (2011), deformation maps to a reference model allow them to measure distances from normality but also to classify the hearts into four distinct pathological classes: healthy, Tetralogy of Fallot, hypertrophic cardiomyopathy and hearts with diastolic dysfunction.

5.2.5 Euclidean distance between images

Instead of using the motion abnormality (Duchateau et al., 2011) or deformation (Ye et al., 2014) maps as the intermediate representation, an even simpler approach exists. This is to use the image intensities of rigidly registered images directly and compute the Euclidean distance between them instead. Such approach was recently presented by Wang et al. (2015) and was validated on a dataset of 209 cardiac images. Their lower-dimensional image representations (Belkin and Niyogi, 2003) derived from this distance make visually appealing visualisation of the dataset and reveals clusters of patients with similar sex, blood pressures, ventricular volumes, but also some appearance similarity. These representations of the dataset were used successfully for binary classification within the three categories (sex, blood

pressure, volume). This is an appealing finding. Images used in their study were all acquired in a controlled environment and on the same MR system, likely with the same acquisition settings.

In our experiments with the multi-centre study DETERMINE (Kadish et al., 2009), we used images cropped similarly to Wang et al. (2015). However, we failed to retrieve images with similar characteristics using Euclidean distances between them. We did not project the images into a low-dimensional embedding though. The Euclidean distance between pixel ints ensities is a measure particularly sensitive to the acquisition differences (see Fig. 3.8 for an illustration of myocardial tissue intensity variability across the dataset) and image artefacts. It is also not necessarily the best measure to address some of the clinical retrieval goals.

To remove the differences between the acquisitions we can strip the images down to the bare minimum, *i.e.*, to segment the images. The other options are to do correct preprocessing or to pick features invariant to irrelevant image differences.

5.2.6 Using decision forests to approximate image similarity

Most of the above mentioned methods use Euclidean or other fixed distances between the fixed intermediate representations. Learning-based techniques can be used to capture the similarity, reduce the semantic gap (André et al., 2012) and pick acquisition invariant features. One way to learn to approximate the similarity is via decision forests.

Pei et al. (2013) recognised that the decision forests can be used to find similar motion patterns for lip reading. They used an unsupervised clustering criterion to train the forest, grouping dense clusters of points together (comparably to the Distance approximating forests in Appendix A). Their approach was termed in Criminisi et al. (2011b) as manifold forests.

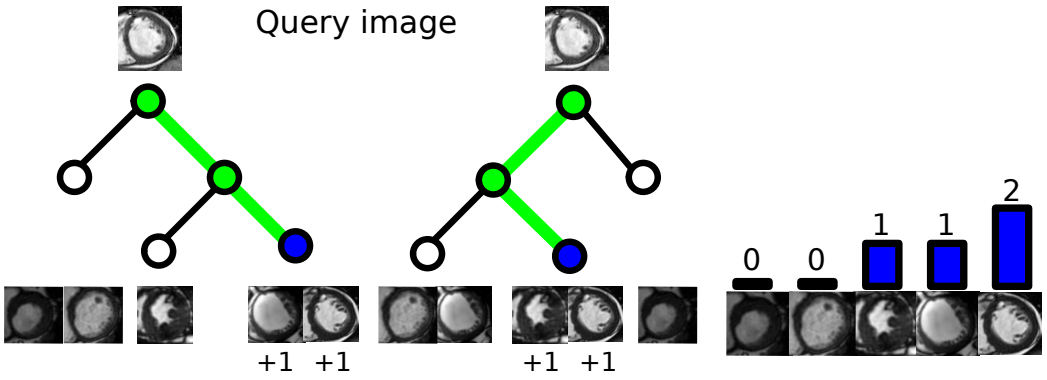
In medical imaging, the decision-forest-based distance was first used by Gray et al. (2013) for cerebral images, to discriminate Alzheimer disease patients from mildly cognitively impaired and healthy controls. Here, two classification forests were trained. One was trained with features derived from volumetric measurements of anatomical regions (after segmentation). Another one with features derived from Positron emission tomography (PET) image intensities at random locations (similarly to our recognition of cardiac acquisition planes with image miniatures in Section 2.5). These forests were trained to classify the brains into Alzheimer disease and healthy. The trained forests were then used to compute the similarity between the images. As it is typical for decision forest-based methods, the algorithm automatically selected relevant discriminative features.

The notion of similarity in the forest-based approaches (the number of times two data points end up in the same leaf) (Gray et al., 2013; Criminisi et al., 2011b; Pei et al., 2013) might seem quite ad-hoc at first. Intuitively, it makes sense to use forests to find similar images. Similar images take similar decisions and pass through similar paths. The selected decisions are adapted during the training phase so that the performance on the target task (*e.g.* classification) is maximised. If the

forest is well trained, the paths are then relevant to the similarity.

5.3 Neighbourhood approximating forests

Konukoglu et al. (2013) showed how decision forests can be trained in a supervised way from pairwise similarity ground-truth to find similar images directly. The similarity is specified in the split criterion. They called this approach the Neighbourhood approximating forest (NAF).



(a) Structuring the dataset with neighbourhood approximating forests.

(b) Similarity to all images in the database.

Figure 5.1: The NAFs are trained to group similar hearts (*e.g.* with respect to similar ejection fraction) together. These similar images end up in the same leaves more often than with the dissimilar ones (Fig. 5.1a). New, previously unseen, images can be used to rapidly query the database by counting the number of common leaves with the database. When querying a new image, the image is passed through decisions of the forest (thick green path) and reaches a set of leaves (blue). The similarity with each database image (Fig. 5.1b) is computed simply as the number of times the query image and the database image co-ocur in the same leaf.

The NAF is a particularly advantageous algorithm for CBIR. We do not need a predefined metric between features. The trees automatically select relevant features from the images and cluster the closer (based on the provided similarity ground-truth) images together to achieve the clinically meaningful goal.

The tree data structure also allows us to rapidly evaluate similarities between the query image and all existing cases in the database. To compare the query image to all previously stored images, we count the number of times the query image co-occurs with each of them in the leaves of the forest. Even for very large datasets, only as many leaves need to be visited as there are trees in the forest. This method is computationally efficient, especially, if the features are also efficiently extracted.

It is possible to insert new cases into the database without retraining or doing additional pairwise comparisons. Once the new image is inserted into the database and passed through the forest, reached leaf indices can be recorded. There is no

longer any need to have direct access to the image for future comparisons. This is compatible with collaborative indexing of personal collections from several clinicians. Sharing sets of leaf indices from images in personal collection would be sufficient to find where the relevant images can be found.

5.3.1 Learning how to structure the dataset

This tree-based method hierarchically structures the dataset such that similar images are clustered closer together while the dissimilar ones are spread more apart. Since this is a supervised method, pairwise distances $\rho(I, J)$ between images I and J in the training set must be known. The ground-truth similarity between the cases can be defined from virtually anything meaningful, *e.g.*, pathology, blood pressure, patient's age, or in our case the ejection fraction, infarct location (Bleton et al., 2015), or the presence of pericardial effusion (Appendix C).

In lieu of the information gain used previously (see Eq. (2.1)), a criterion derived from cluster compactness is used to select split parameters that lead to the best partitioning of the dataset. Cluster compactness C^ρ for a set of images D^s at the node s and with respect to the training image distance ρ is defined as:

$$C^\rho(D^s) = \frac{1}{|D|^2} \sum_{i \in D^s} \sum_{j \in D^s} \rho(I_i, I_j) \tag{5.1}$$

During the training phase, at each node, the parameters of the binary splits θ_s are fixed in order to divide the images at the node into the left I^s_L and the right I^s_R branches. The parameters are chosen such that the following gain measure G^ρ is maximised:

$$G^\rho(I^s, \theta_s) = C^\rho(I^s) - \frac{|I^s_L|}{|I^s|} C^\rho(I^s_L) - \frac{|I^s_R|}{|I^s|} C^\rho(I^s_R) \tag{5.2}$$

To maximise this quantity, the compactness (*e.g.* the pairwise sum of absolute ejection fractions differences for all the pairs within the node) in both left $C^\rho(I^s_L)$ and right $C^\rho(I^s_R)$ branch must be as low as possible. This quantity therefore encourages the forest to pick decisions where similar images (with similar ejection fractions) are put together into the same branch.

5.3.2 Finding similar images

The trained trees can be then used to obtain a measure of similarity S between any two images I and J (even if they were not part of the training set). After passing each image through the trees, a set of leaves is reached (see Fig. 5.1a). The similarity is approximated by counting how many times the same decisions were taken, *i.e.*, by measuring the number of trees where two images both reach the same leaf ($L_t(I) = L_t(J)$) as illustrated in Fig. 5.1b.

$$S(I_i, I_j) = \sum_{t=1}^{n_T} 1(L_t(I_i) = L_t(I_j)) \tag{5.3}$$

Please note that the similarity value is not continuous as one would expect. It is quantised into as many discrete steps as there are trees in the forest (n_T). If two images share no leaf, the similarity between them is null. On the other hand, the similarity between them is equal to the number of trees when the images share all decision paths and always reach the same leaves. For a more detailed description of the NAFs algorithm, please see Konukoglu et al. (2013).

5.3.3 NAFs for post-myocardial infarction hearts

We now apply this technique on a dataset of hearts after an acute myocardial infarction event. In Section 4.2, we described several aspects of the failing hearts that can be captured by cardiac images. Following the acute myocardial infarction the hearts undergo a series of remodelling steps. Some of the most important predictors for the current state and the later remodelling changes (global ventricular dilation, impaired contraction, or a valvular disease) is the infarction position and the amount of preserved cardiac function (Zaliaduonyte-Peksiene et al., 2013; Sun, 2009).

In Bleton et al. (2015), we used NAFs to help with interpretation of previously unseen images of post-myocardial infarction hearts by learning how to approximate similarity between them based on their infarction locations from image segmentations and to predict the infarction position. The amount of preserved cardiac function is another important factor for the course of adverse remodelling after a myocardial infarction. It is therefore important to be able to find weakly contracting hearts when a non-compacting heart on a course of adverse remodelling is queried.

5.4 Learning the functional similarity

The ejection fraction of the LV is one of the most frequently used measures in cardiac reporting to describe the cardiac function and to classify failing hearts.

5.4.1 Cluster compactness based on ejection fraction difference

For this goal, we now train a NAF with absolute ejection fraction (EF) difference as the split criterion. $\mathrm{EF}(I)$ is the ground-truth ejection fraction measure for image I computed from its LV cavity segmentation.

$$\rho_{\mathrm{EF}}(I, J) = |\mathrm{EF}(I) - \mathrm{EF}(J)| \tag{5.4}$$

5.4.2 Preprocessing

We first extract one apical, one mid-cavity, and one basal slice from each image. If there are more slices of a particular type, a single slice is selected from each of these at random. In fact, choosing random combinations of the 3-slice volumes can be seen as a way to augment the dataset.

We then align the images to a reference image with a set of landmarks belonging to the LV epicardium on the mid-cavity slice (as detailed in Appendix B.3) with a similarity transformation model (rigid + scale). The reference image is cropped to 128×128 pixels, such that it wholy contains the target landmarks with some small extra margin. The intensity of the images is normalised (similarly to whitening) by shifting the mean of the cropped images to 127 and by forcing a single standard deviation value across the whole dataset such that the most common intensity ranges (between 0 and 98.5%) fully span the 0 to 255 intensity range.

We strip away the thorax pixels (Section 5.4.2.1) and compute the temporal pseudo-volumetric curves from the image crops. We then extract the end diastolic and end systolic frames from the images (Section 5.4.2.3).

5.4.2.1 Thorax stripping

To reduce the influence of background structures, we focus on the regions adjacent to the moving left ventricle only. Such preprocessing is similar to what is commonly done in brain imaging - skull stripping. Rough estimation of the moving cardiac region can be obtained from temporal intensity changes (see Figure 5.2). We first compute the temporal range of intensity values at each pixel, in other words the difference of maximal and minimal intensity values of the pixel in time. We spatially smoothen the range image and binarise it using the method of Otsu (1979).

Intensities beyond some distance from the largest moving blob can therefore be safely labelled as the background and masked to be removed from the image. It is important to keep some extra region around the heart to avoid removing important cardiac findings such as the pericardial effusion. Smooth blending of the background is obtained by mapping the distance to the blob with a sigmoid.

5.4.2.2 Spatial alignment

In Appendix B, we describe our landmark model and how cardiac landmarks can be estimated. Once the landmarks are estimated, we can align the hearts. Having a landmark-based model allows us to rapidly choose the reference structure of interest. This is a big advantage over the intensity-based registration techniques. The main options include registration to all available landmarks, to both LV endocardial and LV epicardial landmarks, to RV endocardial landmarks alone, or to a specific subset of landmarks, *e.g.*, the antero-septal wall.

In this chapter, we choose the **LV epicardium as the reference**. We register the left ventricular epicardial landmarks to the mean left ventricle epicardium model with similarity (rigid + scale) transformation model. The transformation parameters are estimated with Procrustes analysis (Suinesiaputra et al., 2004). The similarity transformation preserves most of the image information and puts the hearts into correspondence without introducing artefacts from non-linear methods. The standard acquisitions were performed by expert radiologists so the long axis is in general in good alignment with the short axis plane normal vector. However, the

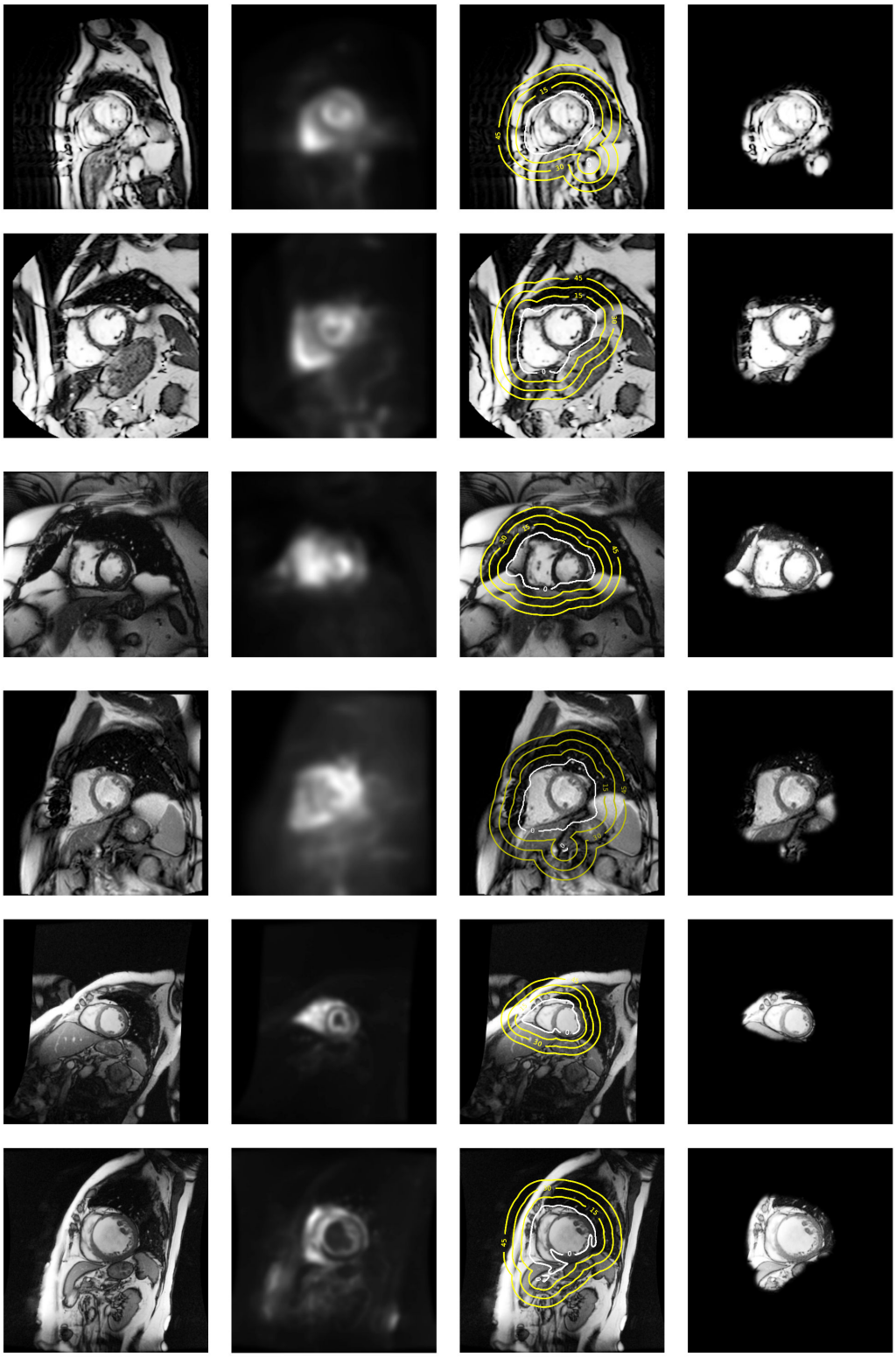


Figure 5.2: Thorax stripping with intensity range images. From left to right: original image, smoothed max-min range image, isocontours to the largest moving object (white contour is the Otsu thresholding of the smoothed image), and final image with faded background beyond 15 mm from the moving blob.

in-plane rotation is arbitrary. We limit the similarity transformation to rotations around the image normal (within the (x-y) plane). This diminishes the effect of image interpolation artefacts along the coarsely sampled z-axis of the CMR images.

Having the images registered allows us to directly compare the different cardiac structures. Pixels at fixed locations roughly correspond to the same anatomical region and their pixel intensities can be compared as done in Section 2.5. Similar approaches have been successfully used for example in classification of patients with multiple sclerosis from depth videos (Kontschieder et al., 2014), or in a face identification pipeline (Taigman et al., 2014). In Kontschieder et al. (2014), the images were rigidly aligned with patients' heads. In Taigman et al. (2014), facial landmarks were first extracted. These were then used to estimate a piecewise affine transformation and to normalise the faces into a reference view.

5.4.2.3 Temporal alignment

Cropping the regions from the estimated landmarks and stripping the thorax also helps us to temporarily align the sequences. Approaches like Dynamic time warping (DTW) (Peyrat et al., 2010) or resampling to a fixed temporal length (Zhang et al., 2010) use the whole cardiac cycle to temporally align the cardiac sequences. To establish unambiguous temporal correspondences between the different cardiac sequences we use only the two main cardiac phases.

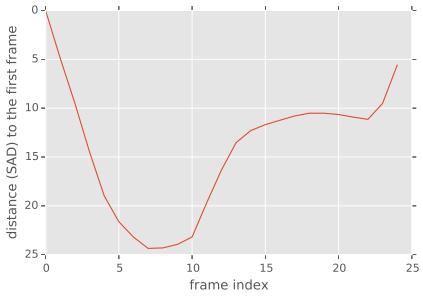


Figure 5.3: Temporal dissimilarity curve (sum of absolute differences over time) of the cropped LV region with respect to the first frame (usually the end diastole). Having cropped regions around the LV helps us to identify the main cardiac phases even without the need for the segmentation. Note the remarkable resemblance of this curve to the typical cardiac volume curve.

The first frame in the sequence is chosen as the end diastolic frame. The end systolic frame is chosen as the most different (with the largest sum of absolute differences) frame from the end-diastole (see Figure 5.3 for an example of the estimated similarity curve).

5.4.3 Spatio-temporal image features

To describe the images and to achieve acceptable query responsiveness, we propose fast temporal intensity-based features (see Fig. 5.4), using no segmentation nor non-rigid image registration. These features are derived from 3D Haar-like features, similar to the ones we used for cardiac image segmentation in Chapter 3, measuring regional intensity statistics. This time, the images are aligned and cropped with the background faded. Measures at fixed positions roughly capture information from the same cardiac regions.

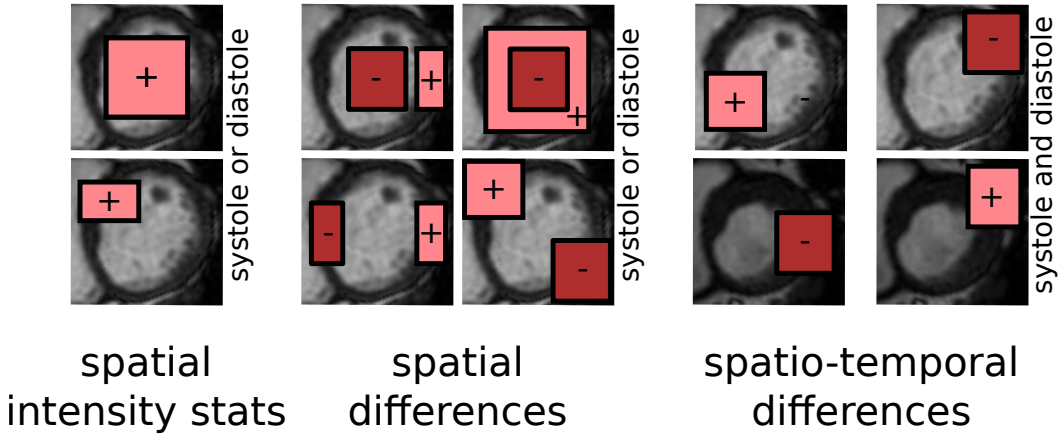


Figure 5.4: Overview of our features. Short axis images are cropped, reoriented and aligned and features are extracted for the NAFs to find similar hearts. These are derived from 3D Haar-like features computing regional intensity averages, standard deviations and maxima, together with their differences. We use several feature families in this work, both spatial only and fully spatio-temporal.

Again, we define several feature families from which we sample the features and we extract not only the **average regional intensities** within the boxes but also **standard deviations** (capturing texture and intensity homogeneity) and **regional maxima**. The first two can be computed very efficiently with integral images as proposed by Viola and Jones (2004). We find the regional maxima directly, but they can be also roughly approximated with integral images using the soft-max function $\log \sum_i \exp(x_i)$. In other words, by exponentiating the voxel intensity values prior to the integral image computation and computing logarithms of the extracted regional sums. Care must be taken to avoid numerical overflows.

In all cases, we limit the parameters of these boxes to fit into the aligned spatio-temporal image crops. We do this by parametrising the spatio-temporal rectangles with coordinates of their two extreme corners (L - lower and U - upper): $(L_x, L_y, L_t, U_x, U_y, U_t)$. Each coordinate is represented as percentage of the cropped image size and is bounded between 0 and 1. This way of defining the rectangles guarantees them to stay within the cropped image and no extra out-of-bounds box evaluation or clipping are necessary. Note, we currently always use the whole range of z values (computing statistics in voxel across all three slices), and only two main

temporal instants (diastole and systole).

In total, we define 30 "semantic" families of features for easier interpretation of the results. We sample 200 random features from each family as we will describe below. This generates a large number of features from several different feature families and assumptions on how they should capture the content. The advantage of the forest-based algorithm is that it can automatically pick features that truly matter to find similar hearts.

5.4.3.1 Intensity statistics of static central rectangles ("centered rect")

First, we propose to measure intensity statistics of the central image region, individually at systole and at diastole. A dilated and poorly contracting ventricle, with large bright blood pool should have higher systolic and diastolic average intensity than a heart whose cavity is much smaller and where the myocardium drives down the average image intensity.

We extract spatial boxes with their centre aligned with the image centre. The image centre coordinates are simply $c = (0.5, 0.5, 0.5)$. These boxes have dimensions proportional to the image crop size (in our case the extracted region is a square and so are the extracted rectangles). Their extent can be anywhere from single central pixel to the extent of the whole crop. In other words, we uniformly sample a single variable d between 0 and 0.5. The rectangles are then defined as $R^{\text{centre}}_{\text{dia}} = (c_x - d, c_y - d, 0, c_x + d, c_y + d, 0.5)$ for diastolic frames, and for systolic frames as $R^{\text{centre}}_{\text{sys}} = (c_x - d, c_y - d, 0.5, c_x + d, c_y + d, 1)$. We use the generated boxes to compute their intensity averages (avg), standard deviations (std) and maxima (max), at both end frames (sys, dia).

This results in 6 feature families (1200 features): **centered rect dia avg**, **centered rect sys avg**, **centered rect dia std**, **centered rect sys std**, **centered rect dia max**, **centered rect sys max**.

5.4.3.2 Intensity statistics of arbitrary static rectangles ("rect")

By measuring regions at arbitrary spatial positions we can capture extra information to help with more fine-grained division of the hearts based on local measurements. Therefore, we also sample randomly-sized diastolic or systolic boxes at arbitrary spatial positions. From a uniform distribution we sample two pairs of spatial coordinates x and y to lie anywhere between 0 and 1. For each coordinate, we pick the smaller of the components to create the lower corner of the box (L) and the larger ones to create the upper corner (U). We can then extract the boxes for systole $R^{\text{spatial}}_{\text{dia}} = (L_x, L_y, 0, U_x, U_y, 0.5)$ and diastole $R^{\text{spatial}}_{\text{sys}} = (L_x, L_y, 0.5, U_x, U_y, 1)$.

The following 6 feature families were extracted (1200 features): **rect dia avg**, **rect sys avg**, **rect dia std**, **rect sys std**, **rect dia max**, **rect sys max**.

5.4.3.3 Full-cycle box statistics ("rectangles full extractor")

We also compute regional statistics of arbitrarily placed rectangles as described above in Section 5.4.3.2, but this time spanning both temporal instants: $R^{\text{full}} = (L_x, L_y, 0, U_x, U_y, 1)$. This results in 3 additional feature families (600 features): **rectangles full extractor avg**, **rectangles full extractor std**, **rectangles full extractor max**.

5.4.3.4 Regional differences ("offset rect spatial diff")

For now, we have computed individual boxes and considered them independently. Comparisons between two spatial regions can help to better discriminate subtler shape differences with single decisions. We compute the spatial differences in regional average intensities, standard deviations and maxima between two boxes of different sizes and at two different positions within the same frame. We sample two independent rectangles as described in Section 5.4.3.2, *i.e.*, $R^{\text{spatial},0}_{\text{dia}} = (L^0_x, L^0_y, 0, U^0_x, U^0_y, 0.5)$ and $R^{\text{spatial},1}_{\text{dia}} = (L^1_x, L^1_y, 0, U^1_x, U^1_y, 0.5)$ for diastolic boxes.

Apart from the boxes being fully contained within the image crop (by construction), we do not restrict their positions and sizes. These boxes can even fully overlap or do not touch each other. Then, we compute the differences in their statistics $(\text{stats}(R^{\text{spatial},0}_{\text{dia}}) - \text{stats}(R^{\text{spatial},1}_{\text{dia}}))$ and use them as additional features. For systolic boxes we add 0.5 to the temporal components.

This results in 6 feature families and 1200 features: **offset rect avg spatial diff dia**, **offset rect avg spatial diff sys**, **offset rect std spatial diff dia**, **offset rect std spatial diff sys**, **offset rect max spatial diff dia**, **offset rect max spatial diff sys**.

5.4.3.5 Central temporal differences ("centered rect temporal diff")

In the end, the function of the heart is best captured with volume changes in time. Temporal intensity changes (between the diastolic and systolic frame) can help to further discriminate the hearts based on the chosen criteria. In particular, shifts in regional image intensities are an excellent mean to roughly capture the amount of motion without registration.

The intuition behind using temporal intensity differences is that average intensities in static or stiff regions do not change temporally much while properly contracting voxels cause a visible drop in average regional intensity due to changed blood-myocardium distribution. Temporal changes in the image histogram distribution due to cardiac contraction have already been used previously to estimate the ejection fraction by Afshin et al. (2012a,b).

The main intensity changes between the ED and ES should happen within the ventricular cavity. Therefore, we first define differences of statistics within equally sized centred rectangles: $\text{stats}(R^{\text{centre}}_{\text{dia}}) - \text{stats}(R^{\text{centre}}_{\text{sys}})$. This yields 3 more feature families (600 features): **centered rect avg temporal diff**, **centered rect std temporal diff**, **centered rect max temporal diff**.

5.4.3.6 Regional temporal changes ("rect temporal diff")

The central temporal differences are limited in the shape of the region they can capture. We therefore measure temporal change of rectangular regions, *i.e.*, the differences between two fixed spatial boxes: $R^0 = (L_x, L_y, 0, U_x, U_y, 0.5)$ and $R^1 = (L_x, L_y, 0.5, U_x, U_y, 1)$. In other words, apart from the temporal coordinate, the box stays fixed for the two frames and the features are computed as $\text{stats}(R^0) - \text{stats}(R^1)$. This adds 3 families (600 features): **rect avg temporal diff**, **rect std temporal diff**, **rect max temporal diff**.

5.4.3.7 Spatio-temporal differences ("offset rect avg temporal diff")

Finally, we combine arbitrary space and time together and include regional spatio-temporal diastole-systole differences of independently placed and sized boxes. One box from the diastolic frame: $R^0 = (L_x^0, L_y^0, 0, U_x^0, U_y^0, 0.5)$ and one from the systolic one: $R^1 = (L_x^1, L_y^1, 0.5, U_x^1, U_y^1, 1)$. Differences in all three box statistics of the two boxes are then computed: $\text{stats}(R^0) - \text{stats}(R^1)$, which makes for the last 3 feature families and 600 features: **offset rect avg temporal diff**, **offset rect std temporal diff**, **offset rect max temporal diff**.

5.5 Validation and results

Dataset

Our image database comes from a multi-centre study and consists of 96 post-myocardial infarction patients with cine short axis MR SSFP acquisitions (Fonseca et al., 2011; Kadish et al., 2009). The infarctions are healed, but the hearts are in arbitrary phases of the remodelling process and with arbitrary myocardial salvage.

5.5.1 Retrieval experiment

For each heart we extracted the 6000 previously described features from the corresponding images to capture the hearts' functional and appearance characteristics. We then trained the NAFs to select the relevant ones and approximate the image similarity with respect to the functional aspect of the heart — the ejection fraction. Our forests consist of 2000 trees with maximum depth 6. All training images were selected to train the trees, 500 randomly selected features and 50 random thresholds (per feature) were tested at each node.

We then queried the database with validation images in order to retrieve the most similar hearts from the training database. We evaluate the quality of the retrieval via the differences in ejection fraction of the query images and mean ejection fraction of their nearest neighbours retrieved by the NAFs.

5.5.1.1 Mean absolute prediction error

First, we divided the dataset into 10 random folds (90 percent of the data for training and 10 for testing). The mean absolute prediction error of the ejection fraction (computed as a mean of ejection fractions of the retrieved four nearest neighbours) across our cross-validation folds was $5.48\% \pm 4.60$. No segmentation was used to measure the ejection fraction. This appears to be slightly better than estimation of the ejection fractions (EFs) by visual inspection (Gudmundsson et al., 2005) with mean variability 7.0%[1].

5.5.1.2 Bucket leave-one-out

To test how well the ejection fractions are predicted across the whole range of values in the dataset, we set up a different experiment. In this step, we sorted the images in the training set with respect to their ejection fractions. We then split the sorted images into 10 equally-sized buckets (each containing roughly the same number of images), for a 10-fold leave-one-out cross-validation. At each fold we picked one of the buckets, and chose the image at its centre as the query image. All other remaining images were left for training.

The retrieval results using this leave-one-out strategy are shown in Fig. 5.5. Especially for the lower and middle EF buckets, the mean EF of the retrieved images closely matches the query ejection fraction (within 4%). The retrieved ejection fractions of buckets at the higher end of the spectrum ($> 49\%$) are much noisier. This is likely because more hearts in the dataset are in the lower EF buckets and there are more examples to learn from. Note also that the most dissimilar images are often located in the opposite side of the functional spectrum.

To visualise how the forest predicts across the whole range of EFs present in the dataset, we fit a linear regression model to the mean predictions of the leave-one-out folds (see Fig. 5.6).

5.5.1.3 Computational time

The full query pipeline (loading the image, predicting the cardiac landmarks, thorax stripping, detecting main cardiac phases, feature extraction, forest traversal and retrieval of similar images) currently runs together in less than 30 seconds. If the features are pre-calculated, the tree traversal and similarity computation with every image in our database is performed in a fraction of a second. Wrapping this tool into a friendly user interface could enable fast interaction with the dataset for exploration.

5.5.2 Feature importance

We mentioned a few times in this thesis that the forest picks the relevant features by itself and allows inspection of the decisions taken. This is a great benefit compared

[1] Try it for yourself: http://www.cardiacejectionfraction.com

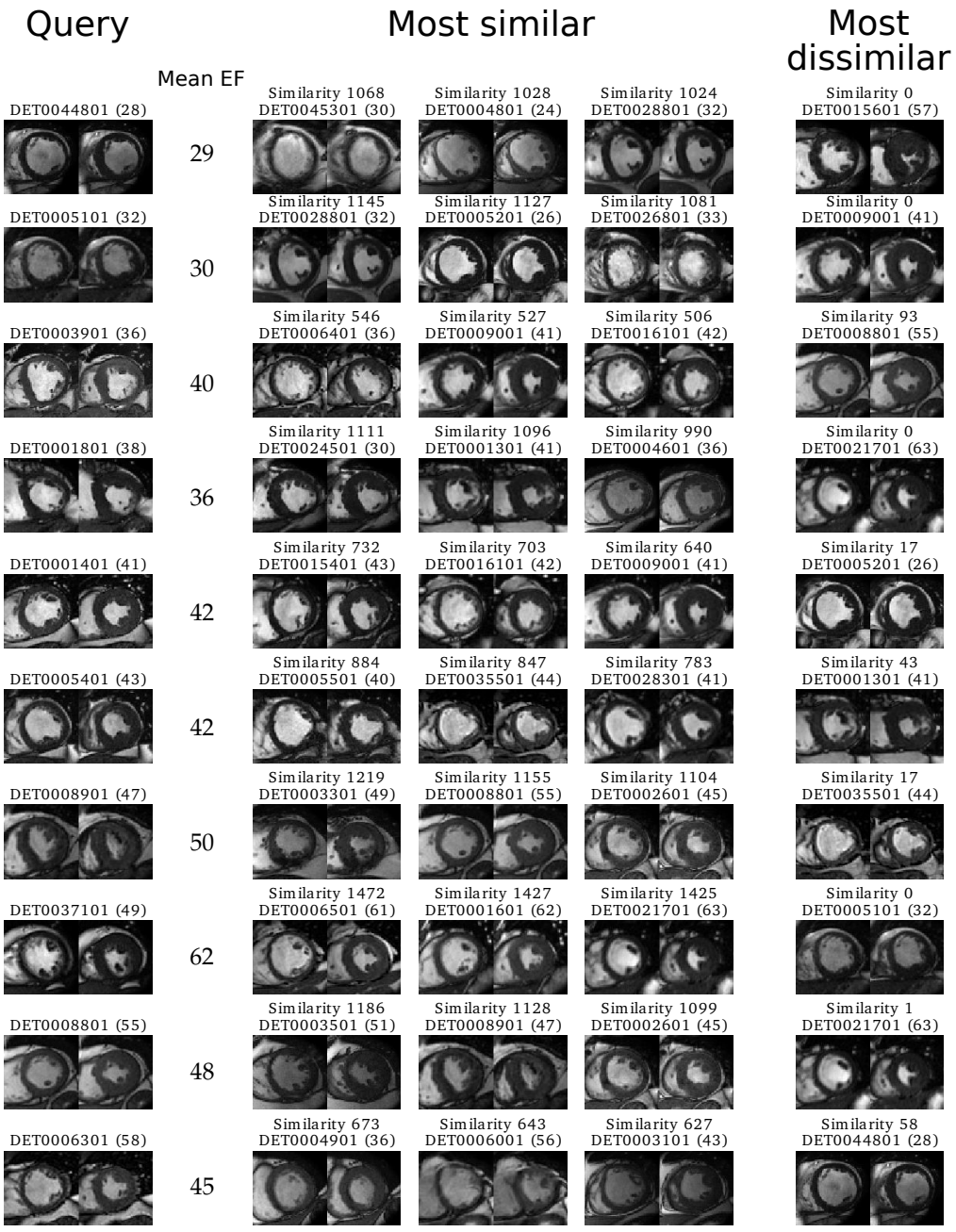


Figure 5.5: Retrieval of similar hearts based on the function criterion with a forest of 2000 trees. Each row represents a query and database retrieval results. The leftmost heart is the query image (shown diastolic and systolic mid cavity slices), next is the mean ejection fraction of the retrieved neighbours, the three most similar hearts, and the most dissimilar heart. Similarity (the number of shared leaves) to the query is marked above the images. The ground-truth ejection fractions are marked next to the case name in the parentheses.

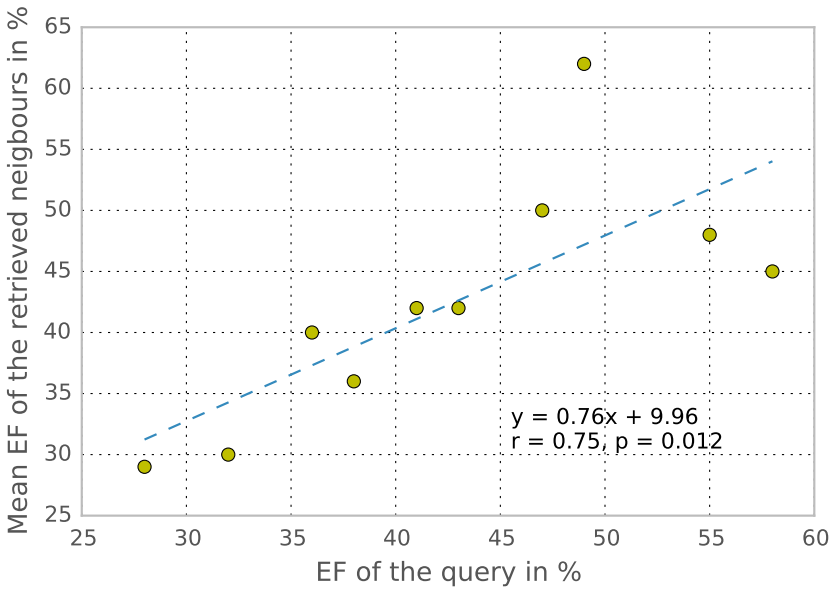


Figure 5.6: Predicting ejection fractions from the retrieved nearest three neighbours for centres of evenly sized EF buckets as queries. Removing the last three buckets significantly boosts the prediction accuracy.

to some of the more "black box" algorithms where the interpretation of the learnt models is more difficult. But what does it mean in practice? There are many different ways to capture and visualise importance of the features.

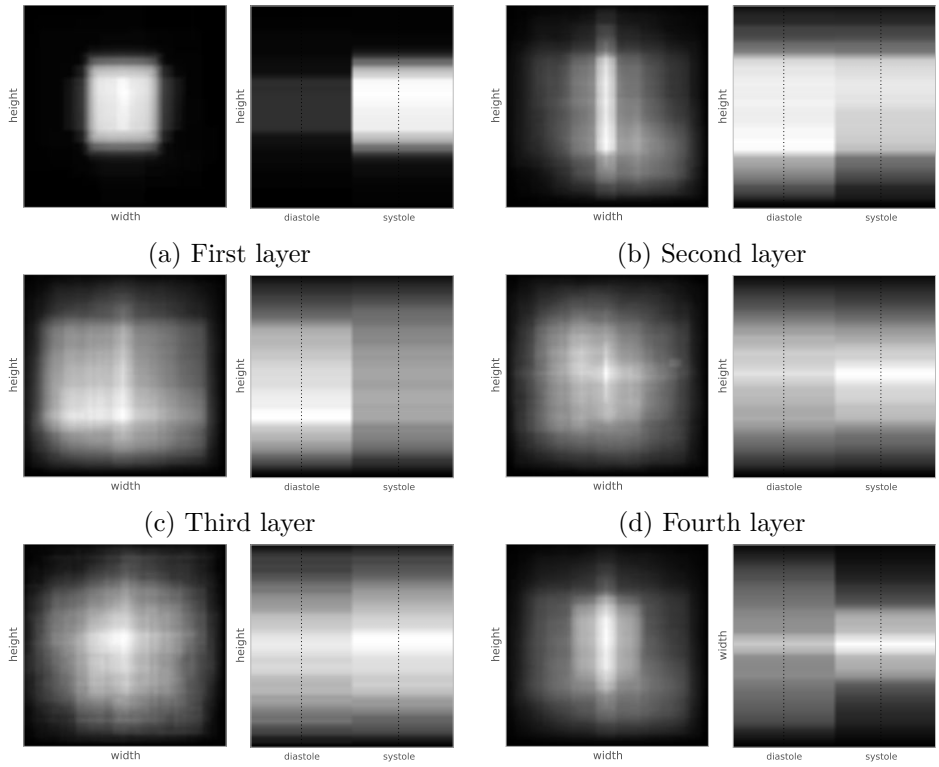
5.5.2.1 Visualising spatio-temporal importance

We draw our inspiration from the work of Kontschieder et al. (2014). Let's start with a visualisation of the first few decision layers (or levels of the trees) of the forest. In Fig. 5.7 we show how often each voxel contributed to the decision across the whole forest (on average). In practice, we count the number of times each voxel was contained in one of the supporting rectangles of the selected features.

Looking at the forest this way slightly resembles the neural network attention model of Vinyals et al. (2014). We can see that with increasing depth the attention of the forests shifts from the central systolic rectangle (Fig. 5.7a) to diastolic frames with more spread out spatial support (Fig. 5.7c). The later layers (Figs. 5.7d and 5.7e) further disperse the regions both spatially and temporally. Overall, the central image region is the most important (Fig. 5.7f).

5.5.2.2 Feature family relevance

The advantage of sampling features from the feature families is that we can easily find out which of these families are relevant for the similarity. The most important features per decision layer of the forest are listed in Fig. 5.8.

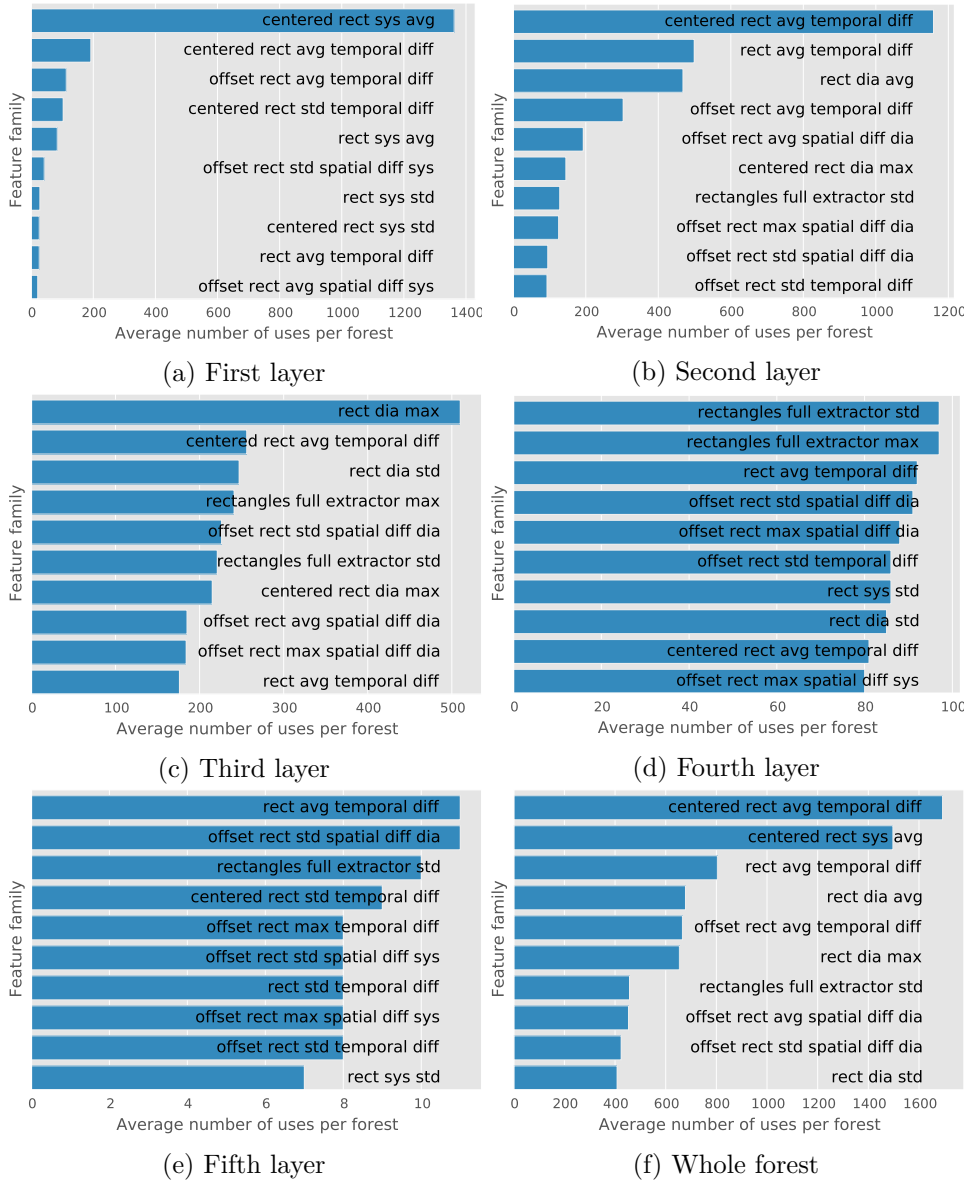


(a) First layer (b) Second layer

(c) Third layer (d) Fourth layer

(e) Fifth layer (f) Whole forest

Figure 5.7: Average feature importance at different depths of the forest (we show maximum intensity projections along the temporal and width axis). This is derived from the number of times a region was selected at the corresponding level of the trees. Brighter intensities mean more frequent use of particular voxel. At first, the central square region at the end systole is almost exclusively sampled. Then, the attention of the forest shifts towards more balanced systole-diastole comparison and more global spatial support. The third layer further increases the spatial support and seems to focus on end diastolic frames. Further levels become even more spatially and temporally spread. Note, we show maximum intensity projection across the temporal axis (left) and spatial as (right). We remind the reader, that our features use only diastolic and systolic frames (see Section 5.4.2.3), and therefore, the temporal axis has only two discrete instants.



(a) First layer (b) Second layer

(c) Third layer (d) Fourth layer

(e) Fifth layer (f) Whole forest

Figure 5.8: The most important feature families by decision layer depth for similarity neighbourhood derived from ejection fraction. We show how many times a family of features was selected by a forest. The first decisions are clearly dominated by mean intensities of the central region at end systole. The temporal changes of this region's mean intensity follow. See Section 5.4.3 for a detailed description of the used features.

In the first layer (see Fig. 5.8a), the most commonly used feature is the mean intensity of central rectangles at systole ("*centered rect sys avg*"). This choice makes perfect sense. For intensity-normalised and aligned images (as in our case) the central region image intensity clearly determines the proportion of myocardium and blood pool. In other words, well contracting heart will cover the region with myocardium and the mean intensity will be lower. The second most frequent decision in the first layer (Fig. 5.8a), the most frequent in the second layer (Fig. 5.8b), and the most popular feature family across the whole forest (Fig. 5.8f) is the intensity change in time of the central region ("*centered rect avg temporal diff*"). The average intensity changes in the central rectangle are indeed intuitive features to capture the ejection fraction similarity. Since the images are pose-normalised by aligning the epicardia at end diastole, the first splits strongly prefer the systolic measures and the diastolic ones are almost completely irrelevant.

The second layer of the forests (Fig. 5.8b) captures mainly the temporal change of mean intensity in the central rectangle and other randomly placed rectangles. But also, the first diastolic features start to appear and their relevance starts to grow (Fig. 5.8c). This is likely to capture complementary information to the systole-dominant decisions in the first layer.

The further layers (Figs. 5.8d and 5.8e) use combinations of a variety of feature families, capturing more and more of the fine-grained image details. The preference of one feature over another becomes less clear. Our analysis could differ for a much larger dataset.

5.6 Discussion and perspectives

We presented an efficient supervised method for content-based image retrieval for cardiac images based on the Neighbourhood approximating forest (NAF) using a set of generic spatio-temporal features. The retrieval similarity was learned from pairwise distance ground-truth derived from semantically meaningful parameters and the NAF automatically picked relevant features. The NAF algorithm has several properties particularly well-suited for content-based image retrieval. The distances to every heart in the database can be rapidly obtained.

The structure of the forest is fixed during the training phase. To insert a new image into the database, it is passed through the decisions and its reference number is recorded at the reached leaves. To compare the query image with images in the database, it is no longer necessary to access the pixel intensities of the stored images. This process is also computationally efficient.

Finally, peeking into the tree decisions can give us insights on which parts of the heart and which features are relevant for a particular goal. However, additional validation will be needed to ensure reliability of this method and its use in other (more interesting) similarity criteria than the EF.

5.6.1 Limitations

The forest-based method has a drawback for small datasets. The recursive division of the dataset creates nodes with only very few examples after a few splits. *E.g.*, in our case, using 96 cases is not much for a forest-based method with binary trees. Perfectly balanced trees recursively divide the dataset and rapidly leave less than 3 cases per node after just 5 decisions. As the number of leaves grows exponentially with tree depth, this reduces the chances of two truly similar images sharing a leaf. Therefore, for smaller datasets, no deep decisions can be taken.

This method is also limited by what the box features can capture. The box features can be seen as a way to aggregate local information. Some finer aspects of the hearts are less distinct for the forest to pick up the signal when using raw image intensity channels.

Since this method is based on image intensities, mixing datasets from other modalities or even CMR acquisition sequences is currently not possible. In particular, the use of the MESA dataset (Bild, 2002) with thousands of asymptomatic volunteers is a great resource to learn the differences between healthy and pathological hearts. In our case, with the intensity-based features, we would likely learn to discriminate the Gradient echo (GRE) from the SSFP sequences. A simple fix (such as the one for correction of cardiac shape biases by (Medrano-Gracia et al., 2013)) is not available.

5.6.2 Perspectives

To address the rapid drop in the number of cases per node, alternative learning structures are worth exploration. For example, Neighbourhood approximating jungles could be derived from the recently introduced Decision jungles (Shotton et al., 2013). In Decision jungles, directed acyclic graphs are used instead of the trees, occasionally merging similar nodes. This can help to keep higher number of points at the nodes, while, at the same time, keeping the decision structure more compact.

Retrieving hearts with similar ejection fractions is just the start. Training this method with more cases should help to improve relevance of the retrieved results, detect more subtle differences between the hearts, or experiment with different similarity criteria, *e.g.*, based on clinical outcome or clinical findings.

Adding additional channels enhancing the pericardial effusion (Appendix C), myocardial thinning, valvular jets, particular motion direction, or other motion-field derived features (such as the acceleration features of Kontschieder et al. (2014)) could help to better capture the appearance and motion information in the images. In addition, Kontschieder et al. (2015) recently proposed a method to train decision forests with backpropagation. This way the optimal convolutional filters and the forest structure could be trained simutaneously.

It is also important to explore how to combine NAFs with different similarity criteria together, so that mixed search criteria could be allowed, and to explore querying by ensembles of example images (*e.g.* by summing up the leaf counts). It

is likely that querying the datasets with these ensembles will enhance leaf counts for the common characteristics while the less common will be “averaged out”.

Chapter 6

Conclusions and perspectives

Over the past few years, several initiatives collecting cardiac images have appeared. These include the Cardiac atlas project (Fonseca et al., 2011) containing the DETERMINE subset (Kadish et al., 2009) we frequently used in this thesis, but also a study of asymptomatic hearts MESA (Bild, 2002). The UK biobank (Petersen et al., 2013) aims to image a large part of British population, and the European Cardiovascular Magnetic Resonance (EuroCMR) registry (Bruder et al., 2013) collecting imaging data from 57 centers in 15 countries. These already quite massive collections are here to grow and will cause impact on cardiac healthcare. We will need to find ways to automatically interpret the information contained within these databases and simplify search.

In this thesis we addressed some of the most important challenges of cardiac data organisation and information extraction from these datasets using machine learning.

6.1 Summary of the contributions

We addressed the following four issues. Here, we show them with our respective contributions and limitations of the proposed methods.

6.1.1 Estimating missing metadata from image content

Many of the DICOM tags are inherently noisy and cannot be realiably used. In this thesis we showed that instead of relying on DICOM, we can estimate some of the metadata from the image content — the cardiac acquisition plane information.

The main contribution of Chapter 2 is that a good cardiac view recogniser (reaching state of the art performance) can be efficiently trained end-to-end without designing features using a convolutional neural network. This is possible by fine-tuning parameters of a CNN previously trained for a large scale image recognition problem. We published our findings in Margeta et al. (2015c, 2014).

- We achieve state of the art performance in cardiac view recognition by learning an end-to-end system from raw image intensities using CNNs
- This is one of the first works to demonstrate the value of features extracted from medical images using CNNs, originally trained on a large scale visual object recognition dataset

- We show that fine-tuning parameters of the CNN pretrained on an object recognition dataset is a good strategy that helps to improve performance and speed up the network training

- We show that the CNNs can be applied to smaller datasets (a common problem in medical imaging) thanks to careful network initialisation and dataset augmentation, even when training the network from scratch

- We also reproduce the observation of Zhou et al. (2012) that the 3-dimensional orientations vectors that can be derived from DICOM tags can be used for cardiac view recognition (if the orientation tags are present)

Fine-tuning of a network is a great way to learn complex models from smaller size medical data and CNNs are powerful machine learning tools that have yet to fully propagate into medical imaging.

We trained the views only on a dataset of post-myocardial infarction hearts. These hearts certainly have some particularities as we showed in the thesis, nevertheless, their appearance and topology are still rather "normal". We have yet to test this method on severely pathological cases or less standard acquisitions (non-centred, oriented with atypical angles).

There are also many more cardiac views left to be recognised (Taylor and Bogaert, 2012) from other acquisition sequences. Such ground-truth has to be collected for all acquisition sequences with which the views are acquired, which leads to a combinatorial problem. Many of these combinations are underrepresented, and learning from very few examples remains a challenging problem.

In addition, the recognition of the views should not stop on describing the views. This work is a cornerstone for recognition of image modalities and MR acquisition protocols and for automatic generation of full image description as done recently for natural images using Recurrent neural networks (Vinyals et al., 2014; Karpathy et al., 2015; Donahue et al., 2014).

6.1.2 Segmentation of cardiac images

Segmentation of cardiac structures allows to index the databases using automatic measurements of tissue volumes and estimation of their masses. In Chapter 3 we propose to segment cardiac MR images with classification forests as in Shotton et al. (2008); Lempitsky et al. (2009); Geremia et al. (2011).

The main contribution of this part of the work is that voxel-wise segmentation of cardiac structures can be learnt from voxel-wise ground-truth.

- We proposed a forest-based segmentation method for LVs from 3D+t cardiac MR sequences and for LAs from 3D MR and CT images

- We introduced a novel two layered approach for standardisation of image intensities and inter-patient cardiac registration

- We designed spatio-temporal features for learning how to segment dynamic sequences
- We proposed to use other image channels such as vesselness, or distances to blood pool (which can be easily segmented) to segment the atrial images
- We also showed (Appendix A), how non-labelled data can help to segment the images in a semi-supervised way and how the best points for labelling can be proposed

This work led to our publications (Margeta et al., 2012, 2013), and we contributed the results from our algorithms to two benchmarking studies (Suinesiaputra et al., 2014b; Tobon-Gomez et al., 2015) for fair evaluation. Our methods are not the best performers in terms of segmentation accuracy. However, they shine in terms of how little knowledge about the problem was hardcoded into the algorithm which makes them quite flexible. The algorithms also get better with more examples. This makes our approaches much more ready to be applied to other cardiac modalities, views and tissues, and for the growing data.

While our methods do not use much prior information, some postprocessing to regularise the solution and large datasets to learn sufficiently robust features is essential. Extra steps such as data augmentation and learning discriminative image channels would likely help to further improve their performance.

6.1.3 Collection of ground-truth for describing the hearts with semantic attributes

In this thesis, we mainly focused on supervised machine learning approaches. This means that the algorithm learns from examples where ground-truth labels are required. The acquisition of these labels is at least as important as the acquisition of the images. Our solution to the annotation collection problem is a web-based crowdsourcing tool for collection of annotations in order to learn semantic descriptions of the hearts described in Chapter 4.

The disease itself is usually not a discrete (yes or no) flag and it manifests itself through several ways. Some of them can be described by clinicians in semantic terms, the cardiac attributes. There is discrepancy between this semantic description and the representation by the computer, making searching for hearts in cardiac databases using such description difficult. Thanks to the pairwise comparisons, less experience in interpretation of cardiac images is needed to correctly answer difficult questions.

- We design a web interface for crowdsourcing of such ground-truth.
- We propose to describe the cardiac images with semantic attributes such as hypertrophy, thinning, kineticity, dilation, or image quality from pairwise image comparisons.

- We show, how concepts from web-ranking can be used to learn these attributes.

Our crowd (3 in-house annotators) size is very limited and so is the amount of ground-truth currently collected and the conclusions we can draw. Annotator variability and quality is also an important aspect to study. When Health Insurance Portability and Accountability Act (HIPAA) compliance is assured, such a tool could be distributed among medical students to generate a rich resource for learning while contributing to the dataset at the same time. Having multiple annotations would also help us to assess how reliable the predictions are.

Presenting random combination of images to the annotators is not ideal as it redundantly wastes their time on simpler questions that the machine can already confidently predict. An active loop should be rather used where human input will be solicited predominantly for the least confident cases.

While we claim that any pathology could be described with semantic attributes, the spectral representations of the hearts cannot capture all of the aspects. Not only we restricted the research to the LVs, for small structures (such as thin walls of the myocardium) the spectral features become unstable due to the change of topology. Moreover, these features depend on good quality image segmentation.

We are also limited by our imagination of the attributes. Attribute discovery techniques are essential in the next steps.

Once more data is collected, learning better feature representations (*e.g.* (Kang et al., 2015; Zhang et al., 2014a)) for prediction of the attributes is crucial for improving the predictions.

6.1.4 Structuring the datasets with clinical similarity

Finally, Chapter 5 discussed how the decision forest structure has desirable properties to structure imaging datasets and can be used to learn similarity for CBIR using a decision forest-based method.

- We show how NAFs can be used to approximate clinical similarity with respect to ejection fraction or infarction location (Bleton et al., 2015) *i.e.* aspects capturing shape, motion and appearance similarity of cardiac sequences.
- We show simple ways to remove static background and estimate the main cardiac phases from the images
- We propose to reuse previously fine-tuned CNN features for cardiac landmark estimation and image alignment (Appendix B).
- We hint, how the pericardial effusion could be enhanced with simple filtering (Appendix C)

In theory, the NAFs can be trained with arbitrary image similarity ground-truth and used for fast exploratory queries within the large cardiac datasets. Using EF

itself does not lead to any benefit over using pure automatic segmentation and image measurements.

Our method needs more validation and larger datasets for improved reliability of the retrieved results and predictions on different similarity criteria. Also, additional image channels will be needed to enhance the relevant information.

While the returned results are not perfect, they are significantly better than just retrieving images at random and the retrieved images are relevant to the query. In the end, collection of real usage data in a feedback loop will be necessary to define which similarity criteria are actually useful and to personalise the retrieval results.

6.2 Perspectives

In this thesis we only scratched the surface of management of large cardiac databases with machine learning, the collection of annotations, or content-based image retrieval of similar cases. There are many challenges lying ahead.

6.2.1 Multimodal approaches

Different cardiac imaging modalities complement each other with the information contained therein. There is a need for multi-modality approaches (Galderisi et al., 2015). None of our tools is modality specific, and the same algorithms could be almost directly applied to these modalities.

In this thesis, we used mainly SAX images, acquired with the SSFP MR protocol. This is clearly not sufficient. Other cardiac views can offer deep insights into different aspects of the heart that are not seen on short axis slices. The Delayed enhancement MRI (DE-MRI) acquisition protocol captures the extent of myocardial infarction, the tagged imaging sequences help to better estimate the motion, the phase contrast imaging captures the blood flow through the valves and in the arteries. New modalities and protocols appear. T1 imaging at 3T magnetic fields can be used to reliably detect and quantify myocardial fibrosis without the need for contrast agent (Kali et al., 2014) (*e.g.* to image infarcted regions in hypertrophic cardiomyopathies, dilated cardiomyopathies, or aortic stenoses). Gradient echo $T2^*$ sequences (Anderson, 2001) capture myocardial iron overload and allow early detection of ventricular dysfunction, before it develops. Other modalities such as CT and US can bring yet another angle into study of cardiac anatomy and motion.

6.2.2 Growing data

The study of the EuroCMR registry (Bruder et al., 2009) scanned 27,309 patients across 57 sites in 15 European Countries according to standard protocol (Kramer et al., 2008). In 61.8% of all subjects, the study found that CMR had significant impact on diagnosis and therapeutic consequences (change in medication, invasive procedure, hospital discharge or admission). About 8.7% of the scans revealed new diagnoses not suspected before.

But we have yet to establish the benefits of CMR on the outcomes. More people will likely be scanned and more data will be generated. Wearable computing has already started to record real-time health information. It will soon become necessary to search through this data with automatic software tools.

With these humongous databases and heterogeneuous data sources, automatic cleaning of the metadata directly from image content and indexing of the databases for fast access to relevant studies become even more relevant than before.

6.2.3 Generating more data

We still face the challenges of expensive data acquisition, patient privacy in data collection and distribution and missing ground-truth. In the meanwhile, if we can generate massive synthetic but realistic data and images using biophysical models, we shall be able to train richer models and only then tune them to the real data problem. This is known as transfer learning (Pan and Yang, 2010) and is one of the main goals of the ERC MedYMA project.

It is significantly cheaper to generate more synthetic data than to acquire and annotate them. There is no need for patients' consent or restriction on data sharing, the underlying parameters used to generate the images (the ground-truth) are known. These *in-sillico* approaches can generate large quantities of images with known ground-truth labels, facilitating training algorithms to reverse the generative process and obtain the labels back from the image data. High quality generators of synthetic cardiac image generators already exist for a wide range of modalities (Alessandrini et al., 2015; Prakosa et al., 2013; De Craene et al., 2014; Tobon-Gomez et al., 2011). Whole platforms on the web are dedicated to the multi-modality medical image simulation to democratise this process (Glatard et al., 2013).

Similarly to how synthetic depth data made real-time pose estimation from depth cameras possible (Shotton et al., 2011), robust detectors were trained on synthesised fluoroscopic images (Heimann et al., 2014). Geremia et al. (2013) learnt to estimate tumour cell density estimation on clinical images by training models on simulated multi-modal MR images using biophysical models growing synthetic cerebral tumours at various positions in the brain. Simulated cardiac images (Prakosa et al., 2013) were used to inverse cardiac electrophysiology (Prakosa et al., 2014) and the models transferred to real images. The opportunities are endless.

6.2.4 Data augmentation

We showed that even a simpler way to generate synthetic but very realistic training examples and to improve performance of the machine learning system is to modify the already existing images such that the labels are preserved. This is called data augmentation or data jittering. The augmentation is a cheap and efficient way to cover some of the dataset variability (this can help if features are not invariant to these changes). Simple image transformations (Decoste and Schölkopf, 2002) and

intensity manipulations are often performed. Image acquisition artefacts such as vignetting or barrel distortion (Wu et al., 2015) or realistic background noise (Hannun et al., 2014) are added.

In general, the problem of augmenting medical images is more difficult as care must be taken only to change the image content and not the associated label (*e.g.* merely changing the image size by an isotropic scale could make a normal heart look like a dilated one) but not impossible. Adding artificial noise, bias fields or image artefacts will increase the robustness of our methods.

We will need machine learning models that have the capacity to learn from these massive data. Ideally, these will be models into which the training data can be iteratively streamed in batches (such as for neural networks trained with stochastic gradient descent, or in bootstrapped forests) so that it will be quite simple to train with new data and update their parameters on the fly. The CNNs as we used for recognition of cardiac planes are a good fit for the growing multi-modal data problem.

6.2.5 Collecting labels through crowdsourcing and gamification

Now, the challenge is not only to keep acquiring and aggregating the data but also to design annotation tools for our algorithms to learn. The solutions are multiple. One is to generate synthetic data with known ground-truth, another one is to ask the clinicians or the "crowd" for them. Many complex questions can be reduced to simpler ones where the crowd can competently answer them.

Pair-wise image annotations we used are not the only way. Large companies like Google are using the crowdsourcing to collect ground-truth for its Street view service house number recognition or to translate books. Galaxy Zoo teaches people astronomy while collecting precious annotations for automated description of galaxies. Duolingo teaches people foreign languages and the users help them to train automatic translation systems in return, gamifying (von Ahn and Dabbish, 2004) the ground-truth collection.

There are many challenges to make these systems more engaging and how to use the often not perfect ground-truth.

6.2.6 Moving from diagnosis to prognosis

In the end, diagnosis is not the end. Finding evidence about "what is likely to happen" should shape clinical practice. Follow-up evidence is needed for cardiac imaging to improve quality of care and the outcomes. Shift in focus from diagnosis to prognosis and outcomes (Timmis et al., 2015) should happen with patients reporting the outcome measures and their quality of life. If prognosis is to replace diagnosis (Croft et al., 2015), this will be impossible without automating the whole process of data management and effectively retrieving the supporting data based on the outcomes and benefits to the patients.

Appendix A

Distance approximating forests for sparse label propagation

Contents

Overview

In this appendix, we present a semi-supervised label propagation algorithm based on the density decision forest to approximate geodesic distances in feature space to labelled points. The algorithm hierarchically partitions the data (both labelled and unlabelled) by a simple unsupervised splitting criterion in order to approximate feature space density. Using the density approximation, we then propagate labels from the nearest annotated points in the dataset to all unlabelled points, using a fast-marching-like propagation strategy.

This allows us to exploit the unlabelled data and use significantly less manual input. Moreover, the obtained distances to the labels can help us to suggest new points to be labelled in an active learning loop. We test this learning method on left ventricle segmentation from CMR.

A.1 Introduction

Large quantities of medical images have been recently made available. Manual inputs (*e.g.* labelling of each voxel in an image or assigning a pathology for each image in the database) can be a tedious, time consuming and expensive process.

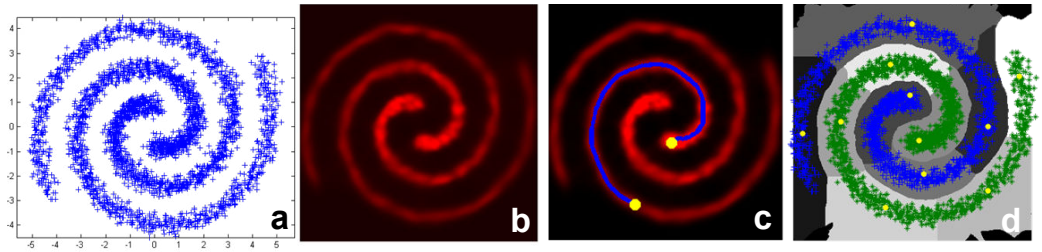


Figure A.1: **(a)** Unlabelled points from a 2D feature space. **(b)** Density estimation with density forests. **(c)** Fast marching shortest path between two points prefers traversing the dense regions. **(d)** Voronoi clustering and label propagation from the nearest sparse labels.

On the other hand, unlabelled data have always been rather inexpensive to acquire. We can represent each image (or voxel in an image) as a point in an N-dimensional feature space. Based on the high density assumption (Singh and Nowak, 2008): if the data points are lying on a manifold, the point densities can help to propagate correct labels to unknown points by finding the closest labels along the density.

It has been illustrated by Criminisi et al. (2011b) that data densities can be efficiently estimated using decision forests. The estimated densities are then used in semi-supervised learning to deal with a lack of labelled data points. Moving along the direction of high density should be considered easier than crossing low density regions (see Fig. A.1). This can be done very efficiently in low dimensional spaces, where the point density is estimated on a Cartesian grid. Algorithms, such as fast marching (Sethian, 1996), then rapidly compute the shortest paths from all labelled points. However, it is not very practical to estimate the density in high-dimensional feature spaces as the size of these grids would rapidly explode with the dimensionality of the feature space. On the other hand, graph-based label propagation approaches are capable to better deal with higher-dimensional spaces, but they suffer for large numbers of points.

A.2 Previous work

Leistner et al. (2009) have proposed a semi-supervised decision forest algorithm, where forest training and data relabelling (with the new predictions) is performed iteratively in a simulated annealing process. This requires computationally demanding retraining at each iteration. Moreover, over the iterations, incorrect labels might leak if some points are classified incorrectly. In general, it is difficult to completely avoid such a leak in semi-supervised learning. Here, we would like to avoid the iterative learning and also to find out when we lose confidence in label assignment and when we are more likely to make an error. Our approach is the most similar to Bachmann (2006) and the semi-supervised forests (Criminisi et al., 2011b). Two other related works use decision forests for metric learning (Xiong et al., 2012) and to approximate (Konukoglu et al., 2012) distance between data points. These

approaches, however, require pairwise point distances as the input.

In Bachmann (2006), an efficient forest-based method called Vantage point forest is used to partition the large high-dimensional data space of pixels from multi-spectral satellite images. They avoid exhaustive point-to-point distance calculations for a subsequent dimensionality reduction algorithm. In their approach, each tree sees a different feature subset and is associated to one landmark point in the image. The authors do not exploit the density direction that can be obtained from a density forest.

In our approach, we propose to avoid the need for feature space rasterisation and use the trees to jointly estimate local point densities. We partition the initial space into several smaller subsets and find approximate geodesic distances from every unlabelled point to the closest label. The tree ensemble then averages the tree distances for a smoother approximation.

In this appendix we propose

1. A method for geodesic distance estimation in the feature space to find candidates for relabelling.

2. A practical implementation of density forests for learning with sparsely labelled data.

A.3 Distance approximating forests

Although a semi-supervised problem like the one illustrated in Fig. A.1 could be solved by the above mentioned methods or traditional manifold learning algorithms, it becomes much more difficult with increasing dimension of the feature space and size of the training set.

By partitioning the data using the "divide and conquer" strategy of the forests, we recursively divide the whole dataset into much smaller chunks of data and work in parallel for each tree. We also use different feature space dimensions at each node for increased computational efficiency while adapting to the the underlying data density.

The training process is very simple. We start at the root node and recursively partition the data S until the stopping criteria (maximum depth, minimum number of points in a leaf) are satisfied. At each node, we partition the data using ρ configurations of binary split functions $\theta_n(x) = \tau_n(\psi_n(\Phi_n(x)))$. These consist of a linear projector $\psi_n(x) = W^T x$ and a threshold $\tau_n(x) = x < T$. Each split function divides the points into two disjoint partitions $V = \{0, 1\}$.

Out of the ρ choices we then select the split that maximizes the total split quality I. In our case, we use a fully unsupervised the differential information gain (Eq. (A.1)) as the cost function, encouraging splits that can be represented with two compact multivariate Gaussians. Here, $|\Lambda(S)|$ is the determinant of the covariance matrix of the Gaussian distribution.

$$I = \log|\Lambda(S_j)| - \sum_{v \in V} \frac{|S_j^v|}{|S_j|} \log|\Lambda(S_j^v)| \tag{A.1}$$

Although the method that primarily inspired this work (Criminisi et al., 2011b) uses a combination of supervised (information gain) and unsupervised criteria, we use only the unsupervised component. For a small number of labelled points the estimation of the supervised cost would be unreliable.

To deal with the high-dimensional feature spaces, at each node n we select a random subset of features with feature selector $\Phi_n(x)$ instead of working with complete feature vectors. This feature subset can be different for each node and the features can be also evaluated "on the fly". Once the forest is trained and the dataset is partitioned into the leaves, we need to find a way to propagate the labels.

A.3.1 Finding shortest paths to the labels

Often, point neighbourhoods for label propagation are approximated using k-nearest Euclidean neighbours. These are then used for graph construction. Small point neighbourhoods are susceptible to noise, whereas large ones increase connectivity of the graph and the search space. In addition, location of the points within the feature space can matter for the optimal choice of the neigbourhood. How to choose the parameter k? We aim to completely avoid the need for it (although we exchange it for the forest depth parameter).

In our case, each tree hierarchically partitions the node data distributions into compact multivariate Gaussian distributions. When the tree structure is fixed, the multivariate Gaussian parameters are stored at each node. We use them to define the local point neighbourhoods within the nodes and to approximate geodesic distances to the nearest labelled points. As every tree is trained with some randomness, the trees differ from each other, and the ensemble neighbourhood of each point can be quite different from the Gaussian shape. Training a distance approximating tree is summarised in Algorithm 3.

A.3.1.1 Leaf covariance and medoid

For every leaf node n we capture the shape and the position of the leaf's point distribution with a covariance matrix Λ_n and mean μ_n. These are computed on the feature space subset (selected with $\Phi_n(x)$ of its parent). In addition, we find leaf's medoid m_n as the closest point to the leaf's centre using the squared Mahalanobis distance[1]:

$$m_n = \arg\min_{p \in n} (\Phi(p) - \mu_n)^T \Lambda_n (\Phi(p) - \mu_n) \tag{A.2}$$

This point will represent the leaf's position in the feature space. We use the medoid (a true dataset point) instead of the leaf's mean (a virtual point), since μ_n

[1] We use the name Mahalanobis distance for this generalized squared interpoint distance.

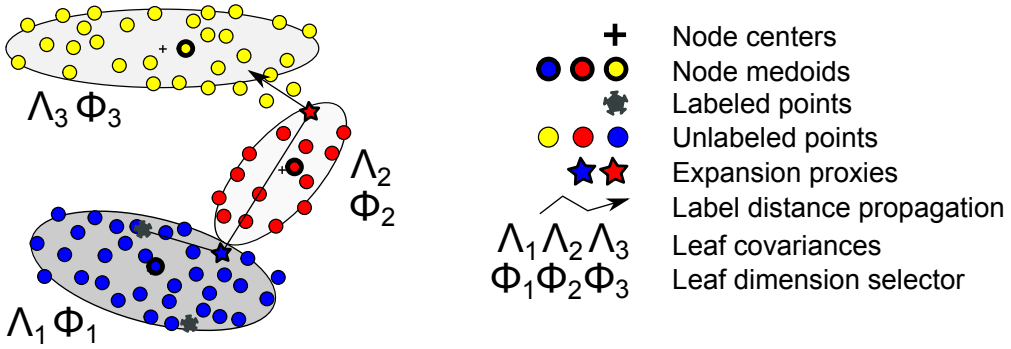


Figure A.2: Distances between points are computed as Mahalanobis distance. Note that the covariances are not necessarily calculated in the same feature subspace. We consider three distances: Distances between labelled points of each class and all unlabelled points within the same leaf, distances between node medoids, and distances to the approximately nearest points (expansion proxies) of the neighbouring nodes.

exists only in the leaf's feature subspace. This permits us to evaluate the complete feature vector for the medoid point and to allow leaf crosstalk.

A.3.1.2 Initialisation of the propagation

To start, we assign a vector of distances to each class label (two scalars for a binary classification problem per tree) to each unlabelled point in the training set. We initialise the values to infinity (*i.e.* the point not reachable). We start at the labelled points with zero distance and propagate the distances to all the unlabelled points within the same leaf, only then we expand the distances towards points in the neighbouring unlabelled leaves until all leaves and all points have an estimate of distance. Two things can happen for the unlabelled points: when the unlabelled point is contained within the same leaf as the class labels, and when not and the label distance has to be propagated from neighbouring leaves.

A.3.1.3 Intra-leaf distance to label propagation

For points in leaves containing the class label, the solution is very simple. Mahalanobis distance to the nearest labelled point is assigned to all unlabelled points of the leaf:

$$D_n(x_i, x_j) = (\Phi_n(x_i) - \Phi_n(x_j))^T \Lambda_n^{-1} (\Phi_n(x_i) - \Phi_n(x_j)) \tag{A.3}$$

A.3.1.4 Out of leaf distance to label propagation

Often the data partitioning happens in such a way that no labelled points are present in the leaf. The distances now need to be propagated to the other unlabelled leaves. The symmetric Mahalanobis distance between points in leaves n and m can be defined as follows (Criminisi et al., 2011b):

$$D_{n,m}(x_i, x_j) = 0.5\,(D_n(x_i, x_j) + D_m(x_i, x_j)) \tag{A.4}$$

This is similar to Jeffries-Matsushita distance (Bachmann, 2006) used for cluster similarity. Note that when n equals to m this becomes equivalent to Eq. (A.3). However, as in the graph-based methods, pairwise point distances to find the best inter-leaf point-to-point combination could be prohibitively expensive for higher number of points to be computed exhaustively. We reduced this problem to finding only the nearest node medoid and then a point of expansion. This can be represented with a much smaller pairwise medoid distance matrix W_t for each tree t. Once computed, this matrix stays fixed.

A.3.1.5 Finding the nearest point from another node

Now, we will discuss how to use this matrix to find the nearest leaves and propagate the label distances. First, the best labelled-unlabelled medoid combination for expansion is found as the minimum pairwise Mahalanobis distance of the medoids. We call leaves containing these two medoids a source (fixed) and a target (expansion) leaf. From the source leaf, where each point is already assigned a distance to the label, we expand the distances to the target leaf.

To reduce another exhaustive pairwise distance problem (*i.e.*, for each point in the target leaf find the closest point in the source leaf), we first find a single proxy point within the target leaf. This proxy represents the connection between the two leaves. Ideally, this should be a point on the boundary of the source leaf distribution and at the same time close enough to the target leaf distribution (see Fig. A.2). We choose it to be a point from the source leaf closest to the target leaf medoid. Now, the symmetric Mahalanobis distances between this point to all points in the target leaf can be calculated. These distances are summed withthe "distance to the label" value of the proxy point. The sum of the two is then stored for each target leaf point.

A.3.1.6 The ensemble decision

This distance propagation procedure is independent for each class (for classification) and each tree. Ultimately, the minimum distances to each class label are assigned to every point in the training set. The final label of each data point is then assigned to be the class with minimal average distance across all forests.

A.4 Results and discussion

A.4.1 Synthetic classification example

We first tested our algorithm on two-class classification problem of a synthetic 3D spirals dataset. We generated two spirals consisting of a total of 1000 points.

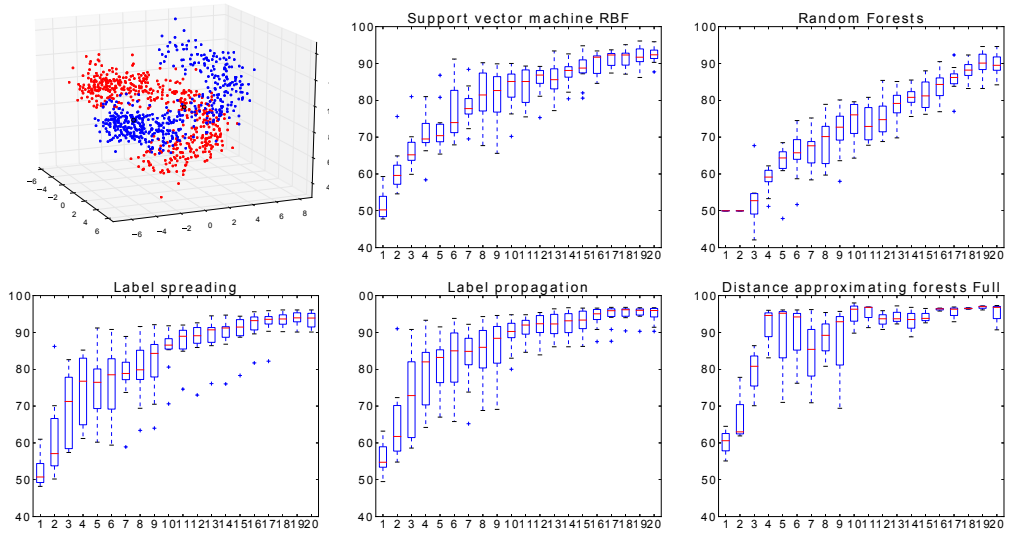


Figure A.3: Comparison of classification accuracy with respect to number of labelled points for various learning algorithms on a 3D spirals dataset. Labelled points are chosen from the dataset at random. Our distance approximating forest reaches higher accuracies significantly faster.

Out of these points we randomly selected 15 points (5.7% of the training data) to be labelled and ran our semi-supervised forest. We trained 300 trees with maximal depth 4 and tested 50 split function configurations per node.

In order to validate the inter-leaf distance propagation we restrict the number of tested dimensions at each node to two (with all three dimensions we were able to rapidly achieve >90% average classification accuracy). Similarly, denser distribution also improves the classification accuracy as the leaf covariances are estimated much more robustly, however, this hides real problems with noisy points. Also when labelled points are placed more uniformly, much better coverage can be obtained. The setup from Fig. A.4 misclassifies only 18.3% of the points.

It is easy to see that the misclassified points in some regions (marked with ellipses on (b) and (c)) have much larger distance to the label than others. This can be exploited in an active learning strategy to propose new points to be labelled.

A.4.2 Towards interactive cardiac segmentation

Traditionally, decision forests have already been used in medical imaging for quite a large range of segmentation problems. This often requires large databases rich with segmentation information. We will show here that our semi-supervised forests can be used when the data is sparsely labelled, *e.g.* in an interactive segmentation.

Prior to the segmentation, we train a fully unsupervised forest using all voxels of the image to be segmented according to the differential entropy splitting cost as described above. This helps us to decompose the image into several configurations

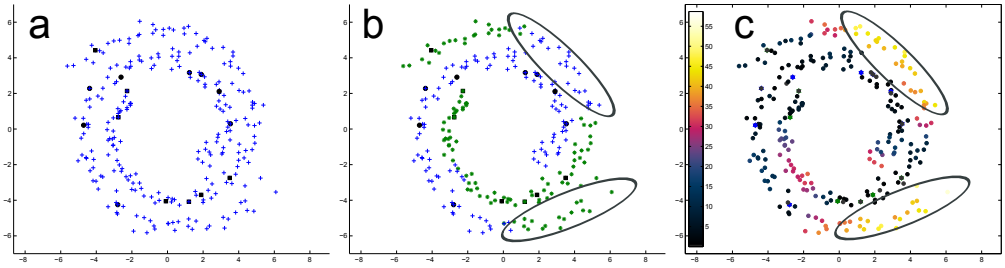


Figure A.4: **(a)** Two 3D spirals in an orthogonal 2D view and labelled points (black squares and circles). **(b)** Labels assigned according to the nearest label. **(c)** Minimum distances to nearest labelled points (encoded by colour) clearly show that the misclassified regions are far from the labels and would be good candidates for relabelling.

of meaningful partitions. Which also means that similar feature measurements are clustered together and can benefit from the geodesic label propagation. Then medoids and their pairwise distance matrices can be precomputed and fixed for each tree in an off-line step. The much faster propagation of the distances to the labels can be then done during the interaction.

A.4.2.1 Features

We use a combination of very simple features for each voxel - image intensity (at two scales of a Gaussian pyramid from the current frame and the frame from the middle of the cardiac sequence). Since we apply the semi-supervision on a single image, there is no need for registration to the reference pose and we use also voxel coordinates as features to spatially regularise the segmentation. For a 2D case, this results in a feature vector of dimension 6.

A.4.2.2 Stabilisation of the covariance matrices

Real MR images pose a problem for the differential entropy. Homogeneous background (air) forms intensity clusters with little variation and the covariance matrices might not be invertible. Therefore, similarly to Levenberg-Marquardt optimisation algorithm (Marquardt, 1963), we regularise our covariance matrices with a scaled identity, *i.e.*, $\Lambda_n^r \leftarrow \Lambda_n + \gamma I$ both for the computation of differential information gain and for the Mahalanobis distance computation. This significantly improves the numerical stability while did not seem to compromise the performance for our experiments. In our experiments we use $\gamma = 0.1$.

The output of our segmentation algorithm is still soft. However, distances to labels are used instead of the label posterior probabilities. The forests were not trained for this particular segmentation problem rather than to decompose the image into something meaningful. Hence, the structure already learned could be used to extend the segmentation to more classes or even regression very easily without retraining - only label propagation step is needed.

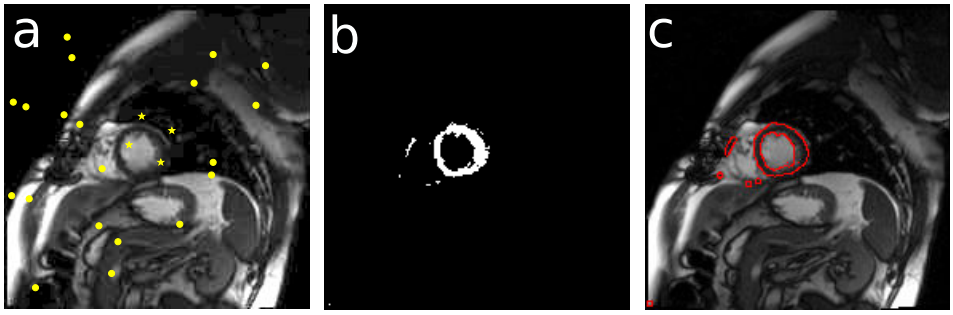


Figure A.5: **(a)** We sparsely labelled 4 points on the myocardium and 20 points from the background. **(b)** Binary mask of points with shorter distance to the nearest myocardial label than to the background label. We used 17 trees, $\rho = 10$, and dimension subset 2. **(c)** Overlay of the final segmentation (Canny edges).

A.5 Conclusions

We proposed a semi-supervised learning algorithm that uses unsupervised forest training to estimate feature space density and to later approximate geodesic distances to the labeled points in this space. Although distances from individual trees are not perfect, they already create a signal of the distance. The distances are then averaged to get the final estimate. As a result we associate the shortest distance to the nearest labelled point of each class to all points in the training dataset. For a two class segmentation problem this means that a 2D vector is associated to each voxel in the image.

We observed that this technique works even if we do a lower-dimensional feature subspace selection, when the subspace is different at each node. This permits to speed up the distance calculation and gives us the possibility to deal with very high-dimensional datasets or even datasets where features are evaluated "on the fly." The decision forests help us to partition the feature space into several smaller regions, each represented by a point subspace, covariance, and a medoid. We can therefore replace expensive pairwise point-to-point with distances between the leaves as a rough estimate of the connectivity. One has to be careful with deep trees as the number of leaves (and hence medoids) grows exponentially with depth (up to $2^{depth-1}$).

We tested this algorithm on a synthetic dataset and for an interactive segmentation of left ventricles from CMR. As the structure of the trees and the medoid distance matrix can be fixed we can do a rapid update of the distances once new labels are added. As the forests are inherently parallel this can be further sped up.

A.5.1 Perspectives

Assigning distances to labels for each points enables us to actively guide point selection process in an active learning loop. Points far from any label should be considered to be labelled first. It is important to note that this approach is not

limited to segmentation and classification problems. In our experiments, we do not yet fully exploit the feature subspace selection for high-dimensional datasets, our maximal feature space dimension was 6. Nevertheless, using the original density forests for the nearest label search using a 6-dimensional Cartesian grid with fast marching would become an extremely memory intensive problem.

The proposed pairwise medoid matrix for inter-leaf label propagation is not the only possibility. At the moment we are completely ignoring the hierarchical structure for the calculation. This might be an interesting area of further exploration. Finally, the sparse labels have very strong influence on the approximated distances. The stability of the algorithm to incorrect labels should be investigated.

Algorithm 3: Training of a single distance approximating tree for one label type.

```
tree ← DensityTree(points, ρ, dim, stopCrit)
nodes, leaves ← tree.nodes, tree.leaves

d ←InfinityArray(points.size)
foreach leaf in tree.leaves do
    medoids[leaf.index] ← CalculateMedoid(leaf.points)
tree.W ← PairwiseDistance(medoids)

foreach leaf in tree.leaves do
    d_l ←InfinityArray(leaf.points.size)

    foreach srcPt in leaf.labelledPoints do
        d_l ←Min(d_l, PropagateDistance(srcPt, leaf.points))

    d[leaf.partition] ←Min(d[leaf.partition], d_l)

fixedMedoids, freeMedoids ←
FindFixedAndFreeMedoids(medoids, tree.W)

while not freeMedoids.Empty() do
    sourceIdx, expansionIdx ← FindMinDistancePair(fixedMedoids,
    freeMedoids, tree.W)

    sourceLeaf ← leaves[sourceIdx]
    expansionLeaf ← leaves[expansionIdx]
    expansionMedoid ← medoids[expansionLeaf.idx]

    proxyPt, proxyPtDist ←
    FindClosestPoint(expansionMedoid, sourceLeaf.points)

    d[expansionLeaf.partition] ←
    proxyPtDist+PropagateDistance(proxyPt, expansionLeaf.points)

    fixedMedoids.Append(expansionMedoid)
    freeMedoids.Remove(expansionMedoid)
```

Appendix B

Regressing cardiac landmarks from CNN features for image alignment

Contents

Overview

To compare relevant cardiac territories across different subjects and cardiac shape variations we must first establish correspondences between the hearts. This is especially necessary since the features we use throughout the thesis (Section 5.4.3) are not invariant even to simple rotations and translations. We show how features from the CNN previously fine-tuned for recognition of acquisition planes (Chapter 2) can be reused for cardiac landmark estimation, and to align the cardiac sequences.

B.1 Introduction

Image registration is one of the obvious choices for registration of cardiac images. Robust inter-subject cardiac alignment via traditional image registration methods is in general quite difficult due to the presence of trabeculations, papillary muscles, extra-cardiac thoracic structures, and MRI acquisition artefacts (Klinke et al., 2013; Naehle et al., 2011; Saremi et al., 2008) frequently appearing on cardiac images. Ou et al. (2012) address the inter-patient cardiac registration by first masking the hearts with manual segmentation. Masking the heart prior to the registration reduces the effect of irrelevant structures but requires cardiac segmentation instead. In Chapter 3, we trained two layers of decision forests, the first one roughly segmented the heart and returned soft probability maps which we used for registering images between different patients. The thorax stripping presented in Section 5.4.2.1 helps

to automatically remove the outer structures by fading the irrelevant non-moving background information without the detailed segmentation.

Even a perfect masking of the images with segmentation does not guarantee accurate correspondences between the cardiac images (except for rigid alignment) using non-linear registration as ambiguities exists for large deformations. Lombaert et al. (2012) suggest that better correspondences between cardiac images for registration can be obtained by registering spectral representations of the images instead of registering the image intensity channels directly. Nevertheless, image registration methods, in general, result in little control on which parts should and which parts should not matter for the alignment.

We argue that landmark-based inter-patient approaches are a more flexible way to define correspondences between the hearts. The advantage of using landmark-based methods for alignment of the cardiac images is the flexibility to rapidly change the correspondence model (rigid, similarity, affine, or piecewise affine) and the reference cardiac structure on the fly.

B.2 Aligning with landmarks and regions of interest

In practice, cardiac images are commonly aligned by automatically or manually placing a set of landmarks. These often include the junction points of the left ventricle, or ventricular barycentres (Wang et al., 2015; Bai et al., 2015). The CAP (Fonseca et al., 2011) provides an interactive tool for fitting much denser left ventricular meshes to the cardiac images. An alternative is drawing tight contours Afshin et al. (2012b) or rectangles (Afshin et al., 2012a) that encapsulate the left ventricle.

For large scale analysis it is desirable to automate this process but still keep the possibility of simple manual corrections for failure and to establish correspondences also for the right ventricle and possibly other structures. It has been shown (*e.g.* by Zhou et al. (2012)), that separate detectors can be trained for the main cardiac landmarks. In theory, it is possible to train a detector for any landmark. This requires to train and test a large number of detectors, to suppress multiple maxima and regularize the final solution.

In this appendix, we propose an automated landmark detection method inspired by the technique of Ren et al. (2014) for face alignment in face recognition. Their method is quite robust to occlusion in natural images, *e.g.*, hair across the eyes and working under varying light conditions. The MR acquisition artefacts can be also modelled as occlusions. The acquisition intensity differences and bias fields are analogous to the variation of the lighting conditions. We therefore approach the alignment of hearts with this technique.

B.3 Definition of cardiac landmarks

Which landmarks to choose for our model and how to consistently establish the correspondences even for the free walls? To compare corresponding cardiac territories

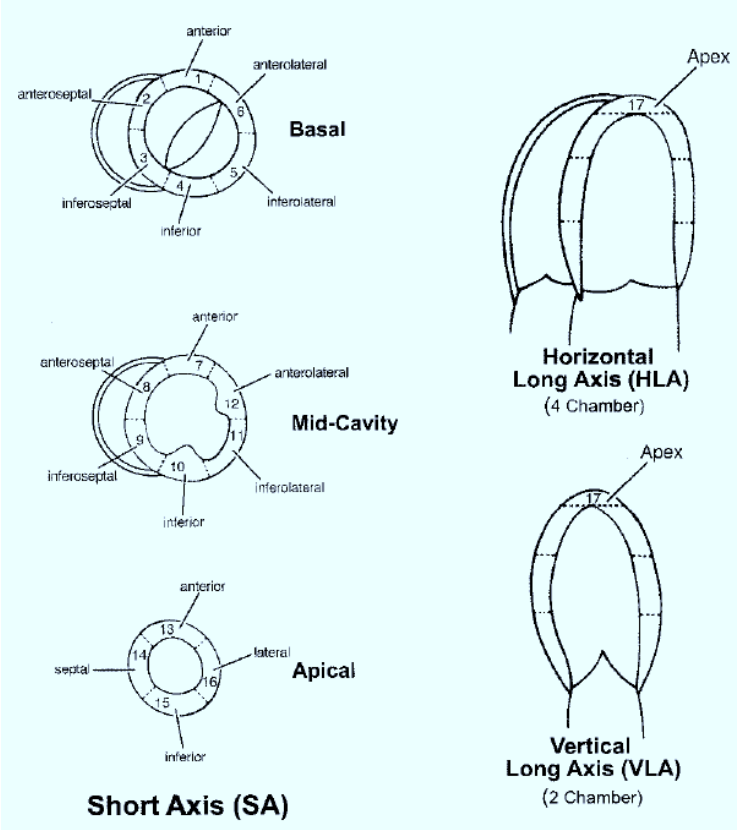


Figure B.1: The standard American heart association 17-zone model (Cerqueira et al., 2002) used for comparative regional analysis.

in a clinical setting, Cerqueira et al. (2002) proposed a now widely adopted 17-zone model of the left ventricle (see Fig. B.1).

Using the left-right ventricle attachment (junction) points with cavity centre for alignment is a solid start for comparative cardiac image analysis as the positions of these are relatively unambiguous. Using more landmarks gives us a more fine-grained control over the alignment model or the choice of the reference cardiac structure. With landmarks we can register anything from local regions (*e.g.* myocardial septa), through parts (*e.g.* left ventricular epicardium) or global hearts with transformation models varying from rigid transformation through similarity and affine transformation, towards piecewise affine and non-rigid models.

We define 68 landmarks in total for mid-cavity slice only. Our landmark (see Figure B.2) model is similar to the frequently used AHA model of Cerqueira et al. (2002). We obtain the landmarks by dividing the manually drawn cardiac contours (LV epicardium, LV endocardium, and right ventricle (RV) endocardium) into several segments. These segments start at the well-defined LV-RV junction points. The curve segments between these two points are uniformly resampled and the landmarks are placed there. Once we have defined a set of training landmark configurations we can train a regressor to place them automatically.

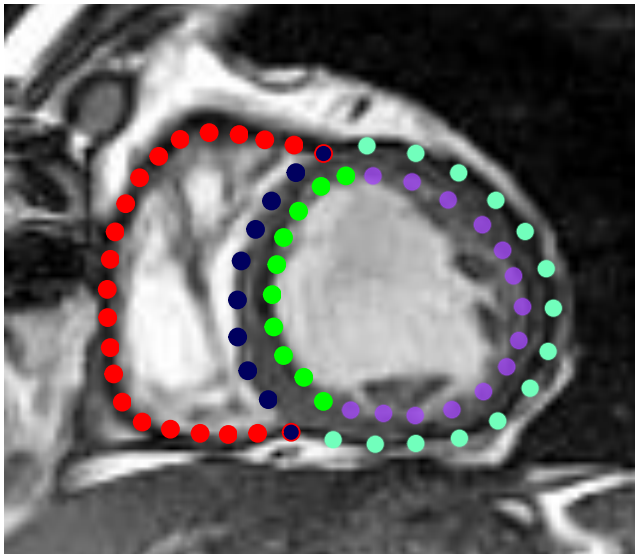


Figure B.2: Our cardiac point distribution model. In total we place 68 landmarks on a mid-cavity slice. We represent the septum with 20 uniformly distributed landmarks, 10 on each side. Two of these 20 points are fixed to on the LV-RV junctions. Then 14 points are uniformly sampled along the left ventricular free wall epicardium. Likewise we uniformly sample another 14 for the endocardium. The right ventricular free wall is represented with 20 landmarks.

B.4 Cascaded shape regression

Posing this problem as discriminative regression we predict the joint positions of all landmarks directly from the images, as was done, for example, by Donner et al. (2013) for estimation of landmarks from X-ray images. The cardiac shape model at iteration t is represented as a vector of flattened landmark coordinates $s^t = [x_1, y_1, x_2, y_2, \ldots, x_{68}, y_{68}]^T$.

Initial shape regression

As we saw earlier in Chapter 2, the heart is often located at the centre of the image (Figure 2.6). To initialise the shape s^0 we first extract CNN features from the central image crops. To describe the image slices we use features from the last fully connected layer ($fc7$) of CardioViewNet network (Section 2.6.5). This CNN was fine-tuned for CMR acquisition plane recognition but not for landmark regression. This gives us a 4096-dimensional feature vector Φ^{cnn} as a byproduct of classifying the image acquisition planes at no additional cost.

Using the extracted features, we can then train a simple linear regressor to estimate the initial shape by linearly projecting the CNN features:

$$s^0 = W^0 \Phi^{cnn} \tag{B.1}$$

Where W^0 is estimated using ridge regression (Golub et al., 1979) as:

$$W^0 = argmin_W \sum_i \|W\Phi^{cnn}(I_i) - s_i\| + \lambda\|W\| \tag{B.2}$$

The regularisation of the weight matrix W^t is necessary for the method to predict realistic cardiac shapes at its output. The best parameter λ was chosen by cross-validation.

Refinement of the prediction with CNN features

The initial estimate (Figure B.3) can then be refined with cascaded regression. Instead of the CNN features, the cascaded regression uses shape-indexed features $\Phi^t(I, s^{t-1})$, *i.e.*, features evaluated from patches centred at landmark positions of the current shape estimate. A new ridge regressor is then trained to predict the shape update jointly from all features. The shape update can be represented as:

$$s^t = s^{t-1} + W^t\Phi^t(I, s^{t-1}) \tag{B.3}$$

At each stage of the cascade a shape estimation matrix W^t is trained via ridge regression. We use local binary features (Ren et al., 2014) for the shape-indexed features.

$$W^t = argmin_W \sum_i \|s_i^{t-1} + W\Phi^t(I_i, s_i^{t-1}) - s_i\| + \lambda\|W\| \tag{B.4}$$

Dataset augmentation

In practice, some differences in hearts' positions and orientations between the acquisitions exist. These are quite arbitrary and depend on the radiologist or the acquisition machine settings. Nevertheless, they do not change the underlying information content of the images.

To better capture variability of the dataset and to increase the training set size, we generate new images indistinguishable from the real ones by rotating and translating the original image slices. Even small changes to the image scale can still generate images of hearts of realistic sizes. We therefore use the original training set and augment it (both the images and the landmarks) with additional scale changes, rotations, and small translations.

The parameters of these transformations were sampled from three Gaussian distributions. Scaling factors were sampled from $N(\mu = 1, \sigma^2 = 0.0025)$, rotation angles (in radians) from $N(\mu = 0, \sigma^2 = 0.0025)$, and translations offsets from a bivariate Gaussian $N_2(\mu = 0, \sigma_x^2 = 4, \sigma_y^2 = 4)$. In total, we generated 10,000 images out of 100 training images.

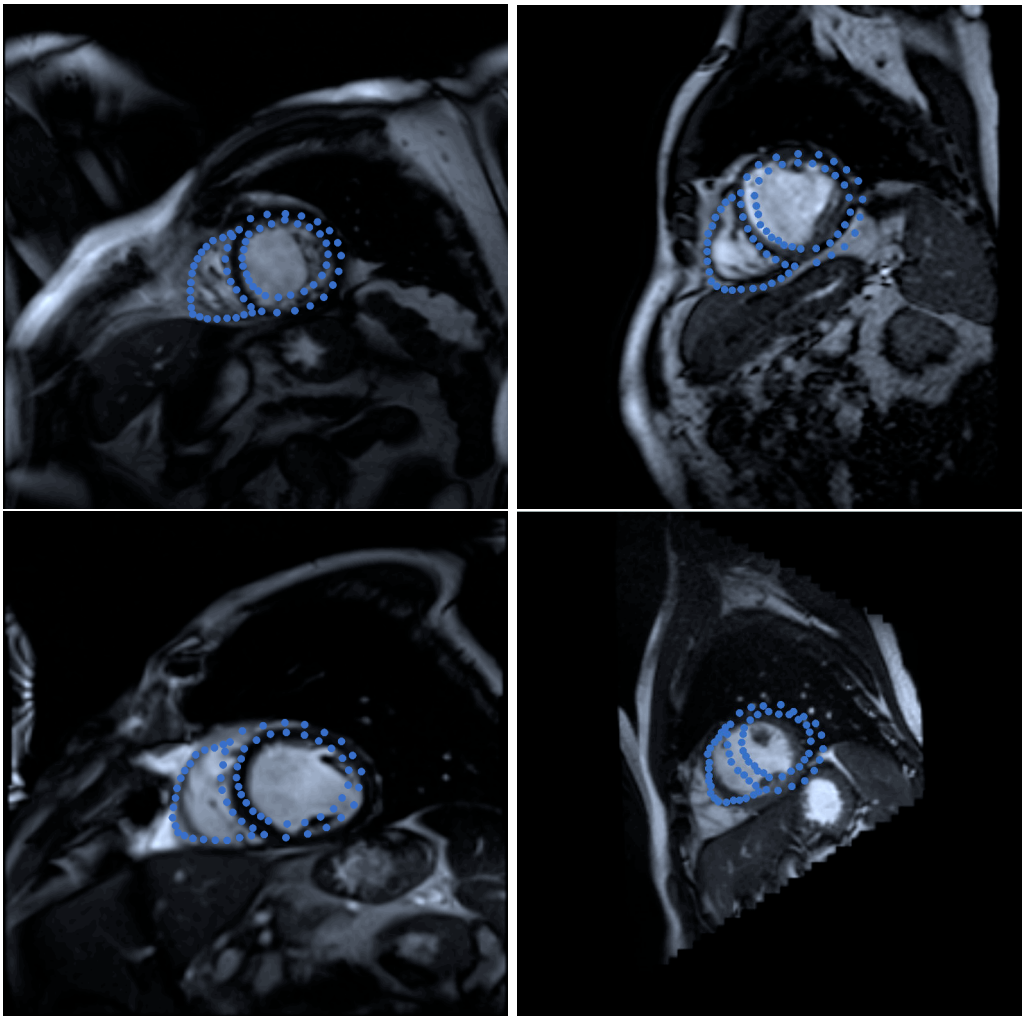


Figure B.3: First stage of the landmark position estimation. Even in the worst case (bottom right corner) we get quite some overlap and global orientation and scale estimation for our model and the heart. The successive stage then further decreases the median prediction error from 6 to 4 pixels (approximately 5 mm).

Validation and results

We did not observe any further performance gains beyond approximately 2,400 example images. This is most likely due to strong correlations between the generated images.

We use 90% of the images for training and the remaining 10% for testing. We split the dataset such that images from patients in the training set do not leak into the testing set. The direct prediction from the CNN features predicts the landmarks with a median error of 6 pixels (approximately 7.5 mm). The median landmark position error then further decreases to 4 pixels (approximately 5 mm) in the second stage.

B.5 Conclusions

We learnt to predict cardiac landmarks by simple regression from CNN-derived features that were adapted to recognise cardiac acquisition planes. For such a simple method, this is quite remarkable. Adding images from different patients, or additional image augmentation techniques such as adding artificial bias fields, could further improve the dataset span for better landmark regression quality for previously unseen images. Of course, fine-tuning is a natural next step to better adapt the CNN parameters and to potentially completely remove the need for the iterative refinement.

Appendix C

A note on pericardial effusion for retrieval

Contents

C.1 Overview

Similarly to the atrial segmentation (Section 3.2), we can use extra image channels to enhance some aspects of the images, such as the pericardial effusion, and potentially use them for image retrieval with the Neighbourhood approximating forest (NAF) algorithm we describe in Chapter 5. In this preliminary study we use an almost identical pipeline to the one we used to retrieve hearts with similar ejection fractions in order.

C.2 Introduction

Pericardial effusion is a common finding in myocardial infarction patients. It manifests itself as an accumulation of liquid in the pericardial sac around the heart. It can be seen on SSFP CMR images as a bright ring around the LV epicardium (see Fig. C.1a). The pericardial effusion is, however, more difficult to detect and to capture. It consists of fewer pixels and has significant variability in appearance. In addition, the pericardial effusion and the pericardial fat are also visually quite similar (apart from small intensity difference).

C.3 Pericardial effusion based similarity

We consider the finding of pericardial effusion as a binary label associated to the image. Currently, we do not distinguish between the pericardial effusion and fat. Since this is a binary label, the compactness for this measure has to be defined differently than for the continuous ejection fraction (Section 5.4.1). We also use an extra image channel enhancing the effusion.

C.3.1 Compactness on effusion

The target cost function $\rho_{effusion}(I, J)$ encourages images with effusion to be put into the same leaves. If both I and J have bright rings around the LV, the pairwise distance is 0 otherwise 1.

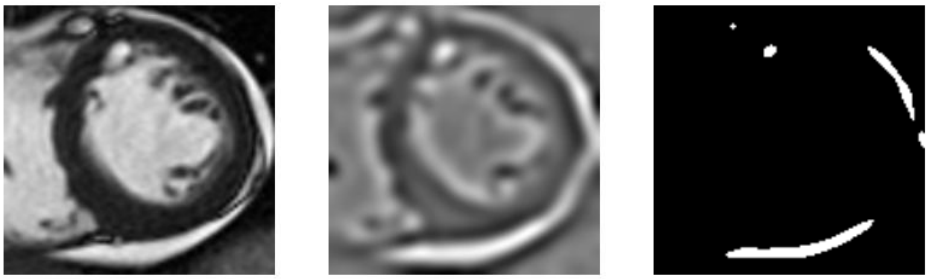
$$\rho_{effusion}(I, J) = \begin{cases} 1 & \text{if Effusion}(I) = \text{Effusion}(J) \\ 0 & \text{otherwise} \end{cases} \tag{C.1}$$

The NAF then attempts to find decisions that separate images without effusion and fat from the images with them.

C.4 An additional image channel

The effusion is a bright ring surrounded by darker tissue. Such structures can be enhanced by convolving the image with a Laplacian of Gaussian (LoG) image filter (see Fig. C.1b).

$$LoG(x, y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2 + y^2}{2\sigma^2}\right] e^{\frac{x^2+y^2}{2\sigma^2}} \tag{C.2}$$



(a) Source image (b) Laplacian filtered (c) Thresholded response

Figure C.1: Pericardial effusion can be enhanced with Laplacian operator ($\sigma = 1.4$ pixels).

C.5 Retrieving similar images

We use the LoG-filter effusion-enhanced images as an additional image channel on which we evaluate the spatio-temporal boxes (Section 5.4.3).

The small number of effusion (and pericardial fat) cases (13) in the dataset and the variability of its appearance was not sufficient to correctly retrieve the affected hearts. Yet, when querying the database for hearts without the effusion, the hearts with it never appeared within the retrieved neighbours and were often the most dissimilar ones (see Fig. C.2).

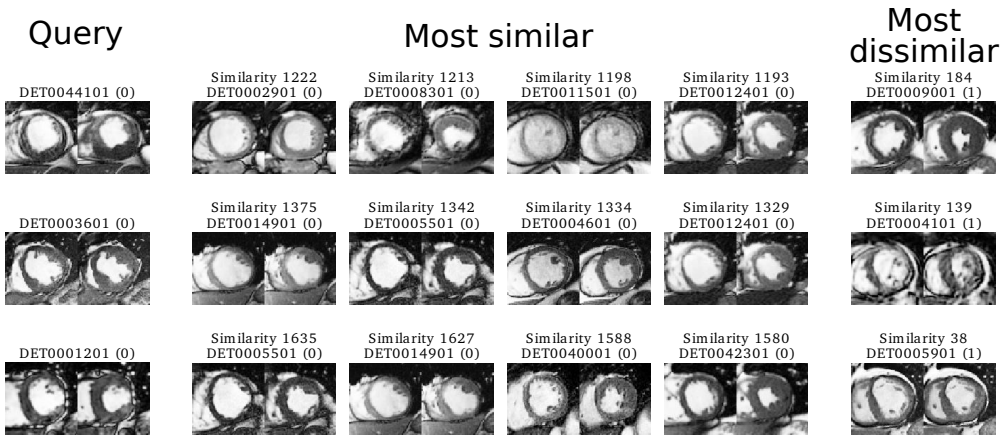


Figure C.2: Retrieval of similar hearts based on the effusion criterion. The leftmost heart is the query image (shown diastolic and systolic mid-cavity slices), next are the four most similar hearts, and rightmost is the most dissimilar heart. The values in the parentheses indicate the presence (1) or absence (0) of the effusion in the images.

C.6 Conclusions and perspectives

Simple convolutional filtering of the images can capture the effusion. Nevertheless, more training examples with expert ground-truth will be necessary to learn good decisions for image retrieval of hearts with effusion.

With filtering, additional aspects of the hearts could be enhanced. Learning these filters can be done with a CNN. Also filters from a pretrained CNN could be explored to create image channels for the NAFs.

Appendix D

Abstracts from coauthored work

Benchmark for Algorithms Segmenting the Left Atrium From 3D CT and MRI Datasets.

C. Tobon-Gomez, A. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, **J. Margeta**, Z. Sandoval, B. Stender, Y. Zheng, M. A. Zuluaga, J. Betancur, N. Ayache, M. A. Chikh, J.-L. Dillenseger, M. Kelm, S. Mahmoudi, S. Ourselin, A. Schlaefer, T. Schaeffter, R. Razavi, and K. Rhode, in IEEE Transactions on Medical Imaging, vol. 34, no. 7, pages 1460 1473, 2015.

Knowledge of left atrial (LA) anatomy is important for atrial fibrillation ablation guidance, fibrosis quantification and biophysical modelling. Segmentation of the LA from Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) images is a complex problem. This manuscript presents a benchmark to evaluate algorithms that address LA segmentation. The datasets, ground truth and evaluation code have been made publicly available through the http://www.cardiacatlas.org website. This manuscript also reports the results of the Left Atrial Segmentation Challenge (LASC) carried out at the STACOM13 workshop, in conjunction with MICCAI13. Thirty CT and 30 MRI datasets were provided to participants for segmentation. Each participant segmented the LA including a short part of the LA appendage trunk and proximal sections of the pulmonary veins (PVs). We present results for nine algorithms for CT and eight algorithms for MRI. Results showed that methodologies combining statistical models with region growing approaches were the most appropriate to handle the proposed task. The ground truth and automatic segmentations were standardised to reduce the influence of inconsistently defined regions (e. g. mitral plane, PVs end points, LA appendage). This standardisation framework, which is a contribution of this work, can be used to label and further analyse anatomical regions of the LA. By performing the standardisation directly on the left atrial surface, we can process multiple input data, including meshes exported from different electroanatomical mapping systems.

A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images.

A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M. P. Jolly, A. H. Kadish, D. C. Lee,

J. Margeta, S. K. Warfield, and A. A. Young, in Medical Image Analysis, vol. 18, no. 1, pages 50 62, 2014.

A collaborative framework was initiated to establish a community resource of ground truth segmentations from cardiac MRI. Multi-site, multi-vendor cardiac MRI datasets comprising 95 patients (73 men, 22 women; mean age 62.73 ś 11.24 years) with coronary artery disease and prior myocardial infarction, were randomly selected from data made available by the Cardiac Atlas Project (Fonseca et al., 2011). Three semi- and two fully-automated raters segmented the left ventricular myocardium from short-axis cardiac MR images as part of a challenge introduced at the STACOM 2011 MICCAI workshop (Suinesiaputra et al., 2012). Consensus myocardium images were generated based on the ExpectationMaximization principle implemented by the STAPLE algorithm (Warfield et al., 2004). The mean sensitivity, specificity, positive predictive and negative predictive values ranged between 0.63 and 0.85, 0.60 and 0.98, 0.56 and 0.94, and 0.83 and 0.92, respectively, against the STAPLE consensus. Spatial and temporal agreement varied in different amounts for each rater. STAPLE produced high quality consensus images if the region of interest was limited to the area of discrepancy between raters. To maintain the quality of the consensus, an objective measure based on the candidate automated rater performance distribution is proposed. The consensus segmentation based on a combination of manual and automated raters were more consistent than any particular rater, even those with manual input. The consensus is expected to improve with the addition of new automated contributions. This resource is open for future contributions, and is available as a test bed for the evaluation of new segmentation algorithms, through the Cardiac Atlas Project (www.cardiacatlas.org).

Confidence-based Training for Clinical Data Uncertainty in Image-based Prediction of Cardiac Ablation Targets.

R. C. Lozoya, **J. Margeta**, L. Le Folgoc, Y. Komatsu, B. Berte, J. Relan, H. Cochet, M. Haïssaguerre, P. Jaïs, N. Ayache, and M. Sermesant, in International Workshop on Medical Computer Vision: Algorithms for Big Data, Held in conjunction with MICCAI 2014, Boston, Lecture Notes in Computer Science, vol. 8848, pages 148 159, B. Menze, G. Langs, A. Montillo, M. Kelm, H. Müller, S. Zhang, W. Cai, and D. Metaxas, Eds., Springer Berlin / Heidelberg, 2014.

Ventricular radio-frequency ablation (RFA) can have a critical impact on preventing sudden cardiac arrest but is challenging due to a highly complex arrhythmogenic substrate. This work aims to identify local image characteristics capable of predicting the presence of local abnormal ventricular activities (LAVA). This can allow, pre-operatively and non-invasively, to improve and accelerate the procedure. To achieve this, intensity and texture-based local image features are computed and random forests are used for classification. However, using machine-learning approaches on such complex multimodal data can prove difficult due to the inherent errors in the training set. In this manuscript we present a detailed analysis of

these error sources due in particular to catheter motion and the data fusion process. We derived a principled analysis of confidence impact on classification. Moreover, we demonstrate how formal integration of these uncertainties in the training process improves the algorithms performance, opening up possibilities for non-invasive image-based prediction of RFA targets.

Local late gadolinium enhancement features to identify the electrophysiological substrate of post-infarction ventricular tachycardia: a machine learning approach.

R. C. Lozoya, **J. Margeta**, L. Le Folgoc, Y. Komatsu, B. Berte, J. S. Relan, H. Cochet, M. Haïssaguerre, P. Jaïs, N. Ayache, and M. Sermesant, in Journal of Cardiovascular Magnetic Resonance, vol. 17, no. Suppl 1, poster 234, 2015.

Most ventricular tachycardias occur on structurally diseased hearts with fibrotic scar, where bundles of surviving tissue promote electrical circuit re-entry. These bundles can be identified on invasive electrophysiological (EP) mapping as local abnormal ventricular activities (LAVA) during sinus rhythm. Although the elimination of LAVAs by radiofrequency ablation was shown to be an efficient therapeutic option, their identification requires is a lengthy and invasive process. Late gadolinium enhancement (LGE) magnetic resonance imaging enables a non-invasive 3D assessment of scar topology and heterogeneity with millimetric spatial resolution. The aim of this work is to identify imaging features associated with LAVA, features that may subsequently be used to target ablation or to stratify the risk of arrhythmia.

Myocardial Infarct Localisation using Neighbourhood Approximation Forests.

H. Bleton, **J. Margeta**, H. Lombaert, H. Delingette, and N. Ayache, in International Workshop on Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, Held in conjunction with MICCAI 2015, Munich, O. Camara, T. Mansi, M. Pop, K. Rhode, M. Sermesant, and A. Young, Eds., 2015.

This paper presents a machine-learning algorithm for the automatic localisation of myocardial infarct in the left ventricle. Our method constructs neighbourhood approximation forests, which are trained with previously diagnosed 4D cardiac sequences. We introduce a new set of features that simultaneously exploit information from the shape and motion of the myocardial wall along the cardiac cycle. More precisely, characteristics are extracted from a hyper surface that represents the profile of the myocardial thickness. The method has been tested on a database of 65 cardiac MRI images in order to retrieve the diagnosed infarct area. The results demonstrate the effectiveness of the NAF in predicting the left ventricular infarct

location in 7 distinct regions. We evaluated our method by verifying the database ground truth. Following a new examination of the 4D cardiac images, our algorithm may detect misclassified infarct locations in the database.

Appendix E

Sommaire en français

E.1 L'aube du Big Data cardiaque

Les développements en cardiologie au cours du dernier siècle (Cooley and Frazier, 2000; Braunwald, 2014) ont étés assez spectaculaires. Beaucoup de révolutions se sont passées depuis le premier électrocardiogramme (ECG) pratique par Einthoven en 1903. Ceux-ci comprennent le cathétérisme cardiaque (1929), la machine coeur-poumon et les premiers modèles animaux dans les années 1950, les chirurgies mini-invasives (1958), et le développement de médicaments (les bêta-bloquants (1962), les statines (1971), et les angiotensines (1974)). L'imagerie diagnostique cardiaque a aussi considérablement améliorée. Le développement après-guerre de l'échographie cardiaque (EC), la tomodensitométrie (TDM) (1970) et l'imagerie par résonance magnétique (IRM) (des années 1980) nous ont aidés à obtenir un regard non-invasif dans le coeur et avec un niveau de détail remarquable.

Toutes ces avancées ont considérablement changées le cours de gestion de la maladie cardio-vasculaire. En 1970, le taux de mortalité liée à ces deux maladies dans les pays à revenu élevé a basculé et a été en baisse constante (Fuster and Kelly, 2010, p52) depuis. Pourtant, les maladies cardiovasculaires demeurent la première cause de mortalité dans le monde (Nichols et al., 2012, p 10;Roger et al., 2011), et causent 47% de tous les décès en Europe. Nous sommes à l'aube de l'âge où de nouvelles techniques cardiaques d'acquisition d'image, des modèles prédictifs (in sillico) cardiaques (Lamata et al., 2014), des simulations d'images réalistes (Glatard et al., 2013; Prakosa et al., 2013; Alessandrini et al., 2015), la surveillance en temps réel du patient (Xia et al., 2013), et les banques d'images cardiaques à grande échelle (Suinesiaputra et al., 2014a; Petersen et al., 2013; Bruder et al., 2013) ont devenus omniprésents et ont la chance d'améliorer la santé cardiaque et notre compréhension. Les données dans ces bases de données ne sont aussi utiles que les questions qu'ils peuvent aider à répondre, les idées qu'ils peuvent engendrer, et les décisions qu'ils permettent de faire. Des grandes études de population avec les recommandations de traitement cliniques peuvent être faites, preuves à l'appui peuvent être adaptées à chaque patient individuellement. Le traitement peut être ajusté en regardant les patients similaires précédemment traités, en comparant leurs résultats et les évolutions de leurs maladies. Des nouveaux outils d'enseignement peuvent aussi être développés en utilisant les données pour créer les patients virtuels comme les études de cas et des simulations de chirurgie sur les modèles imprimé 3D (Bloice et al., 2013; Kim et al., 2008; Jacobs et al., 2008) Ceci pourrait stimuler l'enseignement et la pratique des cardiologues.

E.2 Les défis pour l'organisation des données cardiaques à grande échelle

Les possibilités pour les nouvelles utilisations de grandes bases de données d'image sont innombrables, cependant, l'utilisation de ces bases de données pose de nouveaux défis. Collections cardiaques riches avec des images pertinentes (y compris de nombreuses maladies rares) sont dispersés à travers des milliers de systèmes d'archivage et de transmission d'images (PACS) dans de nombreux pays et hôpitaux. Ces données provenant de sources hétérogènes sont non seulement massives, mais aussi assez non-structurées et bruyantes.

E.2.1 Agrégation des données provenant des études multi-centriques

Les biobanques et consortia internationaux de gestion des bases de données d'imagerie médicale, tels que UK Biobank (Petersen et al., 2013), Cardiac atlas project (CAP) (Fonseca et al., 2011) ou le projet VISCERAL (Langs et al., 2013), ont résolu de nombreux problèmes difficiles dans l'éthique de partage de données d'imagerie médicale et dans l'organisation de la distribution de données — en particulier l'agrégation des données provenant de sources multiples. Dans ces efforts (PACS) ainsi que la norme dimagerie numérique en médecine (DICOM) ont étés inestimables. Des études provenant de plusieurs centres utilisent souvent des nomenclatures spécifiques, suivent des orientations différentes ou utilisent des protocoles d'acquisition différents. Dans ces cas, même ces deux normes ne sont pas suffisantes.

E.2.2 Normalisation des données

Les collections d'images sur des serveurs PACS peuvent être interrogées par l'information du patient (par exemple ID, nom, date de naissance, sexe, taille, poids), la modalité de l'image, la date de l'étude et d'autres balises DICOM, parfois par description de l'étude, des balises personnalisées des cliniciens, des mesures associées (par exemple la pression artérielle ou la fréquence cardiaque) et les codes de maladie ou d'intervention. Voir Fig. 1.1 pour un exemple d'une telle interface.

Il n'y a pas de façon standard pour stocker une partie de l'information d'image importante associée (par exemple, de l'information sur le plan d'acquisition d'images cardiaques). Pour cette information le nom dépend de la mise au point de la station de travail, ou de la langue pratiqué au centre d'imagerie. Même les balises de la norme DICOM contiennent souvent des nomenclatures spécifiques du fournisseur. Par exemple, les mêmes séquences d'acquisition dIRM cardiaque sont marquées différemment à travers des fabricants (Siemens, 2010). Alors que certaines différences de mise en oeuvre existent, elles ne sont pas pertinentes pour l'interprétation des images, et la terminologie pourrait être considérablement simplifiée (Friedrich et al., 2014). L'analyse des publications électroniques avec des images est encore un plus grand défi. Ces images sont rarement en format DICOM et seul le contenu de l'image et la description textuelle sont présents. Ces différences réduisent notre capacité à

interroger et explorer les bases de données pour les images pertinentes de manière efficace. La normalisation peut être appliquée en strictement appliquant les lignes directrices lors de l'acquisition des images, et en utilisant systématiquement les terminologies pour coder les informations associées: tels que SNOMED CT (Stearns et al., 2001). Des précautions doivent être prises pour éliminer les erreurs de saisie manuelle. Les images déjà stockées dans les bases de données sans l'information normalisée devraient être revues pour une meilleure accessibilité.

E.2.3 Récupération des cas similaires

Traverser ces bases de données manuellement et trouver des patients similaires (avec preuves à l'appui) devient donc très chronophage. Récupération des images archivées en PACS est, dans la pratique, assez lent pour une telle utilisation exploratoire. En plus, les données d'imagerie cardiaque stockées dans les bases de données sont fréquemment des séquences 3D+t, et des détails importants peuvent être facilement manqués lors de cette inspection visuelle. Une alternative à cette approche (force brute) est de décrire systématiquement les images avec des représentations plus compactes. Cela prépare les bases de données d'images cardiaques pour la récupération future. Cependant, cette approche limite la recherche sur les cas annotés ou pour les cas connus au clinicien particulier. La plupart de données annotées ne sera donc jamais réutilisée et les données qui ne sont pas utilisées sont les données inutiles.

E.2.4 Annotation des données et le consensus

Lannotation des images simplifie leur réutilisation. Cependant, avec la croissance des données, la demande pour la saisie manuelle devient un fardeau sur les annotateurs experts. Une façon de lapprocher est de réduire les tâches d'annotation en questions simples qui peuvent être très répondues par un plus grand nombre d'évaluateurs moins expérimentés, par exemple via le crowdsourcing. Comme étudié par Suinesiaputra et al. (2015), la variabilité des différents radiologues (experts suivant les mêmes lignes directrices) est non négligeable. Par exemple, dans la segmentation du ventricule gauche, les muscles papillaires sont du tissu myocardique et donc selon Schulz-Menger et al. (2013) devraient idéalement être inclus dans la masse myocardique et exclus du calcul du volume ventriculaire gauche. Les valeurs de référence pour les volumes et les masses correspondantes (Maceira et al., 2006; Hudsmith†et al., 2005) devrait être utilisé dans ce cas. Certains outils incluent les muscles papillaires dans le volume de la cavité à la place. Dans ce cas, un autre ensemble de valeurs de référence doit être envisagée (Natori et al., 2006). Les deux mesures signalées peuvent différer substantiellement. En fin de compte, les muscles papillaires font partie de la progression de la maladie (Harrigan et al., 2008) et méritent une attention individuelle. Les centres d'acquisition sont équipés de différents outils logiciels et tous ces outils ne sont pas aussi capables. Nous avons encore un long chemin à parcourir pour réaliser l'extraction reproductible de mesures basées sur l'image et description cohérente de toutes les informations d'image pertinentes,

surtout étant donné les lignes directrices en constante évolution.

E.2.5 Le besoin d'outils automatisés

Pour réussir dans l'analyse à grande échelle et l'utilisation des données, des moyens efficaces de nettoyage et la description automatique des images cardiaques provenant de plusieurs centres cliniques avec des outils adaptable aux grandes bases de données (Medrano-Gracia et al., 2015) sont primordiales. Comme nous allons le voir sur l'exemple suivant, la conception manuelle de ces outils peut rapidement devenir très difficile.

E.3 Un problème trompeusement simple

Examinez les quatre images mi- ventriculaires petite axe (SAX) obtenues par l'IRM cardiaque avec la séquence dacquisition SSFP sur Fig. 1.2. Ils appartiennent à quatre personnes avec différentes pathologies. L'une d'eux est une image d'un coeur sain, une autre appartient à un patient après un infarctus du myocarde dans la paroi latérale, la troisième à un patient avec un ventricule gauche avec sa fonction cardiaque gravement insuffisant et non-compactant, et la dernière représente un patient avec lépanchement péricardique idiopathique. Pouvez-vous dire correctement lequel est lequel? Cette tâche de lidentification de la pathologie est apparente et sans effort pour une personne expérimentée dans l'interprétation des images cardiaques. Intuitivement, nous pourrions reconnaître le coeur de post-infarctus du myocarde par un amincissement notable de la paroi latérale causé par l'infarctus transmural et une nécrose myocardique subséquente. On peut aussi noter des artefacts de fil de suture sternale d'une chirurgie antérieure. Le coeur non-compactant se manifeste avec une dilatation de la cavité gauche massive, les trabéculations dans la cavité ventriculaire gauche dominantes, et une réduction significative de la contractilité du myocarde (mieux visible sur les séquences d'images cinématographiques). L'épanchement péricardique peut être vu comme un anneau hyper-intense du liquide à l'extérieur du myocarde et le mouvement du coeur pendulaire. Et enfin, le coeur en bonne santé semble "normal".

E.3.1 Automatisation de la tâche

Seulement quand nous essayons d'écrire un programme pour imiter ce raisonnement sur un ordinateur, nous commencerons à apprécier pleinement la véritable complexité des tâches visuelles effectuées par le cerveau. La simplicité de l'extraction d'informations pertinentes à partir des images est très trompeuse. Les concepts intuitifs comme l'amincissement du myocarde, la dilatation de la cavité, la faible contractilité, l'anneau hyper-intense ou le mouvement pendulaire sont inconnus à une machine. Sans oublier la tâche plus globale — de dire automatiquement que toutes ces images proviennent d'une séquence d'acquisition SSFP et sont des coups petite axe. Une des possibilités pour extraire ces informations par un ordinateur est de

commencer à écrire un ensemble de règles. Mesure de l'amincissement du myocarde peut être obtenu comme la longueur de la ligne la plus courte à travers du myocarde, comptant pixels entre deux bords séparant lobjet blanc (le sang dans la cavité) et le environnement gris (sauf pour le cas avec l'épanchement péricardique) à lextérieur du coeur. La dilatation est liée au nombre de voxels au sein de la cavité ventriculaire et son diamètre. Ces deux mesures peuvent être calculées à partir de la segmentation du myocarde ventriculaire gauche. La contractilité peut être estimée à partir du déplacement de pixels, par exemple, par l'intermédiaire du recalage d'images. Les changements subtils que nous voudrions reconnaître sont facilement éclipsés par les différences dacquisition : les images provenant de différentes machines dacquisition, des artefacts d'acquisition et des différences dans l'orientation d'image et la position du coeur dans les images existent. Les images ne sont pas acquises avec des résolutions ni la qualité d'images similaire, les intensités des tissus entre différentes machines ne correspondent pas, des artefacts d'acquisition sont présents ou variations spécifiques aux fournisseurs des protocoles d'acquisition sont utilisées. Nous découvrons bientôt que (pour coder cet ensemble des règles et extraire l'information pertinente pour décrire les images cardiaques n'a pas de frontière).

E.3.2 L'approche de l'apprentissage automatique

L'approche de l'apprentissage automatique est tout à fait différente. Au lieu de coder manuellement les règles, nous spécifions un modèle d'apprentissage et laissons l'algorithme d'apprentissage détermine automatiquement un ensemble de règles en regardant les données, à savoir, pour entraîner le modèle. Dans le cadre de l'apprentissage supervisé, un ensemble d'exemples ainsi que des sorties désirée (par exemple les images et leurs segmentations) sont présentés à l'algorithme de lentraînement. L'algorithme prend alors les règles qui transforment au mieux les entrées aux sorties souhaitées. Il est important que le modèle généralise, autrement dit, quil peut prévoir de façon fiable les sorties pour des images jamais vues, tout en ignorant les différences d'acquisition non pertinentes.

Bien quune bonne prédiction est souhaitable, il est courant d'utiliser les systèmes d'apprentissage automatique "moins que parfait" dans une boucle, et d'améliorer les modèles au fil du temps, lorsque plus de données arrive. Aussi lorsque directives changent, ces algorithmes peuvent être relancés et les images peuvent être traité à nouveau. Prévisions incorrectes peuvent être fixées et ajoutées au nouveau ensemble dentraînement et le modèle peut ensuite être ré-entraîné. L'apprentissage automatique dans l'imagerie médicale est devenu remarquablement présent. Ceci est en partie grâce aux améliorations algorithmiques, mais aussi grâce à la disponibilité de grandes quantités de données. Bien qu'il existe de nombreux algorithmes d'apprentissage automatique, il n'y a pas (encore) un algorithme parfait pour toutes les tâches à accomplir, qui performe bien sur les grands et petits ensembles de données. Tout au long de cette thèse, nous allons utiliser principalement trois familles des algorithmes dapprentissage automatique supervisé: les modèles linéaires (les machines à vecteurs de support (SVM) (Cortes and Vapnik, 1995) et la régression

linaire (Golub et al., 1979).), les forêts de décision (Ho, 1995; Amit and Geman, 1997; Breiman, 1999), et les réseaux de neurones à convolution (CNN) (Fukushima, 1980; LeCun et al., 1989).

E.4 Les questions de recherche de cette thèse

Cette thèse vise à répondre à la question globale suivante: “Comment pouvons-nous simplifier l’utilisation des banques d’images cardiaques pour les cardiologues et chercheurs grâce à l’apprentissage automatique?” Pour nous aider à répondre à cette question, nous avons abordé certains des principaux défis présentés dans la section E.2.

E.4.1 Comment pouvons-nous nettoyer et normaliser les balises DICOM pour rendre le filtrage et le regroupement des séries d’images plus facile ?

Un des premiers problèmes auxquels nous sommes confrontés en imagerie cardiaque lorsquon traite bases de données hétérogènes est le manque de standardisation dans la notation utilisée dans les protocoles d’acquisition (Friedrich et al., 2014) ou la désignation de plans d’acquisition cardiaques. Surtout la connaissance des plans cardiaques est essentielle pour regrouper les images en série et choisir les algorithmes de traitement d’image appropriés. Le chapitre 2 présente nos deux méthodes de nettoyage des métadonnées DICOM en les estiment directement à partir du contenu de l’image. Notre première méthode pour reconnaître les plans d’acquisition utilise les forêts de classification appliquée sur les miniatures d’images. Nous montrons comment les nouvelles images générées à moindre coût peuvent contribuer à améliorer la reconnaissance. Nous montrons ensuite comment modifier une technique de létat de l’art, entraîné initielement pour la reconnaissance d’objets visuels à grande échelle, basée sur les CNN, pour les données d’imagerie cardiaque beaucoup plus petites. Notre deuxième méthode reconnaît vues petite et longue axe (2-, 3- et 4-chambre) avec la performance de reconnaissance très prometteur. Dans l’annexe B, nous montrons comment les fonctionnalités basées sur les CNN peuvent être réutilisés pour la prédiction des distributions de points cardiaques pour l’alignement d’image inter-patient.

E.4.2 Peut-on apprendre à l’ordinateur l’anatomie cardiaque et de segmenter les structures cardiaques à partir d’images IRM ?

Une fois que nous pouvons décrire les images cardiaques en fonction de leurs plans dacquisition et les fusionner en volumes spatio-temporelles 3D+t, nous continuons à apprendre à l’ordinateur les bases de l’anatomie cardiaque — comment segmenter les images cardiaques. La segmentation est essentielle pour extraire les

indices cardiaques basés sur les mesures volumétriques standards telles que le volume systolique et diastolique, fraction d'éjection, et la masse myocardique. Dans le chapitre 3, nous étendons les travaux antérieurs sur la segmentation en utilisant les forêts de classification sémantique (Shotton et al., 2008; Geremia et al., 2011). Nous montrons comment cet algorithme modifié apprend à segmenter les ventricules gauches à partir des séquences IRM SSFP 3D+t SAX sans imposer aucune forme préalable. Notre classificateur est entraîné en deux couches successives, et nous proposons des nouvelles caractéristiques spatio-temporelles pour segmenter les séquences 3D+t. Nous montrons que, cet approche de segmentation nous permet de l'adapter facilement à d'autres structures cardiaques, les oreillettes gauches — la boîte noire du coeur, à la fois de l'IRM et du TDM. Nous avons contribué ces algorithmes à deux études de comparaison pour l'évaluation équitable. Dans l'annexe A, nous proposons une méthode de segmentation semi-supervisé qui exploite les données non labélisées pour apprendre à segmenter à partir de annotations éparses.

E.4.3 Comment peut-on recueillir des données nécessaires à pour les algorithmes d'apprentissage automatique et comment apprendre à décrire les coeurs avec des attributs sémantiquement significatives ?

La plupart des problèmes d'apprentissage automatique pratiques sont actuellement encore résolu d'une manière entièrement supervisé. Par conséquent, il est essentiel d'acquérir la vérité terrain. Le chapitre 4 traite la collection de la vérité terrain pour les algorithmes d'apprentissage automatique. Nous concevons un outil Web de crowdsourcing (approvisionnement par la foule ou production participative) des attributs cardiaques et l'utilisons pour recueillir des annotations d'image par paires. Nous décrivons les formes cardiaques avec leurs signatures spectrales et utilisons un prédicteur linéaire basée sur les SVM pour apprendre à l'ordonner les images en fonction de leurs valeurs d'attribut. Nos résultats préliminaires suggèrent qu'en plus des mesures volumétriques à partir des segmentations cardiaques, les coeurs pourraient être décrits aussi par des attributs cardiaques sémantiques.

E.4.4 Peut-on récupérer automatiquement les coeurs semblables ?

La similarité entre les images dépend de la question clinique. Les requêtes que nous pourrions vouloir demander au système de récupération des images peuvent être très variables. Chapitre 5 se fonde sur la forêt d'approximation du voisinage (NAF) de Konukoglu et al. (2013) et présente notre algorithme pour apprendre la forme, l'apparence et les similitudes de mouvement entre les images cardiaques et comment nous les utilisons pour structurer les ensembles de données cardiaques spatio-temporelles. Nous montrons comment les coeurs avec des propriétés similaires (fraction d'éjection similaire) peuvent être récupérés de la base de données. Dans (Bleton et al., 2015), nous avons ensuite utilisé une technique similaire pour localiser les infarctus cardiaques à partir des formes dynamiques (sans besoin dun agent de

contraste).

E.5 Organisation du manuscrit

Cette thèse est organisée autour de notre travail publié et nos travaux en préparation. Le manuscrit progresse à peu près du niveau global vers la description plus fine des images cardiaques. Chaque chapitre de cette thèse tente de répondre à l'un des objectifs et d'emmener la récupération d'images basée sur le contenu des bases de données à grande échelle CMR plus proche vers la pratique. Tout d'abord, nous entraînons un système de nettoyage des balises d'images qui ne sont pas capturés par DICOM directement à partir du contenu d'image. Dans le chapitre 2, nous montrons comment reconnaître automatiquement les plans dacquisition cardiaques. Dans le chapitre 3, nous proposons une technique de segmentation automatique flexible qui apprend à segmenter les structures cardiaques à partir de données d'image spatio-temporelles, en utilisant de simples cartes des voxels comme la vérité terrain. La segmentation pourrait être utilisée pour l'extraction automatique des mesures clinique. Dans le chapitre 4, nous proposons un moyen de collectionner les annotations nécessaires pour lentraînement des algorithmes automatiques, et pour décrire les images cardiaques avec un ensemble d'attributs sémantiques. Enfin, dans le chapitre 5, nous proposons un algorithme pour structurer les ensembles de données pour y trouver des cas similaires à l'égard de différents critères cliniques. Le chapitre 6 conclut la thèse avec des perspectives et des travaux futurs. Dans les annexes, nous avons illustré comment les données non labélisées peuvent être utilisées pour la segmentation d'images guidée (Annexe A), comment estimer les repères cardiaques pour l'alignement des images (Annexe B), ou comment rehausser épanchement péricardique pour la récupération d'image (Annexe C).

E.6 Sommaires des chapitres

E.6.1 Reconnaisance des plans d'acquisition cardiaques

Lorsqu'on traite des grandes bases de données depuis plusieurs centres et fournisseurs de machines d'acquisition, notations incompatibles sont un facteur limitant pour l'analyse automatisée. Les plans d'acquisition de l'IMR cardiaque sont un très bon exemple d'un tel échec de normalisation de la notation. Sans savoir quel plan cardiaque on traite, l'utilisation des données sans intervention manuelle est limitée. Dans ce chapitre, nous proposons deux techniques d'apprentissage automatique supervisée pour récupérer automatiquement les informations manquantes ou bruyantes et de prédire les cinq vues (ou plans d'acquisition) cardiaques les plus courants. Nous montrons que les plans d'acquisitions cardiaques sont à peu près alignés pour situer le coeur au centre de l'image. Nous l'utilisons pour apprendre prédicteurs des plans d'acquisition cardiaques à partir d'images 2D.

Dans notre première méthode nous entraînons une forêt de classification sur les miniatures d'image. L'augmentation de l'ensemble de données avec des transfor-

mations géométriques (qui préservent les étiquettes) est un moyen efficace qui nous aide à améliorer la précision de la classification sans des acquisitions ou des données supplémentaires. Nous améliorons la performance de la forêt en affinant un réseau de neurones à convolution profond, à l'origine pré-entrainé sur une grande base de données pour la reconnaissance d'images naturelles (ImageNet LSVRC 2012). Nous transférons la représentation apprise à la reconnaissance des vues cardiaques.

Nous comparons ces approches avec les prédictions à l'aide des caractéristiques extraites des images en utilisant le CNN entraîné directement sur ImageNet, et avec un CNN entraîné à partir de zéro. Nous montrons que peaufinage est une approche viable pour adapter les paramètres des grands CNN pour les petits problèmes. Nous validons respectivement cet algorithme sur deux études cardiaques différentes avec 200 patients et 15 témoins sains. Celui-ci provient d'un ensemble de données cardiaques en libre accès qui simplifie la comparaison directe avec des techniques similaires dans l'avenir. Nous montrons qu'il y a une valeur significative à peaufiner un modèle entrainé pour des images naturelles et de le transférer aux images médicales. Les approches présentées sont tout à fait générales et peuvent être appliquées à toute tâche de reconnaissance d'image. Notre meilleure approche réalise un score de F1 moyen de $97,66\%$ et améliore l'état de l'art sur la reconnaissance de vues cardiaques en utilisant le contenu d'image seulement. Elle évite les annotations supplémentaires et apprend la représentation de la fonction voulue automatiquement.

Ceci est une composante important pour organiser et filtrer des grandes collections de données cardiaques avant de les analyser. Elle nous permet de fusionner des études à partir de plusieurs centres, afin de permettre le filtrage d'image plus intelligent, la sélection des algorithmes de traitement d'image les plus appropriés, et une amélioration de la visualisation d'ensembles de données cardiaques, et pour la recherche d'image par le contenu et un contrôle de qualité.

E.6.2 Segmentation d'images cardiaques

Les indices cliniques les plus couramment utilisés pour la quantification des maladies cardiaques sont souvent basés sur les mesures géométriques des structures anatomiques, telles que les volumes des cavités, des masses du myocarde ou des épaisseurs de paroi. Il serait très utile de les avoir stockées dans les bases de données cardiaques à côté des images et les utiliser pour sélectionner les cohortes de patients pour les études cliniques. Ces mesures sont souvent calculées à partir des segmentations de ces structures cardiaques. La segmentation constitue également la base pour la compréhension de l'anatomie cardiaque par un ordinateur.

Dans ce chapitre, nous présentons une méthode d'apprentissage automatique flexible pour la segmentation d'images IRM cardiaques. Nous segmentons deux des structures cardiaques les plus importantes. Tout d'abord, nous apprenons comment segmenter les ventricules gauches des séquences dynamiques 3D+t SAX directement à partir de la vérité-terrain par voxel en utilisant les forêts de classification avec le contexte spatio-temporel. Ensuite, nous montrons que cette méthode peut être

facilement adaptée à d'autres structures du coeur, tel que l'oreillette gauche.

E.6.3 Approvisionnement par la foule des attributs semantiques

Dans chapitre 2, nous avons montré comment automatiquement prédire les métadonnées d'image de grâce aux CNNs. Nous pouvons facilement récupérer, par exemple, toutes les coupes SAX du ventricule gauche dans la base de données. Les méthodes automatiques de segmentation d'image (tels que celles présentés dans le chapitre 3) nous permettent d'indexer et d'interroger les bases de données en utilisant des mesures géométriques simples (par exemple les volumes des cavités, la masse myocardique ou l'épaisseur du mur). Ces deux outils sont des étapes importantes vers la récupération d'images cardiaque en fonction du contenu automatisée.

Quelques images, cependant, ne peuvent être directement décrits et répertoriés sur si simple mesures géométriques seules. Considérons une requête un peu plus complexe: "récupérer toutes les images de petite axe des coeurs akinétiques avec un amincissement significatif de la paroi, et le limiter que sur les images de qualité diagnostique".

Nous visons à atteindre la description informatisée des images cardiaques avec un ensemble des attributs sémantiques pour la forme, le mouvement et l'apparence. Dans ce chapitre, nous nous concentrons sur les images de coeurs d'infarctus du myocarde avec une légère à modérée dysfonction ventriculaire gauche. Nous développons un outil qui va permettre de filtrer les bases de données cardiaques avec un tel ensemble d'attributs. Nous apprenons la machine à décrire les images en fonction des caractéristiques extraites de l'image et les comparaisons par pair. La vérité terrain se compose de deux images et un indicateur disant quelle image des deux dispose d'une présence inférieure ou supérieure de l'attribut. Nous avons conçu une interface web permettant de recueillir telles annotations de vérité-terrain via approvisionnement par la foule.

E.6.4 Recherche d'image par le contenu

Dans ce chapitre, nous proposons une méthode pour apprendre à rapprocher la similitude entre les images pour la récupération automatique d'images basée sur le contenu des coeurs sémantiquement similaires. Nous nous appuyons sur les forets d'approximation du voisinage (NAF) de (Konukoglu et al., 2013), un algorithme que nous entraînons pour capturer les similitudes entre les images cardiaques. Il permet les récupérations efficaces des coeurs les plus similaires basés sur des critères cliniques. Nous illustrons son utilisation sur une base de données des patients après un infarctus du myocarde.

En Bleton et al. (2015), nous avons déjà montré comment des voisins cardiaque peuvent être utilisés pour localiser infarctus sans injecter l'agent de contraste à partir des segmentations dynamiques. Ici, nous combinons les attributs d'images spatio-temporelles avec les NAFs et utilisons la similitude dérivée de la fraction d'éjection pour trouver les coeurs avec une fonction de pompage similaire. Aucune

segmentation d'image n'est plus nécessaire.

E.7 Conclusions et perspectives

Au cours des dernières années, plusieurs initiatives de collecte des images cardiaques ont apparus. Ceux-ci incluent Cardiac Atlas Project (Fonseca et al., 2011) contenant le sous-ensemble DETERMINE (Kadish et al., 2009) fréquemment utilisé dans cette thèse, mais aussi une étude sur les coeurs asymptomatiques MESA (Bild, 2002). UK Biobank (Petersen et al., 2013) vise à imager une grande partie de la population, et le registre EuroCMR (Bruder et al., 2013) collectionne de données d'imagerie à partir de 57 centres dans 15 des pays. Ces collections déjà très massives sont ici pour grandir et vont avoir un impact sur la santé cardiaque. Nous devrons trouver des façons d'automatiquement interpréter l'information contenue dans ces bases de données et de simplifier la recherche dedans.

Dans cette thèse, nous avons abordé certains des défis les plus importants dans l'organisation des données cardiaques et dans l'extraction d'informations à partir de ces ensembles de données grâce à l'apprentissage automatique.

E.7.1 Synthèse des contributions

Nous avons abordé les quatre questions suivantes. Ici, nous les remontrons avec nos contributions et les limites des méthodes proposées.

E.7.1.1 Estimation des métadonnées manquantes à partir de contenu de l'image

Les balises DICOM sont intrinsèquement bruyantes et ne peuvent pas être utilisées avec confiance. Dans cette thèse, nous avons montré qu'au lieu de compter sur DICOM, nous pouvons estimer une partie des métadonnées (les plans d'acquisition cardiaques) à partir du contenu de l'image.

La contribution principale du chapitre 2 est qu'un algorithme robuste de reconnaissance des vues cardiaques peut être efficacement entraîné de bout en bout sans l'aide d'une conception des attributs avec un réseau de neurones à convolution. Ceci est possible en peaufinant des paramètres du réseau préalablement formés pour un problème de reconnaissance d'image à grande échelle. Nous avons publié nos résultats dans Margeta et al. (2015c, 2014).

- Nous atteignons la performance de l'état de l'art en reconnaissance des vues cardiaques par l'apprentissage d'un système de bout en bout à partir des intensités d'image grâce à l'aide d'un CNN

- Ceci est l'un des premiers travaux à démontrer la valeur de l'extraction des attributs à partir des images médicales utilisant un CNN initialement entraîné pour la reconnaissance d'objets visuels avec une large quantité de données

- Nous montrons que le peaufinage des paramètres d'un CNN pré-entraîné pour la reconnaissance d'objets visuels est une bonne stratégie qui aide à améliorer la performance et à accélérer l'entraînement du réseau
- Nous montrons qu'un CNN peut être appliquée à des ensembles de données plus petits (un problème commun en imagerie médicale) grâce à l'initialisation du réseau prudent et l'augmentation artificielle de l'ensemble de données, même lors de l'entraînement du réseau à partir de zéro
- Nous reproduisons également l'observation de Zhou et al. (2012) que les vecteur d'orientation peuvent être tirés de balises DICOM pour une reconnaissance de vue cardiaque (si ces balises d'orientation sont présentes)

Le peaufinage d'un réseau est un excellent moyen d'apprendre des modèles complexes avec des données médicales d'une plus petite taille et les CNNs sont des outils puissants d'apprentissage automatique qui devraient encore complètement propager dans l'imagerie médicale. Nous avons entraîné le modèle que sur les vues cardiaques uniquement et sur un ensemble de données des coeurs de l'infarctus du myocarde. Ces coeurs ont certainement quelques particularités que nous avons montré dans la thèse, néanmoins, leurs apparences et leurs topologies sont encore assez "normales". Nous avons encore à tester cette méthode sur des cas pathologiques ou des acquisitions moins standards (non-centrées, orientées avec des angles atypiques). Il y a aussi beaucoup plus de vue cardiaques et d'autres séquences d'acquisition pour être reconnus (Taylor and Bogaert, 2012). La vérité terrain doit être recueilli pour toutes ces séquences d'acquisition avec les vues, ce qui conduit à un problème combinatoire. Beaucoup de ces combinaisons sont sous-représentées, et l'apprentissage avec très peu exemples toujours reste un problème difficile. En outre, la reconnaissance des images cardiaques ne devrait pas s'arrêter sur la description des vues. Ce travail est une pierre angulaire pour la reconnaissance des modalités d'image, des protocoles d'acquisition et pour la génération automatique des descriptions d'image comme fait récemment pour des images naturelles à l'aide récurrente de neurones réseaux (Vinyals et al., 2014; Karpathy et al., 2015; Donahue et al., 2014).

E.7.1.2 Segmentation d'images cardiaques

Segmentation des structures cardiaques permet d'indexer les bases de données à l'aide automatique des mesures des volumes de tissus et l'estimation de leurs masses. Dans le chapitre 3 nous proposons de segmenter les images d'IRM cardiaque avec les forêts de classification comme dans Shotton et al. (2008); Lempitsky et al. (2009); Geremia et al. (2011). La principale contribution de cette partie de l'ouvrage est que la segmentation des structures cardiaques peut être réalisée à partir de la vérité terrain par voxel.

- Nous avons proposé une méthode de segmentation basée sur la forêt de classification pour ventricule gauche (LV) à partir des séquences cardiaques d'IRM 3D+t et de l'oreillette gauche à partir des images 3D d'IRM et de TDM

- Nous avons introduit une nouvelle approche en deux couches pour la normalisation des intensités de l'image et le recalage cardiaque inter-patients
- Nous avons conçu caractéristiques spatio-temporelles pour apprendre à segmenter les séquences dynamiques
- Nous avons proposé d'utiliser d'autres canaux d'image tels que la vascularité, ou distances aux contours sanguines pour segmenter les images auriculaires
- Nous avons également montré (chapitre A), comment les données non-étiquetées peuvent aider à segmenter les images d'une manière semi-supervisé et comment les meilleurs points pour étiquetage peuvent être proposés

Ce travail a abouti à deux publications (Margeta et al., 2012, 2013), et nous avons envoyé les résultats de nos algorithmes à deux études comparatives (Suinesiaputra et al., 2014b; Tobon-Gomez et al., 2015) pour l'évaluation équitable. Nos méthodes ne sont pas les meilleurs performeurs en termes de précision de segmentation. Cependant, elles brillent en termes de la façon dont peu de connaissance sur le problème étaient codé en dur dans l'algorithme ce qui les rend assez flexible. Les algorithmes se comporteront également mieux avec plus d'exemples. Cela rend nos approches beaucoup plus prêt à être appliquées à d'autres modalités cardiaques, les vues et les tissus différents, et pour les données en pleine croissance. Bien que nos méthodes n'utilisent pas beaucoup d'informations à priori, certain post-traitement pour la régularisation et les grands ensembles de données sont essentiels pour apprendre une solution robuste. Des mesures supplémentaires telles que l'augmentation des données et l'apprentissage des canaux d'image discriminants probablement aideraient à encore améliorer la performance.

E.7.1.3 Collection de vérité-terrain pour décrire les coeurs avec les attributs sémantiques

Dans cette thèse, nous nous sommes concentrés principalement sur les approches d'apprentissage machine supervisé. Cela signifie que l'algorithme apprend à partir d'exemples où la vérité terrain est requise. L'acquisition de la vérité terrain est au moins aussi importante que l'acquisition des images. Notre solution au problème de la collecte d'annotation est l'approvisionnement par la foule, un outil qui collecte les annotations sur internet afin d'apprendre les descriptions sémantiques des coeurs décrits dans le chapitre 4.

La maladie elle-même n'est généralement pas un processus discret (oui ou non) et se manifeste elle-même à travers des plusieurs façons. Certains d'entre eux peuvent être décrits par les cliniciens en termes sémantiques, les attributs cardiaques. Il y a une discordance entre cette description sémantique et la représentation par l'ordinateur, qui rend la recherche pour les coeurs dans les bases de données à l'aide de telle description difficile. Grâce aux comparaisons par paires, moins d'expérience dans l'interprétation des images cardiaques est nécessaires pour répondre correctement à des questions difficiles.

- Nous concevons une interface web pour le crowdsourcing de cette vérité-terrain.
- Nous proposons de décrire les images cardiaques avec des attributs sémantiques tels que l'hypertrophie, l'amincissement, la kineticité, la dilatation ou la qualité d'image à partir des comparaisons par paires.
- Nous montrons, comment les concepts du Web ranking peuvent être utilisés pour apprendre ces attributs.

Notre experts (3 annotateurs internes) restent limités par la quantité de vérités terrains productibles, ce qui affecte les conclusions que nous pouvons tirer. La variabilité et la qualité des annotateurs est également un aspect important à étudier. Lorsque HIPAA est assurée, un tel outil pourrait être réparti entre les étudiants en médecine pour générer une ressource précieuse pour l'apprentissage tout en contribuant à l'ensemble de données en même temps. Ayant de multiples annotations nous aiderait également à évaluer la fiabilité des prédictions. La présentation d'une combinaison aléatoire des images pour les annoter n'est pas idéal non plus car il gaspille leur temps sur des questions potentiellement trop simples que la machine peut déjà prévoir avec certitude. Une boucle active devrait être plutôt utilisée, l'intervention humaine serait sollicitée principalement pour les cas les moins confiants.

Alors que nous affirmons que toute pathologie pourrait être décrite avec des attributs sémantiques, les représentations spectrales des coeurs ne peuvent pas capturer tous les aspects. Non seulement nous avons limité la recherche aux ventricules gauches mais aussi pour les petites structures (tels que parois du myocarde très minces) les caractéristiques spectrales deviennent instables à cause du changement de la topologie. En outre, ces représentations dépendent d'une segmentation de bonne qualité.

Nous sommes également limités par notre imagination des attributs. Les techniques de découverte des attributs sont essentielles dans les prochaines étapes. Une fois plus de données sont collectées, l'apprentissage de meilleures représentations d'entités (par exemple (Kang et al., 2015; Zhang et al., 2014a)) pour la prédiction des attributs est crucial pour améliorer les prédictions.

E.7.1.4 Structurer les ensembles de données avec similitude clinique

Enfin, chapitre 5 discute comment les forêts de décision approximant le voisinage (NAF) possèdent des propriétés souhaitables pour structurer des ensembles des données d'imagerie et peut être utilisé pour apprendre la similarité pour la recherche d'image base sur le contenu.

- Nous montrons comment NAF peut être utilisée pour calculer la similarité clinique à l'égard de la fraction d'éjection (EF) ou l'emplacement du myocarde (Bleton et al., 2015): les aspects qui captent la forme, le mouvement et l'apparence des séquences cardiaques.

- Nous montrons des façons simples de supprimer l'arrière-plan statique et d'estimer les principales phases cardiaques à partir des images
- Nous proposons de réutiliser le CNN précédemment peaufiné pour l'estimation des repères cardiaques et alignement de l'image (chapitre B).
- Nous indiquons, comment l'épanchement péricardique pourrait être amélioré avec le filtrage simple (chapitre C)

En théorie, la NAF peut être entraînée avec une vérité terrain de similitude arbitraire et utilisé pour les requêtes exploratoires rapides au sein de la grande base de données cardiaque. L'utilisation de la fraction d'éjection elle-même ne conduit pas à une prestation supérieure à l'usage de la segmentation automatique et extraction des mesures.

Notre méthode a besoin de plus de validation et de grands ensembles de données pour améliorer la fiabilité des résultats récupérés et effectuer des prédictions sur différentes critères de similitude. En outre, les canaux d'image supplémentaires seront nécessaires pour mieux capter l'information pertinente. Alors que les résultats retournés ne sont pas parfaits, ils sont nettement mieux que simplement récupérer des images au hasard. Les images récupérées sont pertinents pour la question clinique. En fin de compte, la collecte des données d'utilisation réelle dans une boucle sera nécessaire pour définir quels critères de similarité sont réellement utiles et pour personnaliser les résultats de récupération.

E.7.2 Perspectives

Dans cette thèse, nous avons seulement gratté la surface de la gestion de grandes bases de données cardiaques avec l'apprentissage automatique, la collecte des annotations, ou la recherche d'image par le contenu des cas similaires. Il y a encore beaucoup de défis à relever.

E.7.2.1 Les approaches multimodales

Différentes modalités d'imagerie cardiaque se complètent mutuellement avec l'information qui y est contenues. Les approches multi-modalités (Galderisi et al., 2015) seront indispensables. Aucun de nos outils n'est pas spécifique à la modalité, et les mêmes algorithmes pourraient être presque directement appliqués à ces modalités.

Dans cette thèse, nous avons utilisé principalement les images d'IRM SAX, acquises avec le protocole SSFP. Ceci n'est clairement pas suffisant. Autres vues cardiaques peuvent offrir un éclairage précieux sur les différents aspects du coeur qu'on ne voit pas sur de courtes images SAX. Le protocole d'acquisition IRM avec le rehaussement retardé capte l'ampleur de l'infarctus du myocarde, les séquences d'imagerie marqués aident à mieux estimer le mouvement, l'imagerie à contraste de phase capte le flux sanguin à travers des veines et des artères. De nouvelles modalités et protocoles apparaissent. Imagerie T1 à la puissance du champ magnétiques à 3T peut être utilisé pour détecter de manière fiable et quantifier la fibrose du

myocarde sans avoir besoin d'agent de contraste (Kali et al., 2014) (par exemple aux régions infarcis d'image dans les cardiomyopathies hypertrophiques dilatées, les cardiomyopathies, ou sténoses aortiques). Les séquences de $T2^*$ (Anderson, 2001) captent la surcharge de fer et permettent la détection précoce de la dysfonction ventriculaire, avant qu'elle se développe. D'autres modalités telles que la TDM et l'EC peuvent apporter un autre point de vue dans étude de l'anatomie cardiaque et du mouvement.

E.7.2.2 Données croissantes

L'étude du registre EuroCMR (Bruder et al., 2009) a imagé 27309 patients à travers de 57 sites dans 15 pays européens selon un protocole standard (Kramer et al., 2008). Dans $61,8\%$ de tous les sujets, l'étude a révélé que l'IRM avaient l'impact significative sur le diagnostic et les conséquences thérapeutiques (changement de médication, procédure invasive, sortie de l'hôpital ou de l'admission). Au tour de $8,7\%$ des analyses ont révélés des nouvelles diagnoses ne pas soupçonnés avant.

Mais nous avons encore à établir les avantages de l'IRM cardiaque sur les résultats. Plus de gens seront probablement scannés et plus de données seront générés. Les ordinateurs corporels (wearable computing) ont déjà commencés à enregistrer des informations de santé en temps réel. Il deviendra bientôt nécessaire de chercher dans ces données avec des outils automatiques.

Avec ces bases de données faramineuses et ces sources de données hétérogènes, le nettoyage automatique des métadonnées directement à partir du contenu de l'image et de l'indexation des bases de données pour un accès rapide aux études pertinentes deviennent d'autant plus importants que jamais.

E.7.2.3 Génération de plus de données

Nous sommes toujours confrontés aux défis de l'acquisition de données coûteuse, de l'intimité du patient, de la collecte de données et leur distribution et de l'absence de la vérité terrain. En attendant, si nous pouvions générer des données synthétiques, massives et réalistes à l'aide des modèles biophysiques, nous serons en mesure de former des modèles plus riches et seulement après les accorder aux problèmes de données réelles. Ceci est connu comme l'apprentissage par transfert (Pan and Yang, 2010) et est l'un des principaux objectifs du projet ERC MedYMA.

Il est nettement moins cher à produire des données synthétiques que de les acquérir et les annoter. Il n'y a pas besoin d'obtenir le consentement ou la restriction des patients sur le partage de données, les paramètres sous-jacents utilisés pour générer les images (la vérité terrain) sont connu. Ces approches *in sillico* peuvent produire de grandes quantités d'images avec des étiquettes connues, facilitant algorithmes d'entraînement pour inverser le processus génératif et d'obtenir les étiquettes à partir du contenu d'image. Générateurs d'images cardiaques synthétiques de haute qualité existent déjà pour une large gamme de modalités (Alessandrini et al., 2015; Prakosa et al., 2013; De Craene et al., 2014; Tobon-Gomez et al., 2011).

Il y a des plates-formes sur le Web qui sont dédiées à la simulation multi-modalité d'image médicale qui permettent de démocratiser ce processus (Glatard et al., 2013).

Les données synthétiques de profondeur ont rendu l'estimation de la pose en temps réel possible (Shotton et al., 2011) et les détecteurs robustes ont été entraînés sur les images radioscopiques synthétisés (Heimann et al., 2014). Geremia et al. (2013) a appris à estimer la densité cellulaire des tumeurs sur les images cliniques grâce à la simulation des images multimodales en utilisant les modèles biophysiques de croissance des tumeurs cérébrales synthétiques à diverses positions dans le cerveau. Les images cardiaques simulées (Prakosa et al., 2013) ont été utilisées pour inverser l'électrophysiologie cardiaque (Prakosa et al., 2014) et puis les modèles ont été transférés à des images réelles. Les possibilités sont infinies.

E.7.2.4 Augmentation de données

Nous avons montré que même un moyen plus simple de générer des images synthétique mais aussi très réalistes pour améliorer la performance du système d'apprentissage automatique: modifier les images déjà existantes, telles que les étiquettes sont conservées. Ceci est appelé l'augmentation de données. L'augmentation est un moyen pas cher et efficace pour couvrir une partie de la variabilité dans l'ensemble de données (cela peut aider si les attributs extraites ne sont pas invariantes à ces changements). Transformations d'image simples (Decoste and Schölkopf, 2002) et manipulations d'intensité sont souvent effectués. Artefacts d'acquisition d'image tels que vignettage ou la distorsion en barillet (Wu et al., 2015) ou bruit de fond réaliste (Hannun et al., 2014) sont fréquemment ajoutés.

En général, le problème d'augmentation des images médicales est plus difficile parce que soins doivent être prises pour ne changer que le contenu de l'image et pas l'étiquette associée (par exemple un changement simple de la taille de l'image par une échelle isotrope pourrait faire un coeur normal ressembler plus à un coeur dilaté). L'ajout du bruit artificiel, du champ de biais d'IRM synthetique ou des artefacts d'images vont tous augmenter la robustesse de nos méthodes.

Nous aurons besoin de modèles d'apprentissage automatique qui ont la capacité d'apprendre à partir de ces données massives. Idéalement, ceux-ci seront des modèles dans lesquels les données de l'entraînement peuvent être itérativement streamés en lots (tels que les CNNs entraîné avec la descente de gradient stochastique, ou dans les forêts bootstrapés) de sorte que ce sera assez simple à entraîner avec des nouvelles données et mettre à jour leurs paramètres à la volée. Le CNN que nous avons utilisé pour la reconnaissance de plans cardiaques sont un bon outil pour le problème de données multimodal de croissance.

E.7.2.5 Collecte d'étiquettes par crowdsourcing et gamification

Maintenant, le défi est non seulement de maintenir l'acquisition et l'agrégation des données, mais également de concevoir des outils d'annotation pour nos algorithmes à apprendre. Les solutions sont multiples. La première consiste à générer des données synthétiques avec une vérité terrain connue, une autre est de demander aux

cliniciens ou la foule pour eux. Beaucoup de questions complexes peuvent être remplacées par des questions plus simples où la foule a assez de compétence de les répondre.

Les annotations des images par pairs que nous avons utilisées ne sont pas le seul moyen. Les grandes entreprises comme Google utilisent l'approvisionnement par la foule pour recueillir la vérité terrain pour ses produits *e.g.* Street View (reconnaissance du numéro du bâtiment) ou la traduction automatisée. Galaxy Zoo enseigne au public l'astronomie tout en collectant des annotations précieuses pour la description automatisé des galaxies. Duolingo enseigne aux gens des langues étrangères et les utilisateurs les aide à entraîner des systèmes de traduction automatique en retour, grâce à la collecte de la vérité terrain amusante (von Ahn and Dabbish, 2004). Il y a beaucoup de défis pour rendre ces systèmes plus engageante et comment utiliser la vérité terrain souvent pas parfait.

E.7.2.6 Du diagnostic au pronostic

En fin de compte, le diagnostic n'est pas la fin. Trouver des preuves à propos de " ce qui est susceptible de se produire " devrait conduire la pratique clinique. Le suivi est nécessaire pour l'imagerie cardiaque pour améliorer la qualité des soins et les résultats. Changement d'orientation du diagnostic aux pronostic et aux résultats (Timmis et al., 2015) devrait apparaître avec les patients signalant les mesures des résultats et leur qualité de vie. Si le pronostic est de remplacer le diagnostic (Croft et al., 2015), ce sera impossible sans automatisation de l'ensemble du processus de gestion des données et une récupération efficace des données avec l'appui basé sur les résultats et les avantages pour les patients.

Bibliography

Afshin, M., Ben Ayed, I., Islam, A., Goela, A., Peters, T. M., and Li, S. (2012a). Global assessment of cardiac function using image statistics in Mri. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 2):535–43. (Cited on pages 97 and 126.)

Afshin, M., Ben Ayed, I., Islam, A., Goela, A., Ross, I., Peters, T., and Li, S. (2012b). Estimation of the Cardiac Ejection Fraction from image statistics. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 824–827. IEEE. (Cited on pages 97 and 126.)

Alessandrini, M., De Craene, M., Bernard, O., Giffard-Roisin, S., Allain, P., Weese, J., Saloux, E., Delingette, H., Sermesant, M., and D' hooge, J. (2015). A pipeline for the Generation of Realistic 3D synthetic Echocardiographic Sequences: Methodology and Open-access Database. *IEEE Transactions on Medical Imaging*, pages 1–1. (Cited on pages 6, 112, 141 and 156.)

Amit, Y. and Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, 9(7):1545–1588. (Cited on pages 11, 24 and 146.)

Anderson, L. (2001). Cardiovascular T2-star magnetic resonance for the early diagnosis of myocardial iron overload. *European Heart Journal*, 22(23):2171–2179. (Cited on pages 111 and 156.)

André, B., Vercauteren, T., Buchner, A. M., Wallace, M. B., and Ayache, N. (2011a). A smart atlas for endomicroscopy using automated video retrieval. *Medical Image Analysis*, 15(4):460–476. (Cited on page 86.)

André, B., Vercauteren, T., Buchner, A. M., Wallace, M. B., and Ayache, N. (2011b). Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos. In Fichtinger, G., Martel, A., and Peters, T., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2011*, volume 6893 LNCS, pages 297–304. Springer. (Cited on pages 68 and 69.)

André, B., Vercauteren, T., Buchner, A. M., Wallace, M. B., and Ayache, N. (2012). Learning Semantic and Visual Similarity for Endomicroscopy Video Retrieval. *IEEE Transactions on Medical Imaging*, 31(6):1276–1288. (Cited on pages 69, 74, 86 and 88.)

Bachmann, C. M. (2006). Improved manifold coordinate representations of large-scale hyperspectral scenes. *Geoscience and Remote Sensing, Transactions on*, 44(10):2786–2803. (Cited on pages 116, 117 and 120.)

Bai, W., Peressutti, D., Oktay, O., Shi, W., ORegan, D. P., King, A. P., and Rueckert, D. (2015). Learning a Global Descriptor of Cardiac Motion from a Large Cohort of 1000+ Normal Subjects. In *8th International Conference on Functional Imaging and Modeling of the Heart (FIMH) 2015*, volume 1, pages 3–11. (Cited on page 126.)

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396. (Cited on page 87.)

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (Cited on pages 30 and 37.)

Bergeest, J.-P. and Jäger, F. (2008). A comparison of Five Methods for Signal Intensity Standardization in Mri. In Tolxdorff, T., Braun, J., Deserno, T. M., Horsch, A., Handels, H., Meinzer, H.-P., and Brauer, W., editors, *Bildverarbeitung für die Medizin 2008*, pages 36–40. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on page 52.)

Bernardis, E., Konukoglu, E., and Ou, Y. (2012). Temporal shape analysis via the spectral signature. In *Medical Image Computing and Computer-Assisted Intervention*. (Cited on pages 69, 77 and 87.)

Beymer, D., Syeda-Mahmood, T., and Wang, F. (2008). Exploiting spatio-temporal information for view recognition in cardiac echo videos. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. (Cited on page 22.)

Bild, D. E. (2002). Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *American Journal of Epidemiology*, 156(9):871–881. (Cited on pages 105, 107 and 151.)

Bleton, H., Margeta, J., Lombaert, H., Delingette, H., and Ayache, N. (2015). Myocardial Infarct Localization using Neighbourhood Approximation Forests. In *Statistical Atlases and Computational Models of the Heart*. (Cited on pages 12, 83, 84, 86, 90, 91, 110, 147, 150 and 154.)

Bloice, M. D., Simonic, K.-M., and Holzinger, A. (2013). On the usage of health records for the design of virtual patients: a systematic review. *BMC medical informatics and decision making*, 13(1):103. (Cited on pages 6 and 141.)

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, Heidelberg. (Cited on pages 30 and 33.)

Bottou, L. (2012). Stochastic Gradient Descent Tricks. *Neural Networks: Tricks of the Trade*, 1(1):421–436. (Cited on page 33.)

Braunwald, E. (2014). The ten advances that have defined modern cardiology. *Trends in Cardiovascular Medicine*, 24(5):179–183. (Cited on pages 5 and 141.)

Breiman, L. (1999). Random forests-random features. Technical report, University of California, Berkeley. (Cited on pages 11, 24 and 146.)

Bruder, O., Schneider, S., Nothnagel, D., Dill, T., Hombach, V., Schulz-Menger, J., Nagel, E., Lombardi, M., van Rossum, A. C., Wagner, A., Schwitter, J., Senges, J., Sabin, G. V., Sechtem, U., and Mahrholdt, H. (2009). EuroCmr (European Cardiovascular Magnetic Resonance) Registry. Results of the German Pilot Phase. *Journal of the American College of Cardiology*, 54(15):1457–1466. (Cited on pages 111 and 156.)

Bruder, O., Wagner, A., Lombardi, M., Schwitter, J., van Rossum, A., Pilz, G., Nothnagel, D., Steen, H., Petersen, S., Nagel, E., Prasad, S., Schumm, J., Greulich, S., Cagnolo, A., Monney, P., Deluigi, C. C., Dill, T., Frank, H., Sabin, G., Schneider, S., and Mahrholdt, H. (2013). European Cardiovascular Magnetic Resonance (EuroCmr) registry–multi national results from 57 centers in 15 countries. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 15(1):9. (Cited on pages 6, 107, 141 and 151.)

Buechel, E. V., Kaiser, T., Jackson, C., Schmitz, A., and Kellenberger, C. J. (2009). Normal right- and left ventricular volumes and myocardial mass in children measured by steady state free precession cardiovascular magnetic resonance. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 11:19. (Cited on page 44.)

Burggraaff, J., Dorn, J., D` Souza, M., Kamm, C. P., Morrison, C., Kontschieder, P., Tewarie, P., Steinheimer, S., Sellen, A., Criminisi, A., Dahlke, F., Kappos, L., and Uitdehaag, B. (2015). Video-based pairwise comparison is a consistent granular method of rating movement function. In *Congress of the European Committee for the Treatment and Research in Multiple Sclerosis*. (Cited on pages 72 and 80.)

Cerqueira, M. D., Imaging, C., Cerqueira, M. D., Weissman, N. J., Dilsizian, V., Jacobs, A. K., Kaul, S., Laskey, W. K., Pennell, D. J., Rumberger, J. A., Ryan, T., and Verani, M. S. (2002). Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart: A statement for Healthcare Professionals From the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. *Circulation*, 105(4):539–542. (Cited on pages 20 and 127.)

Chuang, M. L., Gona, P., Hautvast, G., Salton, C. J., Blease, S. J., Yeon, S. B., Breeuwer, M., O' Donnell, C. J., and Manning, W. J. (2011). Impact of left

ventricular trabeculations and papillary muscles on measures of cavity volume and ejection fraction. *Journal of Cardiovascular Magnetic Resonance*, 13(Suppl 1):P36. (Cited on page 44.)

Chuang, M. L., Gona, P., Hautvast, G. L. T. F., Salton, C. J., Breeuwer, M., O' Donnell, C. J., and Manning, W. J. (2014). CMr reference values for left ventricular volumes, mass, and ejection fraction using computer-aided analysis: the Framingham Heart Study. *Journal of magnetic resonance imaging : JMRI*, 39(4):895–900. (Cited on page 44.)

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220. (Cited on page 30.)

Ciresan, D. C., Meier, U., and Masci, J. (2011). Flexible, high performance convolutional neural networks for image classification. *International Joint Conference on Artificial Intelligence*, pages 1237–1242. (Cited on page 30.)

Cireşan, D. C., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 32, pages 3642–3649. (Cited on pages 22 and 30.)

Cooley, D. A. and Frazier, O. H. (2000). The Past 50 Years of Cardiovascular Surgery. *Circulation*, 102(Supplement 4):IV–87–IV–93. (Cited on pages 5 and 141.)

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. (Cited on pages 11, 35, 76 and 145.)

Criminisi, A., Juluru, K., and Pathak, S. (2011a). A discriminative-Generative Model for Detecting Intravenous Contrast in Ct Images. In *Medical Image Computing and Computer-Assisted Intervention*, pages 49–57. (Cited on page 42.)

Criminisi, A. and Shotton, J. (2011). Regression forests for efficient anatomy detection and localization in Ct studies. *Medical Computer Vision*, pages 106–117. (Cited on page 24.)

Criminisi, A., Shotton, J., and Konukoglu, E. (2011b). Decision Forests: A unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. (Cited on pages 24, 26, 40, 88, 116, 118 and 119.)

Croft, P., Altman, D. G., Deeks, J. J., Dunn, K. M., Hay, A. D., Hemingway, H., LeResche, L., Peat, G., Perel, P., Petersen, S. E., Riley, R. D., Roberts, I., Sharpe, M., Stevens, R. J., Van Der Windt, D. a., Von Korff, M., and Timmis, A. (2015). The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen"should shape clinical practice. *BMC Medicine*, 13(1):1–8. (Cited on pages 113 and 158.)

Crow, F. C. (1984). Summed-area tables for texture mapping. *ACM SIGGRAPH Computer Graphics*, 18(3):207–212. (Cited on page 51.)

Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkill Through Time : Revisiting the History of Chess. *Advances in Neural Information Processing Systems*, 20:1–8. (Cited on page 72.)

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems 27*, pages 2933–2941. (Cited on page 30.)

De Craene, M., Alessandrini, M., Allain, P., Marchesseau, S., Waechter-Stehle, I., Weese, J., Saloux, E., Morales, H. G., Cuingnet, R., Delingette, H., Sermesant, M., Bernard, O., and D' hooge, J. (2014). Generation of ultra-realistic synthetic echocardiographic sequences. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 73–76. IEEE. (Cited on pages 112 and 156.)

Decoste, D. and Schölkopf, B. (2002). Training invariant support vector machines. *Machine Learning*, 46(1-3):161–190. (Cited on pages 112 and 157.)

Deng, J., Dong, W., Socher, R., Li, L.-j., Li, K., and Fei-fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *In CVPR*. (Cited on page 30.)

Depa, M., Sabuncu, M. R. M., Holmwang, G., Nezafat, R., Schmidt, E. E. J., Golland, P., Holmvang, G., Nezafat, R., Schmidt, E. E. J., and Golland, P. (2010). Robust Atlas-Based Segmentation of Highly Variable Anatomy: Left Atrium Segmentation. In Camara, O., Pop, M., Rhode, K., Sermesant, M., Smith, N., and Young, A., editors, *Statistical Atlases and Computational Models of the Heart*, volume 6364 of *Lecture Notes in Computer Science*, pages 85–94. Springer Berlin Heidelberg. (Cited on page 59.)

Depeursinge, A., Vargas, A., Gaillard, F., Platon, A., Geissbuhler, A., Poletti, P.-A., and Müller, H. (2012). Case-based lung image categorization and retrieval for interstitial lung diseases: clinical workflows. *International journal of computer assisted radiology and surgery*, 7(1):97–110. (Cited on page 86.)

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271. (Cited on page 56.)

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2014). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *arXiv*. (Cited on pages 108 and 152.)

Donner, R., Menze, B. H., Bischof, H., and Langs, G. (2013). Fast anatomical structure localization using top-down image patch regression. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence*

and Lecture Notes in Bioinformatics), volume 7766 LNCS, pages 133–141. (Cited on page 128.)

Duchateau, N., De Craene, M., Piella, G., Silva, E., Doltra, A., Sitges, M., Bijnens, B. H., and Frangi, A. F. (2011). A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. *Medical Image Analysis*, 15(3):316–328. (Cited on pages 69 and 87.)

Ecabert, O., Peters, J., Schramm, H., Lorenz, C., von Berg, J., Walker, M. J., Vembar, M., Olszewski, M. E., Subramanyan, K., Lavi, G., and Weese, J. (2008). Automatic model-based segmentation of the heart in Ct images. *IEEE transactions on medical imaging*, 27(9):1189–201. (Cited on pages 45, 46 and 59.)

Egloff, D. (2005). Weighted P2 quantile, Boost Accumulators 1.46 (www.boost.org). (Cited on page 53.)

Eslami, A., Karamalis, A., Katouzian, A., and Navab, N. (2012). Segmentation By Retrieval With Guided Random Walks: Application To Left Ventricle Segmentation in Mri. *Medical Image Analysis*. (Cited on pages 69 and 86.)

Fahmy, A. S., Al-Agamy, A. O., and Khalifa, A. (2012). Myocardial segmentation using contour-constrained optical flow tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7085 LNCS:120–128. (Cited on page 55.)

Foncubierta-Rodríguez, A., Depeursinge, A., and Müller, H. (2012). Using Multiscale Visual Words for Lung Texture Classification and Retrieval. In Müller, H., Greenspan, H., and Syeda-Mahmood, T., editors, *Medical Content-Based Retrieval for Clinical Decision Support*, volume 7075 of *Lecture Notes in Computer Science*, pages 69–79. Springer Berlin Heidelberg. (Cited on page 68.)

Foncubierta-Rodríguez, A. and Müller, H. (2012). Ground Truth Generation in Medical Imaging A crowdsourcing –based Iterative Approach. In *Workshop on Crowdsourcing for Multimedia, ACM Multimedia*, pages 1–6. (Cited on page 74.)

Fonseca, C. G., Backhaus, M., Bluemke, D. A., Britten, R. D., Chung, J. D., Cowan, B. R., Dinov, I. D., Finn, J. P., Hunter, P. J., Kadish, A. H., Lee, D. C., Lima, J. A. C., Medrano-Gracia, P., Shivkumar, K., Suinesiaputra, A., Tao, W., and Young, A. A. (2011). The Cardiac Atlas Project- an Imaging Database for Computational Modeling and Statistical Atlases of the Heart. *Bioinformatics*, 27(16):2288–2295. (Cited on pages 6, 28, 38, 47, 98, 107, 126, 142 and 151.)

Fratz, S., Chung, T., Greil, G. F., Samyn, M. M., Taylor, A. M., Valsangiacomo Buechel, E. R., Yoo, S.-J., and Powell, A. J. (2013). Guidelines and protocols for cardiovascular magnetic resonance in children and adults with congenital heart disease: ScmR expert consensus group on congenital heart disease. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 15(1):51. (Cited on page 44.)

Friedrich, M. G., Bucciarelli-Ducci, C., White, J. a., Plein, S., Moon, J. C., Almeida, A. G., Kramer, C. M., Neubauer, S., Pennell, D. J., Petersen, S. E., Kwong, R. Y., Ferrari, V. a., Schulz-Menger, J., Sakuma, H., Schelbert, E. B., Larose, É., Eitel, I., Carbone, I., Taylor, A. J., Young, A., de Roos, A., and Nagel, E. (2014). Simplifying cardiovascular magnetic resonance pulse sequence terminology. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 16:3960. (Cited on pages 7, 11, 42, 142 and 146.)

Friedrich, M. G., Larose, E., Patton, D., Dick, A., Merchant, N., and Paterson, I. (2012). Canadian Society for Cardiovascular Magnetic Resonance (CanScmR) Recommendations for Cardiovascular Magnetic Resonance Image Analysis and Reporting. (Cited on page 44.)

Fukushima, K. (1980). Neocognitron: A self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological cybernetics*, (36):193–202. (Cited on pages 11, 29 and 146.)

Fuster, V. and Kelly, B. B., editors (2010). *Promoting Cardiocascular Health in the Developing World: A Critical Challenge to Achieve Global Health*. Washington, D.C.: National Academies Press. (Cited on pages 6 and 141.)

Galderisi, M., Cardim, N., D' Andrea, A., Bruder, O., Cosyns, B., Davin, L., Donal, E., Edvardsen, T., Freitas, A., Habib, G., Kitsiou, A., Plein, S., Petersen, S. E., Popescu, B. a., Schroeder, S., Burgstahler, C., Lancellotti, P., Document Reviewers, Sicari, R., Muraru, D., Lombardi, M., Dulgheru, R., and Gerche, a. L. (2015). The multi-modality cardiac imaging approach to the Athlete's heart: an expert consensus of the European Association of Cardiovascular Imaging. *European Heart Journal - Cardiovascular Imaging*. (Cited on pages 111 and 155.)

Geremia, E., Clatz, O., Menze, B. H., Konukoglu, E., Criminisi, A., and Ayache, N. (2011). Spatial decision forests for Ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390. (Cited on pages 12, 24, 46, 48, 49, 60, 108, 147 and 152.)

Geremia, E., Menze, B. H., Prastawa, M., Weber, M., Criminisi, A., and Ayache, N. (2013). Brain Tumor Cell Density Estimation from Multi-modal Mr Images Based on a Synthetic Tumor Growth Model. *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, 7766:273–282. (Cited on pages 112 and 157.)

Glatard, T., Lartizien, C., Gibaud, B., Da Silva, R. F., Forestier, G., Cervenansky, F., Alessandrini, M., Benoit-Cattin, H., Bernard, O., Camarasu-Pop, S., Cerezo, N., Clarysse, P., Gaignard, A., Hugonnard, P., Liebgott, H., Marache, S., Marion, A., Montagnat, J., Tabary, J., and Friboulet, D. (2013). A virtual imaging platform for multi-modality medical image simulation. *IEEE Transactions on Medical Imaging*, 32(1):110–118. (Cited on pages 6, 112, 141 and 157.)

Glatard, T., Montagnat, J., and Magnin, I. E. (2004). Texture based medical image indexing and retrieval : application to cardiac imaging. In *MIR '04 Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 135 – 142. (Cited on pages 69 and 86.)

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 9:249–256. (Cited on page 30.)

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223. (Cited on pages 11, 129 and 146.)

Gray, K. R., Aljabar, P., Heckemann, R. a., Hammers, A., and Rueckert, D. (2011). Random forest-based manifold learning for classification of imaging data in dementia. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7009 LNCS, pages 159–166. Springer. (Cited on page 24.)

Gray, K. R., Aljabar, P., Heckemann, R. a., Hammers, A., and Rueckert, D. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, 65:167–175. (Cited on page 88.)

Gudmundsson, P., Rydberg, E., Winter, R., and Willenheimer, R. (2005). Visually estimated left ventricular ejection fraction by echocardiography is closely correlated with formal quantitative methods. *International Journal of Cardiology*, 101(2):209–212. (Cited on page 99.)

Gueld, M. O., Kohnen, M., Keysers, D., Schubert, H., Wein, B. B., Bredno, J., and Lehmann, T. M. (2002). Quality of DicOm header information for image categorization. In *Proceedings of SPIE*, volume 4685, pages 280–287. (Cited on pages 19 and 85.)

Han, Y., Olson, E., Maron, M. S., Manning, W. J., and Yeon, S. B. (2008). 2075 Papillary muscles and trabeculations significantly impact ventricular volume, ejection fraction, and regurgitation assessment by cardiovascular magnetic resonance in patients with hypertrophic cardiomyopathy. *Journal of Cardiovascular Magnetic Resonance*, 10(Suppl 1):A344. (Cited on page 44.)

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arxiv*, pages 1–12. (Cited on pages 30, 113 and 157.)

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621. (Cited on page 78.)

Harrigan, C. J., Appelbaum, E., Maron, B. J., Buros, J. L., Gibson, C. M., Lesser, J. R., Udelson, J. E., Manning, W. J., and Maron, M. S. (2008). Significance of Papillary Muscle Abnormalities Identified by Cardiovascular Magnetic Resonance in Hypertrophic Cardiomyopathy. *American Journal of Cardiology*, 101(5):668–673. (Cited on pages 8 and 143.)

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *ArXiv e-prints*. (Cited on pages 30 and 37.)

Heimann, T., Mountney, P., John, M., and Ionasec, R. (2014). Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data. *Medical Image Analysis*, 18(8):1320–1328. (Cited on pages 112 and 157.)

Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkill (Tm): A bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, pages 569–576. (Cited on pages 72 and 80.)

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv: 1207.0580*, pages 1–18. (Cited on pages 22, 30, 33 and 35.)

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282. (Cited on pages 11, 24 and 146.)

Hoogendoorn, C., Whitmarsh, T., Duchateau, N., Sukno, F. M., De Craene, M., and Frangi, A. F. (2010). A groupwise mutual information metric for cost efficient selection of a suitable reference in cardiac computational atlas construction. In *SPIE Medical Imaging*, volume 7962, pages 76231R–76231R–9. (Cited on page 54.)

Huang, W., Zhang, P., and Wan, M. (2013). A novel similarity learning method via relative comparison for content-based medical image retrieval. *Journal of Digital Imaging*, 26(5):850–865. (Cited on page 75.)

Hudsmith†, L., Petersen†, S., Francis, J., Robson, M., and Neubauer, S. (2005). Normal Human Left and Right Ventricular and Left Atrial Dimensions Using Steady State Free Precession Magnetic Resonance Imaging. *Journal of Cardiovascular Magnetic Resonance*, 7(5):775–782. (Cited on pages 8, 44 and 143.)

Iglesias, J. E., Konukoglu, E., Montillo, A., Tu, Z., and Criminisi, A. (2011). Combining Generative and Discriminative Models for Semantic Segmentation of Ct Scans via Active Learning. In Székely, G. and Hahn, H. K., editors, *Information Processing in Medical Imaging*, volume 6801 of *LNCS*. Springer. (Cited on page 54.)

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Arxiv*. (Cited on pages 30 and 42.)

Jacobs, S., Grunert, R., Mohr, F. W., and Falk, V. (2008). 3D-imaging of cardiac structures using 3D heart models for planning in heart surgery: a preliminary study. *Interactive cardiovascular and thoracic surgery*, 7(1):6–9. (Cited on pages 6 and 141.)

Jain, R. and Chlamtac, I. (1985). The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Communications of the ACM*, 28(10):1076–1085. (Cited on page 53.)

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*. (Cited on pages 34 and 35.)

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, pages 133–142. (Cited on page 75.)

Jolly, M.-P., Guetter, C., Lu, X., Xue, H., and Guehring, J. (2012). Automatic Segmentation of the Myocardium in Cine Mr Images Using Deformable Registration. volume 7085, pages 98–108. (Cited on page 55.)

Kadish, A. H., Bello, D., Finn, J. P., Bonow, R. O., Schaechter, A., Subacius, H., Albert, C., Daubert, J. P., Fonseca, C. G., and Goldberger, J. J. (2009). Rationale and design for the Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation (DetErmIne) trial. *Journal of cardiovascular electrophysiology*, 20(9):982–7. (Cited on pages 28, 37, 38, 70, 78, 88, 98, 107 and 151.)

Kali, A., Cokic, I., Tang, R. L. Q., Yang, H.-J., Sharif, B., Marbán, E., Li, D., Berman, D. S., and Dharmakumar, R. (2014). Determination of location, size, and transmurality of chronic myocardial infarction without exogenous contrast media by using cardiac magnetic resonance imaging at 3 T. *Circ Cardiovasc Imag*, 7(3):471–81. (Cited on pages 111 and 156.)

Kang, K., Loy, C. C., and Wang, X. (2015). Deeply Learned Attributes for Crowded Scene Understanding. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. (Cited on pages 110 and 154.)

Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., and Winnemoeller, H. (2014). Recognizing Image Style. In Valstar, M., French, A., and Pridmore, T., editors, *Proceedings of the British Machine Vision Conference*, pages 1–20. BMVA Press. (Cited on pages 22 and 34.)

Karim, R., Mohiaddin, R., and Rueckert, D. (2007). Automatic Segmentation of the Left Atrium. In *Medical Image Under standing And Analysis Conference*. (Cited on pages 59 and 61.)

Karpathy, A., Li, F.-F., and Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. (Cited on pages 108 and 152.)

Keraudren, K., Oktay, O., Shi, W., Hajnal, J. V., and Rueckert, D. (2014). Endocardial 3D ultrasound Segmentation using Autocontext Random Forests. In *Challenge on Endocardial Three-dimensional Ultrasound Segmentation - MICCAI 2014*. (Cited on page 66.)

Kim, M. S., Hansgen, A. R., and Carroll, J. D. (2008). Use of Rapid Prototyping in the Care of Patients with Structural Heart Disease. *Trends in Cardiovascular Medicine*, 18(6):210–216. (Cited on pages 6 and 141.)

Klinke, V., Muzzarelli, S., Lauriers, N., Locca, D., Vincenti, G., Monney, P., Lu, C., Nothnagel, D., Pilz, G., Lombardi, M., van Rossum, A. C., Wagner, A., Bruder, O., Mahrholdt, H., and Schwitter, J. (2013). Quality assessment of cardiovascular magnetic resonance in the setting of the European Cmr registry: description and validation of standardized criteria. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 15(1):55. (Cited on pages 78 and 125.)

Kontschieder, P., Dorn, J., Morrison, C., Corish, R., Zikic, D., Sellen, A. K., DSouza, M., Kamm, C., Burggraaff, J., Tewarie, P., Vogel, T., Azzarito, M., Chin, P., Dahlke, F., Polman, C., Kappos, L., Uitdehaag, B., and Criminisi, A. (2014). Quantifying Progression of Multiple Sclerosis via Classification of Depth Videos. In *Medical Image Computing and Computer-Assisted Intervention*. (Cited on pages 94, 101 and 105.)

Kontschieder, P., Fiterau, M., Criminisi, A., Bul, S. R., and Kessler, F. B. (2015). Deep Neural Decision Forests. In *International Conference on Computer Vision*. (Cited on pages 66 and 105.)

Konukoglu, E., Criminisi, A., Pathak, S., Robertson, D., White, S., Haynor, D., and Siddiqui, K. (2011). Robust linear registration of Ct images using random regression forests. In *SPIE Medical Imaging*, volume 7962, pages 79621X–79621X–8. (Cited on page 54.)

Konukoglu, E., Glocker, B., Criminisi, A., and Pohl, K. M. (2012). WEsd - Weighted Spectral Distance for Measuring Shape Dissimilarity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 1–35. (Cited on pages 77, 87 and 116.)

Konukoglu, E., Glocker, B., Zikic, D., and Criminisi, A. (2013). Neighbourhood approximation using randomized forests. *Medical image analysis*, 17(7):790–804. (Cited on pages 12, 84, 89, 91, 147 and 150.)

Kovashka, A., Parikh, D., and Grauman, K. (2012). WhittleSearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980. IEEE. (Cited on page 80.)

Kramer, C. M., Barkhausen, J., Flamm, S. D., Kim, R. J., and Nagel, E. (2008). Standardized cardiovascular magnetic resonance imaging (Cmr) protocols, society for cardiovascular magnetic resonance: board of trustees task force on standardized protocols. *Journal of Cardiovascular Magnetic Resonance*, 10(1):35. (Cited on pages 111 and 156.)

Kramer, C. M., Barkhausen, J., Flamm, S. D., Kim, R. J., and Nagel, E. (2013). Standardized cardiovascular magnetic resonance (Cmr) protocols 2013 update. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 15(1):91. (Cited on pages 42 and 44.)

Krishnan, K., Ibanez, L., Turner, W., and Avila, R. (2009). Algorithms, architecture, validation of an open source toolkit for segmenting Ct lung lesions. In Brown, M., de Bruijne, M., van Ginneken, B., Kiraly, A., Kuhnigk, J.-M., Lorenz, C., McClelland, J. R., Mori, K., Reeves, A., and Reinhardt, J. M., editors, *MICCAI Workshop on Pulmonary Image Analysis*, pages 365–375. CreateSpace. (Cited on page 58.)

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. (Cited on pages 22, 30 and 33.)

Kules, B., Capra, R., Banta, M., and Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface? *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 313–322. (Cited on page 85.)

Kutra, D., Saalbach, A., Lehmann, H., Groth, A., Dries, S. P. M., Krueger, M. W., Dössel, O., and Weese, J. (2012). Automatic multi-model-based segmentation of the left atrium in cardiac Mri scans. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 2):1–8. (Cited on page 59.)

Lamata, P., Casero, R., Carapella, V., Niederer, S. a., Bishop, M. J., Schneider, J. E., Kohl, P., and Grau, V. (2014). Images as drivers of progress in cardiac computational modelling. (Cited on pages 6, 59 and 141.)

Langs, G., Hanbury, A., Menze, B., and Müller, H. (2013). VIscEraL: Towards large data in medical imaging - Challenges and directions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7723 LNCS:92–98. (Cited on pages 6 and 142.)

Le Folgoc, L., Delingette, H., Criminisi, A., and Ayache, N. (2013). Current-based 4D shape analysis for the mechanical personalization of heart models. In *Lecture*

Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 7766 LNCS, pages 283–292. (Cited on page 69.)

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*. (Cited on pages 11, 29 and 146.)

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324. (Cited on page 36.)

Leistner, C., Saffari, A., Santner, J., and Bischof, H. (2009). Semi-supervised random forests. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 506–513. (Cited on page 116.)

Lempitsky, V., Verhoek, M., Noble, J., and Blake, A. (2009). Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3D echocardiography. In Ayache, N., Delingette, H., and Sermesant, M., editors, *Functional Imaging and Modeling of the Heart*, volume 5528 of *LNCS*, pages 447–456. Springer. (Cited on pages 46, 49, 50, 53, 108 and 152.)

Li, B., Liu, Y., Occleshaw, C. J., Cowan, B. R., and Young, A. a. (2010). In-line automated tracking for ventricular function with magnetic resonance imaging. *JACC: Cardiovascular Imaging*, 3(8):860–866. (Cited on page 55.)

Lin, T.-y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft CocO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *European Conference on Computer Vision –ECCV 2014*, volume 8693, pages 740–755. Springer. (Cited on page 30.)

Lombaert, H., Grady, L., Pennec, X., Ayache, N., and Cheriet, F. (2012). Spectral demons - Image registration via global spectral correspondence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7573 LNCS, pages 30–44. (Cited on pages 77 and 126.)

Lombaert, H., Zikic, D., Criminisi, A., and Ayache, N. (2014). Laplacian Forests: Semantic Image Segmentation by Guided Bagging. In *Medical Image Computing and Computer-Assisted Intervention –MICCAI 2014 - 17th International Conference*, pages 496–504. (Cited on pages 48 and 66.)

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. (Cited on page 66.)

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110. (Cited on page 68.)

Lu, X., Wang, Y., Georgescu, B., Littman, A., and Comaniciu, D. (2011). Automatic Delineation of Left and Right Ventricles in Cardiac Mri Sequences Using a Joint Ventricular Model. In Metaxas, D. and Axel, L., editors, *Functional Imaging and Modeling of the Heart*, volume 6666 of *LNCS*, pages 250–258. Springer. (Cited on page 46.)

Maceira, a. M., Prasad, S. K., Khan, M., and Pennell, D. J. (2006). Normalized left ventricular systolic and diastolic function by steady state free precession cardiovascular magnetic resonance. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 8(3):417–426. (Cited on pages 8, 44 and 143.)

Mahapatra, D. (2014). Automatic Cardiac Segmentation Using Semantic Information from Random Forests. *Journal of Digital Imaging*, 27(6):794–804. (Cited on page 66.)

Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H. G., Eisenmann, M., and Speidel, S. (2014). Can Masses of Non-Experts Train Highly Accurate Image Classifiers? In Golland, P., Hata, N., Barillot, C., Hornegger, J., and Howe, R., editors, *Medical Image Computing and Computer-Assisted Intervention –MICCAI 2014*, volume 8674, pages 438–445. Springer, Boston, MA. (Cited on page 74.)

Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905. (Cited on page 58.)

Malik, J. and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America. A, Optics and image science*, 7(5):923–932. (Cited on page 30.)

Mantilla, J., Garreau, M., Bellanger, J.-J., and Paredes, J. L. (2013). Machine Learning Techniques for Lv Wall Motion Classification Based on Spatio-temporal Profiles from Cardiac Cine Mri. *2013 12th International Conference on Machine Learning and Applications*, 2(04):167–172. (Cited on page 69.)

Margeta, J., Bleton, H., Lee, D. C., Criminisi, A., and Ayache, N. (2015a). Learning to retrieve semantically similar hearts. *in preparation for submission*. (Cited on page 83.)

Margeta, J., Cabrera Lozoya, R., Lee, D. C., Criminisi, A., and Ayache, N. (2015b). Crowdsourcing cardiac attributes. *in preparation for submission*. (Cited on page 67.)

Margeta, J., Criminisi, A., and Ayache, N. (2013). Decision forests for segmentation of left atria from 3D mRi . *Statistical Atlases and Computational Models of the*

Heart. Imaging and Modelling Challenges, pages 2–5. (Cited on pages 43, 109 and 153.)

Margeta, J., Criminisi, A., Cabrera Lozoya, R., Lee, D., and Ayache, N. (2015c). Fine-tuned convolutional neural nets for cardiac Mri acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging \& Visualization*, pages 1–11. (Cited on pages 17, 107 and 151.)

Margeta, J., Criminisi, A., Lee, D. C., and Ayache, N. (2014). Recognizing cardiac magnetic resonance acquisition planes. In *Conference on Medical Image Understanding and Analysis (MIUA 2014)*, Egham. Reyes-Aldasoro, Constantino Carlos and Slabaugh, Gregory. (Cited on pages 17, 23, 24, 29, 38, 107 and 151.)

Margeta, J., Geremia, E., Criminisi, A., and Ayache, N. (2012). Layered Spatio-temporal Forests for Left Ventricle Segmentation from 4D cardiac Mri Data. In Camara, O., Konukoglu, E., Pop, M., Rhode, K., Sermesant, M., and Young, A., editors, *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges*, volume 7085 of *Lecture Notes in Computer Science*, pages 109–119. Springer Berlin / Heidelberg, Toronto. (Cited on pages 43, 109 and 153.)

Markonis, D., Holzer, M., Dungs, S., Vargas, a., Langs, G., Kriewel, S., and Müller, H. (2012). A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548. (Cited on pages 78 and 84.)

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied*. (Cited on page 122.)

McLeod, K., Sermesant, M., Beerbaum, P., and Pennec, X. (2015). Spatio-Temporal Tensor Decomposition of a Polyaffine Motion Model for a Better Analysis of Pathological Left Ventricular Dynamics. *IEEE Transactions on Medical Imaging*, 34(7):1562–1575. (Cited on page 69.)

Medrano-Gracia, P., Cowan, B. R., Bluemke, D. a., Finn, J. P., Kadish, A. H., Lee, D. C., Lima, J. A., Suinesiaputra, A., and Young, A. a. (2013). Atlas-based analysis of cardiac shape and function: correction of regional shape bias due to imaging protocol for population studies. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 15(1):80. (Cited on page 105.)

Medrano-Gracia, P., Cowan, B. R., Suinesiaputra, A., and Young, A. a. (2014). Atlas-based anatomical modeling and analysis of heart disease. (Cited on pages 69 and 87.)

Medrano-Gracia, P., Cowan, B. R., Suinesiaputra, A., and Young, A. a. (2015). Challenges of Cardiac Image Analysis in Large-Scale Population-Based Studies. *Current Cardiology Reports*, 17. (Cited on pages 8 and 144.)

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. a. (2011). On oblique random forests. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6912 LNAI, pages 453–469. (Cited on page 40.)

Mihl, C., Dassen, W. R. M., and Kuipers, H. (2008). Cardiac remodelling: concentric versus eccentric hypertrophy in strength and endurance athletes. *Netherlands Heart Journal*, 16(4):129–133. (Cited on page 70.)

Naehle, C. P., Kreuz, J., Strach, K., Schwab, J. O., Pingel, S., Luechinger, R., Fimmers, R., Schild, H., and Thomas, D. (2011). Safety, feasibility, and diagnostic value of cardiac magnetic resonance imaging in patients with cardiac pacemakers and implantable cardioverters/defibrillators at 1.5 T. *American heart journal*, 161(6):1096–105. (Cited on page 125.)

Natori, S., Lai, S., Finn, J. P., Gomes, A. S., Hundley, W. G., Jerosch-Herold, M., Pearson, G., Sinha, S., Arai, A., Lima, J. a. C., and Bluemke, D. a. (2006). Cardiovascular function in multi-ethnic study of atherosclerosis: Normal values by age, sex, and ethnicity. *American Journal of Roentgenology*, 186(6 SUPPL. A). (Cited on pages 8 and 143.)

Nichols, M., Townsend, N., Luengo-Fernandez, R., Leal, J., Gray, A., Scarborough, P., Rayner, M., and Luengo-Fernandez, R. (2012). European Cardiovascular Disease Statistics 2012. Technical report. (Cited on pages 6 and 141.)

Nyúl, L. G. and Udupa, J. K. (1999). On standardizing the Mr image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081. (Cited on pages 51, 52 and 60.)

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of Mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150. (Cited on page 52.)

O' Donovan, P., Lībeks, J., Agarwala, A., and Hertzmann, A. (2014). Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics*, 33(4):1–9. (Cited on page 68.)

Otey, M., Bi, J., Krishna, S., and Rao, B. (2006). Automatic view recognition for cardiac ultrasound images. In *International Workshop on Computer Vision for Intravascular and Intracardiac Imaging*, pages 187–194. (Cited on page 22.)

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, C(1):62–66. (Cited on pages 61 and 92.)

Ou, Y., Ye, D. H., Pohl, K. M., and Davatzikos, C. (2012). Validation of DraMms among 12 popular methods in cross-subject cardiac Mri registration. *Lecture*

Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7359 LNCS:209–219. (Cited on pages 45 and 125.)

Ourselin, S., Roche, A., Prima, S., and Ayache, N. (2000). Block matching: A general framework to improve robustness of rigid registration of medical images. In Delp, S., DiGoia, A., and Jaramaz, B., editors, *Medical Image Computing and Computer-Assisted Intervention 2000*, volume 1935 of *LNCS*. Springer. (Cited on page 54.)

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. (Cited on pages 112 and 156.)

Papavassiliu, T., Kühl, H. P., Schröder, M., Süselbeck, T., Bondarenko, O., Böhm, C. K., Beek, A., Hofman, M. M. B., and van Rossum, A. C. (2005). Effect of endocardial trabeculae on left ventricular measurements and measurement reproducibility at cardiovascular Mr imaging. *Radiology*, 236(1):57–64. (Cited on page 44.)

Parikh, D. and Grauman, K. (2011). Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, number Iccv, pages 503–510. Ieee. (Cited on pages 75 and 76.)

Park, J. H., Zhou, S. K., Simopoulos, C., Otsuki, J., and Comaniciu, D. (2007). Automatic cardiac view classification of echocardiogram. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8. Ieee. (Cited on page 22.)

Pauly, O. (2012). *Random Forests for Medical Applications*. PhD thesis, Technical University Munich. (Cited on pages 24 and 26.)

Pauly, O., Glocker, B., and Criminisi, A. (2011). Fast multiple organ detection and localization in whole-body Mr Dixon sequences. *Medical Image Computing and Computer-Assisted Intervention*, pages 239–247. (Cited on page 24.)

Pavani, S.-K., Delgado, D., and Frangi, A. F. (2010). Haar-like features with optimally weighted rectangles for rapid object detection. *Pattern Recognition*, 43(1):160–172. (Cited on page 51.)

Pedregosa, F. and Gramfort, A. (2012). Learning to rank from medical imaging data. In *Third International Workshop on Machine Learning in Medical Imaging - MLMI 2012*. (Cited on page 75.)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. (Cited on page 28.)

Pei, Y., Kim, T.-K., and Zha, H. (2013). Unsupervised Random Forest Manifold Alignment for Lipreading. *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 129–136. (Cited on page 88.)

Petersen, S. E., Matthews, P. M., Bamberg, F., Bluemke, D. a., Francis, J. M., Friedrich, M. G., Leeson, P., Nagel, E., Plein, S., Rademakers, F. E., Young, A. a., Garratt, S., Peakman, T., Sellors, J., Collins, R., and Neubauer, S. (2013). Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of Uk Biobank - rationale, challenges and approaches. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 15(1):46. (Cited on pages 6, 107, 141, 142 and 151.)

Petitjean, C., Dacher, J.-n., Petitjean, C., and A, J.-n. D. (2011). A review of segmentation methods in short axis cardiac Mr images . A review of segmentation methods in short axis cardiac Mr images. (Cited on page 45.)

Peyrat, J.-M., Delingette, H., Sermesant, M., Xu, C., and Ayache, N. (2010). Registration of 4D cardiac Ct sequences under trajectory constraints with multichannel diffeomorphic demons. *IEEE transactions on medical imaging*, 29(7):1351–1368. (Cited on page 94.)

Prakosa, A., Sermesant, M., Allain, P., Villain, N., Rinaldi, C. A., Rhode, K., Razavi, R., Delingette, H., and Ayache, N. (2014). Cardiac electrophysiological activation pattern estimation from images using a patient-specific database of synthetic image sequences. *IEEE Transactions on Biomedical Engineering*, 61(2):235–245. (Cited on pages 112 and 157.)

Prakosa, A., Sermesant, M., Delingette, H., Marchesseau, S., Saloux, E., Allain, P., Villain, N., and Ayache, N. (2013). Generation of synthetic but visually realistic time series of cardiac images combining a biophysical model and clinical images. *IEEE Transactions on Medical Imaging*, 32:99–109. (Cited on pages 6, 112, 141, 156 and 157.)

Ren, S., Cao, X., Wei, Y., and Sun, J. (2014). Face Alignment at 3000 Fps via Regressing Local Binary Features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, volume 1, pages 1685 – 1692, Columbus, OH. IEEE. (Cited on pages 126 and 129.)

Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Adams, R. J., Berry, J. D., Brown, T. M., Carnethon, M. R., Dai, S., de Simone, G., Ford, E. S., Fox, C. S., Fullerton, H. J., Gillespie, C., Greenlund, K. J., Hailpern, S. M., Heit, J. a., Ho, P. M., Howard, V. J., Kissela, B. M., Kittner, S. J., Lackland, D. T., Lichtman, J. H., Lisabeth, L. D., Makuc, D. M., Marcus, G. M., Marelli, A., Matchar, D. B.,

McDermott, M. M., Meigs, J. B., Moy, C. S., Mozaffarian, D., Mussolino, M. E., Nichol, G., Paynter, N. P., Rosamond, W. D., Sorlie, P. D., Stafford, R. S., Turan, T. N., Turner, M. B., Wong, N. D., and Wylie-Rosett, J. (2011). Heart disease and stroke statistics–2011 update: a report from the American Heart Association. *Circulation*, 123(4):e18–e209. (Cited on pages 6 and 141.)

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. page 37. (Cited on pages 22, 29, 30 and 34.)

Sagristà-Sauleda, J., Mercé, A. S., and Soler-Soler, J. (2011). Diagnosis and management of pericardial effusion. *World journal of cardiology*, 3(5):135–143. (Cited on page 71.)

Saremi, F., Grizzard, J. D., and Kim, R. J. (2008). Optimizing cardiac Mr imaging: practical remedies for artifacts. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 28(4):1161–87. (Cited on page 125.)

Sato, Y., Nakajima, S., Atsumi, H., Koller, T., Gerig, G., Yoshida, S., and Kikinis, R. (1997). 3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. In Troccaz, J., Grimson, E., and Mösges, R., editors, *CVRMed-MRCAS'97 SE - 22*, volume 1205 of *Lecture Notes in Computer Science*, pages 213–222. Springer Berlin Heidelberg. (Cited on page 62.)

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117. (Cited on page 29.)

Schulz-Menger, J., Bluemke, D. a., Bremerich, J., Flamm, S. D., Fogel, M. a., Friedrich, M. G., Kim, R. J., Knobelsdorff-Brenkenhoff, F. V., Kramer, C. M., Pennell, D. J., Plein, S., and Nagel, E. (2013). Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (ScmR) Board of Trustees Task Force on Standardized Post Processing. *Journal of Cardiovascular Magnetic Resonance*, 15(1):35. (Cited on pages 8, 44, 59 and 143.)

Sethian, J. A. (1996). A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of*. (Cited on page 116.)

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., and Arbel, T. (2010). Evaluating intensity normalization on Mris of human brain with multiple sclerosis. *Medical image analysis*, 15(2):267–282. (Cited on page 52.)

Shaker, M. S., Wael, M., Yassine, I. A., and Fahmy, A. S. (2014). Cardiac Mri view classification using autoencoder. In *Biomedical Engineering Conference (CIBEC), 2014 Cairo International*, pages 125–128. (Cited on page 22.)

Sharif, A., Hossein, R., Josephine, A., Stefan, S., and Royal, K. T. H. (2014). CNn Features off-the-shelf : an Astounding Baseline for Recognition. In *CVPRW '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519. (Cited on pages 22 and 34.)

Shi, W., Zhuang, X., Wang, H., Duckett, S., Oregan, D., Edwards, P., Ourselin, S., and Rueckert, D. (2011). Automatic Segmentation of Different Pathologies from Cardiac Cine Mri Using Registration and Multiple Component Em Estimation. In Metaxas, D. and Axel, L., editors, *Functional Imaging and Modeling of the Heart*, volume 6666 of *LNCS*, pages 163–170. Springer. (Cited on pages 45, 46, 51 and 53.)

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *In CVPR*, volume 2, page 3. (Cited on pages 112 and 157.)

Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. (Cited on pages 12, 45, 108, 147 and 152.)

Shotton, J., Sharp, T., and Kohli, P. (2013). Decision Jungles: Compact and Rich Models for Classification. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 234–242. Curran Associates, Inc. (Cited on page 105.)

Siemens (2010). MRi Acronyms. *Siemens Healthcare*, pages 1–6. (Cited on pages 7 and 142.)

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR 2015*, pages 1–10. (Cited on page 42.)

Singh, A. and Nowak, R. (2008). Unlabeled data: Now it helps, now it doesn't. *Advances in Neural Information Processing*, pages 1–8. (Cited on page 116.)

Sivic, J. and Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. *Proceedings Ninth IEEE International Conference on Computer Vision*, (Iccv):2–9. (Cited on page 68.)

Stearns, M. Q., Price, C., Spackman, K. a., and Wang, a. Y. (2001). SNomEd clinical terms: overview of the development process and project status. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 662–666. (Cited on pages 7 and 143.)

Suinesiaputra, A., Bluemke, D. a., Cowan, B. R., Friedrich, M. G., Kramer, C. M., Kwong, R., Plein, S., Schulz-Menger, J., Westenberg, J. J. M., Young, A. a., and Nagel, E. (2015). Quantification of Lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *Journal of Cardiovascular Magnetic Resonance*, 17(1):63. (Cited on pages 8, 44 and 143.)

Suinesiaputra, A., Cowan, B., Medrano-Gracia, P., and Young, A. (2014a). Big Heart Data: Advancing Health Informatics through Data Sharing in Cardiovascular Imaging. *IEEE Journal of Biomedical and Health Informatics*, 2194(c):1–1. (Cited on pages 6 and 141.)

Suinesiaputra, A., Cowan, B. R., Al-Agamy, A. O., Elattar, M. A., Ayache, N., Fahmy, A. S., Khalifa, A. M., Medrano-Gracia, P., Jolly, M. P., Kadish, A. H., Lee, D. C., Margeta, J., Warfield, S. K., and Young, A. A. (2014b). A collaborative resource to build consensus for automated left ventricular segmentation of cardiac Mr images. *Medical Image Analysis*, 18(1):50–62. (Cited on pages 43, 55, 56, 57, 58, 109 and 153.)

Suinesiaputra, A., Üzümcü, M., Frangi, A. F., Kaandorp, T. A. M., Reiber, J. H. C., and Lelieveldt, B. P. F. (2004). Detecting Regional Abnormal Cardiac Contraction in Short-Axis Mr Images Using Independent Component Analysis. In *Medical Image Computing and Computer-Assisted Intervention*, pages 737–744. (Cited on page 92.)

Sun, Y. (2009). Myocardial repair/remodelling following infarction: roles of local factors. *Cardiovascular research*, 81(3):482–90. (Cited on page 91.)

Swee, J. K. and Grbić, S. (2014). Advanced Transcatheter Aortic Valve Implantation (TavI) Planning from Ct with ShapeForest. *Medical Image Computing and Computer-Assisted Intervention –MICCAI 2014*, 8674:1–8. (Cited on page 24.)

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv preprint arXiv:1409.4842*, (2):1879–1886. (Cited on page 42.)

Taigman, Y., Yang, M., Ranzato, M. A., and Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8. (Cited on pages 30 and 94.)

Tamuz, O. and Belongie, S. (2011). Adaptively Learning the Crowd Kernel. (Cited on page 80.)

Taylor, A. M. and Bogaert, J. (2012). Cardiovascular Mr Imaging Planes and Segmentation. In Bogaert, J., Dymarkowski, S., Taylor, A. M., and Muthurangu, V., editors, *Clinical Cardiac MRI SE - 333*, Medical Radiology, pages 93–107. Springer Berlin Heidelberg. (Cited on pages 19, 108 and 152.)

Timmis, a., Flather, M., and Gale, C. (2015). European Heart Journal - Quality of Care and Clinical Outcomes: a new journal for the 21st century. *European Heart Journal - Quality of Care and Clinical Outcomes*, pages 16–17. (Cited on pages 113 and 158.)

Tobon-Gomez, C., De Craene, M., McLeod, K., Tautz, L., Shi, W., Hennemuth, A., Prakosa, A., Wang, H., Carr-White, G., Kapetanakis, S., Lutz, A., Rasche, V.,

Schaeffter, T., Butakoff, C., Friman, O., Mansi, T., Sermesant, M., Zhuang, X., Ourselin, S., Peitgen, H.-O., Pennec, X., Razavi, R., Rueckert, D., Frangi, a. F., and Rhode, K. S. (2013a). Benchmarking framework for myocardial tracking and deformation algorithms: An open access database. *Medical image analysis*, 17(6):632–648. (Cited on page 38.)

Tobon-Gomez, C., Geers, A. J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z., Stender, B., Zheng, Y., Zuluaga, M. a., Betancur, J., Ayache, N., Chikh, M. A., Dillenseger, J.-L., Kelm, B. M., Mahmoudi, S., Ourselin, S., Schlaefer, A., Schaeffter, T., Razavi, R., and Rhode, K. S. (2015). Benchmark for Algorithms Segmenting the Left Atrium From 3D cT and Mri Datasets. *IEEE Transactions on Medical Imaging*, 34(7):1460–1473. (Cited on pages 43, 60, 64, 109 and 153.)

Tobon-Gomez, C., Peters, J., Weese, J., Pinto, K., Karim, R., Schaeffter, T., Razavi, R., and Rhode, K. S. (2013b). Left Atrial Segmentation Challenge : a unified benchmarking framework. In *Statistical atlases and computational models of the heart*. (Cited on page 60.)

Tobon-Gomez, C., Sukno, F. M., Bijnens, B. H., Huguet, M., and Frangi, a. F. (2011). Realistic simulation of cardiac magnetic resonance studies modeling anatomical variability, trabeculae, and papillary muscles. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 65:280–288. (Cited on pages 112 and 156.)

Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 Million Tiny Images: a Large Data Set for Nonparametric Object and Scene Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–70. (Cited on page 26.)

Tu, Z. and Bai, X. (2010). Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 32(10):1744–57. (Cited on page 46.)

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and Tell: A neural Image Caption Generator. (Cited on pages 101, 108 and 152.)

Viola, P. and Jones, M. (2004). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518. IEEE Comput. Soc. (Cited on pages 45, 51, 53 and 95.)

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, pages 319–326. (Cited on pages 80, 113 and 158.)

Wang, H., Shi, W., Bai, W., de Marvao, A. M. S. M., Dawes, T. J. W., ORegan, D. P., Edwards, P., Cook, S., and Rueckert, D. (2015). Prediction of Clinical Information from Cardiac Mri Using Manifold Learning. In *8th International Conference on Functional Imaging and Modeling of the Heart (FIMH) 2015*, volume 1, pages 91–98. (Cited on pages 87, 88 and 126.)

Wang, Z., Bhatia, K., Glocker, B., Marvao, A. D., Misawa, K., Mori, K., and Rueckert, D. (2014). Geodesic Patch-based Segmentation. In *MICCAI 2014*, pages 1–8. (Cited on page 66.)

Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (StaPle): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921. (Cited on page 56.)

Winter, M. M., Bernink, F. J., Groenink, M., Bouma, B. J., van Dijk, A. P., Helbing, W. a., Tijssen, J. G., and Mulder, B. J. (2008). Evaluating the systemic right ventricle by Cmr: the importance of consistent and reproducible delineation of the cavity. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, 10:40. (Cited on page 44.)

Wong, S. P., French, J. K., Lydon, A.-M., Manda, S. O. M., Gao, W., Ashton, N. G., and White, H. D. (2015). Relation of left ventricular sphericity to 10-year survival after acute myocardial infarction. *American Journal of Cardiology*, 94(10):1270–1275. (Cited on page 69.)

Wu, R., Yan, S., Shan, Y., Dang, Q., and Sun, G. (2015). Deep Image: Scaling up Image Recognition. *Arxiv*. (Cited on pages 28, 113 and 157.)

Xia, H., Asif, I., and Zhao, X. (2013). Cloud-Ecg for real time Ecg monitoring and analysis. *Computer Methods and Programs in Biomedicine*, 110(3):253–259. (Cited on pages 6 and 141.)

Xiong, C., Johnson, D., Corso, J. J., Xu, R., Corso, J. J., and Xu, R. (2012). Random Forests for Metric Learning with Implicit Pairwise Position Dependence. *ArXiv e-prints*, pages 1–13. (Cited on page 116.)

Ye, D. H., Desjardins, B., Ferrari, V., Metaxas, D., and Pohl, K. M. (2014). Auto-Encoding of Discriminating Morphometry from Cardiac Mri. In *IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 217–221, Beijing. IEEE. (Cited on page 87.)

Zaliaduonyte-Peksiene, D., Simonyte, S., Lesauskaite, V., Vaskelyte, J., Gustiene, O., Mizariene, V., Jurkevicius, R., Jariene, G., Tamosiunas, A., and Zaliunas, R. (2013). Left ventricular remodelling after acute myocardial infarction: Impact of clinical, echocardiographic parameters and polymorphism of angiotensinogen gene. *Journal of the renin-angiotensin-aldosterone system : JRAAS*. (Cited on page 91.)

Zhang, H., Wahle, A., Johnson, R. K., Scholz, T. D., and Sonka, M. (2010). 4-D cardiac Mr image analysis: Left and right ventricular morphology and function. *IEEE Transactions on Medical Imaging*, 29(2):350–364. (Cited on page 94.)

Zhang, N., Paluri, M., Ranzato, M. A., Darrell, T., and Bourdev, L. (2014a). PAndA: Pose Aligned Networks for Deep Attribute Modeling. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644. IEEE. (Cited on pages 110 and 154.)

Zhang, X., Cowan, B. R., Bluemke, D. a., Finn, J. P., Fonseca, C. G., Kadish, A. H., Lee, D. C., Lima, J. a. C., Suinesiaputra, A., Young, A. a., and Medrano-Gracia, P. (2014b). Atlas-based quantification of cardiac remodeling due to myocardial infarction. *PloS one*, 9(10):e110243. (Cited on pages 69, 70 and 87.)

Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D., Guerra, E., de Lara, J., Malizia, A., and Díaz, P. (2008). Four-chamber heart modeling and automatic segmentation for 3-D cardiac Ct volumes using marginal space learning and steerable features. *IEEE Transactions on Medical Imaging*, 27(11):1668–1681. (Cited on pages 45 and 46.)

Zhou, Y., Peng, Z., and Zhou, X. S. (2012). Automatic view classification for cardiac Mri. In *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1771–1774, Barcelona. IEEE. (Cited on pages 22, 23, 38, 40, 108, 126 and 152.)

Zikic, D., Glocker, B., and Criminisi, A. (2013). Atlas encoding by randomized forests for efficient label propagation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8151 LNCS:66–73. (Cited on page 48.)

Apprentissage Automatique pour Simplifier l'Utilisation de Banques d'Images Cardiaques

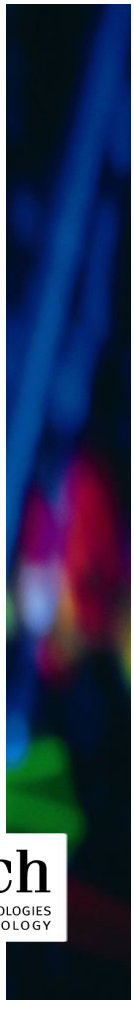
RÉSUMÉ : L'explosion récente de données d'imagerie cardiaque a été phénoménale. L'utilisation intelligente des grandes bases de données annotées pourrait constituer une aide précieuse au diagnostic et à la planification de thérapie. En plus des défis inhérents à la grande taille de ces banques de données, elles sont difficilement utilisables en l'état. Les données ne sont pas structurées, le contenu des images est variable et mal indexé, et les métadonnées ne sont pas standardisées. L'objectif de cette thèse est donc le traitement, l'analyse et l'interpretation automatique de ces bases de données afin de faciliter leur utilisation par les spécialistes de cardiologie. Dans ce but, la thèse explore les outils d'apprentissage automatique supervisé, ce qui aide à exploiter ces grandes quantités d'images cardiaques et trouver de meilleures représentations. Tout d'abord, la visualisation et l'interpretation d'images est améliorée en développant une méthode de reconnaissance automatique des plans d'acquisition couramment utilisés en imagerie cardiaque. La méthode se base sur l'apprentissage par forêts aléatoires et par réseaux de neurones à convolution, en utilisant des larges banques d'images, où des types de vues cardiaques sont préalablement établies. La thèse s'attache dans un deuxième temps au traitement automatique des images cardiaques, avec en perspective l'extraction d'indices cliniques pertinents. La segmentation des structures cardiaques est une étape clé de ce processus. A cet effet une méthode basée sur les forêts aléatoires qui exploite des attributs spatio-temporels originaux pour la segmentation automatique dans des images 3D et 3D+t est proposée. En troisième partie, l'apprentissage supervisé de sémantique cardiaque est enrichi grâce à une méthode de collecte en ligne d'annotations d'usagers. Enfin, la dernière partie utilise l'apprentissage automatique basée sur les forêts aléatoires pour cartographier des banques d'images cardiaques, tout en établissant les notions de distance et de voisinage d'images. Une application est proposée afin de retrouver dans une banque de données, les images les plus similaires à celle d'un nouveau patient.

Mots clés : Résonance magnétique cardiaque, apprentissage automatique, forêts de décision, réseaux de neurones à convolution, nettoyage automatique de données, segmentation d'image, recherche d'image par le contenu

Machine Learning for Simplifying the Use of Cardiac Image Databases

ABSTRACT : The recent growth of data in cardiac databases has been phenomenal. Clever use of these databases could help find supporting evidence for better diagnosis and treatment planning. In addition to the challenges inherent to the large quantity of data, the databases are difficult to use in their current state. Data coming from multiple sources are often unstructured, the image content is variable and the metadata are not standardised. The objective of this thesis is therefore to simplify the use of large databases for cardiology specialists with automated image processing, analysis and interpretation tools. The proposed tools are largely based on supervised machine learning techniques, i.e. algorithms which can learn from large quantities of cardiac images with groundtruth annotations and which automatically find the best representations. First, the inconsistent metadata are cleaned, interpretation and visualisation of images is improved by automatically recognising commonly used cardiac magnetic resonance imaging views from image content. The method is based on decision forests and convolutional neural networks trained on a large image dataset. Second, the thesis explores ways to use machine learning for extraction of relevant clinical measures (e.g. volumes and masses) from 3D and 3D+t cardiac images. New spatio-temporal image features are designed and classification forests are trained to learn how to automatically segment the main cardiac structures (left ventricle and left atrium) from voxel-wise label maps. Third, a web interface is designed to collect pairwise image comparisons and to learn how to describe the hearts with semantic attributes (e.g. dilation, kineticity). In the last part of the thesis, a forest-based machine learning technique is used to map cardiac images to establish distances and neighborhoods between images. One application is retrieval of the most similar images.

Keywords : Cardiac magnetic resonance, machine learning, decision forests, convolutional neural networks, data munging, segmentation, content-based image retrieval



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY