Universitat de Girona

DEFORMABLE OBJECT SEGMENTATION IN ULTRA-SOUND IMAGES

Joan MASSICH VALL

Dipòsit legal: Gi. 67-2014
http://hdl.handle.net/10803/128329

PhD Thesis

Deformable object segmentation in Ultra-Sound images

Joan Massich Vall

2013

PhD Thesis

Deformable object segmentation in Ultra-Sound images

Joan Massich Vall

2013

Doctoral Programme in Technology

Supervised by

Joan Martí Bonmatí

Fabrice Meriaudeau

Thesis submitted in partial fulfilment of the requirements for the Degree of
Doctor in Philosophy at the University of Girona and the
University of Burgundy

A la meva princesa maca,
i en memòria d'un dels meus herois.

Dr. Joan Martí, from Universitat de Girona, and
Dr. Fabrice Meriaudeau, from Université de Bourgogne,

DECLARE

That the work entitled *Deformable object segmentation in Ultra-Sound images*, presented by *Joan Massich* to obtain the degree in Doctor of Philosophy, has been developed under our supervision and complies with the requirements needed to obtain the International Mention.

Therefore, in order to certify the aforesaid statement, we sign this document.

Girona, October 2013.

Agraïments

Silent gratitude isn't much use to anyone.

G.B. Stern

Utilitzant la mateixa fórmula que utilitza algú que durant aquest període de doctorat ha passat a ser important per mi, m'agradaria agrair la feina que aquí presento a totes aquelles persones que fan possible que els meus somnis es facin realitat. Però en aquest cas no només voldria agrair-ho a qui m'ajuda, em guia i crea reforç positiu per tal d'aconseguir que surti el millor de mi, sinó que no voldria oblidar-me dels altres, de tots aquells que em fan serrar les dents i resar dos em cago en déu, perquè sense ells potser tampoc hi hagués arribat.

Un cop dit això tan genèric, m'agradaria agrair de forma més explícita a Joan Martí Bonmatí i Fabrice Meriaudeau, els meus dos directors de tesi, que hagin tingut la valentia (o falta de seny) de posar-se a dirigir una nau ingovernable, sempre a la deriva de l'última idea sense sentit, i que, a més a més, ho hagin fet sota la meva conducta incendiària, sota el crit de "foc a bordu" o "això és una merda" mentre m'he dedicat, dia per altre, i durant quatre anys, a cremar-li les veles.

En la mateixa línia també m'agradaria agrair al professor Hamed Sari-Sarraf la mateixa valentia o falta de seny per acollir-me dues vegades dins l'equip d'investigadors de l'Applied Vision Lab a la Universitat de Texas Tech (guns up, riders!).

Tampoc voldria oblidar-me d'agrair la paciència infinita de Sergi Ganau i Rosalia Aguilar, el personal d'UDIAT amb qui he treballat per entendre, recolectar i catalogar les imatges amb les quals realitzem les nostres tasques d'investigació. Estic d'acord que les colaboracions, en més o menys mesura, formen part de les obligacions de tots, però repeteixo que la paciència infinita que han tingut a UDIAT per formar, divendres rere divendres, a un analfabet mèdic com jo per tal que pogués llegir les imatges d'ultrasò adequadament i veure més enllà de simples taques, és un esforç que els agraeixo molt.

També m'agradaria agrair el suport econòmic obtingut de la Universitat de Girona a través de les beques BR-GR, al Ministerio a través dels projectes TIN2007-60553 i TIN2011-23704, i al Consell regional de la Bourgogne, ja que dels diners que han aportat entre els uns i els altres és d'on han anat sortint totes les misses.

Un cop dit tot això, podria anar agraint un per un a tothom fins arribar a agrair de forma personal al repartidor que porta les pizzes a la universitat si et fas passar per l'Andrés. Però si he de fer això fins arribar als repartidors de Girona, Lubbock i Le Creusot no acabaré mai, així que doneu-vos tots per agraïts.

Faré l'esforç, però, d'agrair el recolzament a tots els companys d'entrenament d'aquí i d'allà. Potser també a algú més perquè sinó m'ho sentiria a dir i en definitiva els agraiments és l'únic que llegireu.

A en Pueyo, per tot el que representa; al carcamal d'en Quintana, que ha passat de ser un vadell a haber de passar dues vegades per veure'l i, si la carretera fa pujada, haver-me'n d'oblidar de veure'l; a en Guilloume quan no està en forma; als nens com n'Enric, en Gubern o en Gamussilla per les sortides a peu, encara que facin el que els doni la gana i després es trenquin. A en Valverde, perquè em fa sentir menys tarat. A l'avi Cufí, pel ritme que tant m'agrada, o a en Robert, que encara és més còmode. Nois, seguiu pujant a Le Creusot perquè els d'allà estan taradíssims. A en Fabrice, en Micha, n'Olivier, n'Albhan, en Cedric i tots amb els que anem a córrer i que, un cop despenjat, passen a recollir-me per seguir torturant-me, els molt animals. A l'equip de triatló de Texas i al de ciclisme. Als dos equips amb els quals vam fer els ironmans. A la Sílvia, que em compensa les càrregues d'entrenament i feina amb dietes a base de plàtans, tot i saber que en sóc alèrgic.

Tampoc voldria oblidar-me d'en Ricard Prados, qui es pensa que l'he estat putejant amb estima durant quatre anys quan, en realitat, majoritariament tot han estat maniobres molt ben estructurades per n'Albert i executades per en Quintana. I sabeu que puc demostrar-ho perquè en tinc proves gràfiques.

M'agradaria agraïr també, però de forma seriosa encara que ells siguin uns catxondos, les converces i la feina feta amb l'Arnau, en Xavi, en Christian, en Gerard, en Desiré, n'Ian, n'Arunkumar o la poligonera de la Meri, que al final no mossega tant com ens vol fer creure perquè és un tros de pà beneit que quan sàpiga que us ho he explicat ja la sentiré.

I finalment en aquests agraiments no hi pot faltar, i ell ja sap perquè, en Miki que encara que no us ho creieu m'ha convencut més d'un cop en afluixar i passar per l'aro.

List of publications

Here it can be found a list of scientific contributions already published during the course of this phd along with a list of undergoing publications as outcome of this dissertation.

published contributions

- Massich, Joan, Fabrice Meriaudeau, Elsa Pérez, Robert Martí, Arnau Oliver, and Joan Martí. "Lesion segmentation in breast sonography." In Digital Mammography, pp. 39-45. Springer Berlin Heidelberg, 2010.

- Massich, Joan, Fabrice Meriaudeau, Elsa Pérez, Robert Martí, Arnau Oliver, and Joan Martí. "Seed selection criteria for breast lesion segmentation in Ultra-Sound images." In MICCAI Workshop on Breast Image Analysis, pp 55-64. Toronto, Canada. 2011

- J.Martí, A.Gubern-Mérida, J.Massich, A.Oliver, J.C.Vilanova, J.Comet, E.Pérez, M.Arzoz, and R.Martí. Ultrasound Image Analysis. Methods and Applications. Recent Advances in Biomedical Signal Processing, pp 216-230. Eds: J.M.Górriz, E.W.Lang, and J.Ramírez, Bentham Science Publishers. 2011.

- Massich, Joan, Fabrice Meriaudeau, Melcior Sentís, Sergi Ganau, Elsa Pérez, Robert Martí, Arnau Oliver, and Joan Martí. "Automatic seed placement for breast lesion segmentation on US images." In Breast Imaging, pp. 308-315. Springer Berlin Heidelberg, 2012.

- Massich, Joan, Fabrice Meriaudeau and Joan Martí. "Segmentation techniques applied to breast ultrasound imaging: A review." to be published. *Submitted to Medical Image Analysis, Elsevier.*

- Massich, Joan, Fabrice Meriaudeau and Joan Martí. "A superpixel based technique for breast lesion delineation in sonographic data." *Submitted to Computers in Biology and Medicine, Elsevier*

List of Tables

List of Figures

List of Acronyms

ABUS Automated whole Breast Ultra-Sound 6
ACM Active Contour Model 37
ACR American College of Radiology 14
ACWE Active Contour Without Edges 66
ADF Anisotropic Diffusion Filter 37
AMED Average Minimum Euclidian Distance 57
AOV Area Overlap 36
ARD Average Radial Derivative 36
ARE Average Radial Error 55
BI-RADS Breast Imaging-Reporting and Data System 20
BoF Bag-of-Features 91
BoW Bag-of-Words 91
BUS Breast Ultra-Sound xxii
CAD Computer Aided Diagnosis xxii
CADe Computer Aided Detection 27
CADx Computer Aided Diagnosis 27
CC Cranio-Caudal 3
CRF Conditional Random Field 42
CV Computer Vision 77
DIC Ductal Infiltrating Carcinoma 113
DICOM Digital Imaging and Communications in Medicine 112
DM Digital Mammography xxi
DPM Deformable Part Model 42

MRF Markov Random Field 37
MRI Magnetic Resonance Image 2
NC Normalized Cuts 48
NPV Negative Predictive Value xxii
NRV Normalized Residual Value 54
OF Overlap Fraction 54
PCA Principal Component Analysis 93
PDE Partial Differential Equation 44
PDF Probability Density Function 45
PD Proportional Distance 57
PET Position Emission Tomography 10
PPV Positive Predictive Value xxii
PR Pattern Recognition 99
QC Quadratic-Chi 85
QS Quick-Shift 115
RBF Radial Basis Function 104
RGI Radial Gradien Index 38
RGI Radial Gradient Index 38
RG Region Growing 45
ROI Region Of Interest 37
SA Simulated Annealing 106
SIFT Self-Invariant Feature Transform 91
SI Similarity Index 53
SLIC Simple Linear Iterative Clustering
STAPLE Simultaneous Truth and Performance Level Estimation 39
SVM Support Vector Machine 104
TN True Negative 51
TPR True-Positive Ratio 54
TP True Positive 40
US Ultra-Sound xxi

Contents

Abstract

Breast cancer is the second most common cancer (1.4 million cases per year, 10.9% of diagnosed cancers) after lung cancer, followed by colorectal, stomach, prostate and liver cancers [1]. In terms of mortality, breast cancer is the fifth most common cause of cancer death. However, it place as the leading cause of cancer death among females both in western countries and in economically developing countries [2].

Medical imaging plays an important role in breast cancer mortality reduction, contributing to its early detection through screening, diagnosis, image-guided biopsy, treatment follow-up and suchlike procedures [3]. Although Digital Mammography (DM) remains the reference imaging modality, Ultra-Sound (US) imaging has proven to be a successful adjunct image modality for breast cancer screening [3], [4], specially as a consequence of the discriminative capabilities that US offers for differentiating between solid lesions that are benign or malignant [5] so that the amount of unnecessary biopsies, which is estimated to be between $65 \sim 85\%$ of the prescribed biopsies [6], can be reduced [7] in replacing them by short-term US screening follow-up [8].

Regardless of the clinical utility of the US images, such image modality suffers from different inconveniences due to strong noise natural of US imaging and the presence of strong US artifacts, both degrading the overall image quality [9] which compromise the performance of the radiologists. Radiologists infer health state of the patients based on visual inspection of images which by means of some screening technique (e.g. US) depict physical properties of the screened body. The radiologic diagnosis error rates are similar to those found in any other tasks requiring human visual inspection, and such errors, are subject to the quality of the images and the ability of the reader to interpret the physical properties depicted on them[10].

Therefore the major goals of medical imaging researchers in general, and also in particular for breast lesion assessment using US data, has been to provide better instrumentation for improving the image quality, as well as,

methodologies and procedures in order to improve the interpretation of the image readings. In image interpretation unified terms for characterizing, describing and reporting the lesions have been developed [5], [11]–[13] in order to reduce diagnosis inconsistencies among readers [14]. Such unifying terms so called lexicons are proven to be a useful framework for the radiologists when analyzing Breast Ultra-Sound (BUS) images. The Positive Predictive Value (PPV) and Negative Predictive Value (NPV) which represent the percentage of properly diagnosed cases [15] achieved when describing lesions with these lexicon tools turned them into the standard for human reading and diagnosis based on BUS images.

A common framework allows managing the US imaging inconveniences such as strong noise or artifacts by allowing the comparison of double readings done by several specialized observers. The major inconvenience for double reading is the elevated time required from the radiologists. Thus, since a single observer using Computer Aided Diagnosis (CAD) as a second opinion has been proven to achieve comparable results [16], CAD systems are used to aliviate the time demand from the radiologists. However these descriptors are subject to an accurate delineation of the lesion which when read by an expert radiologist is instantly understood but in a CAD system a computerized system is required.

This thesis analyzes the current strategies to segment breast lesions in US data and proposes a fully automatic methodology for generating accurate segmentations of breast lesions in US data with low false positive rates. The proposed approach targets the segmentation as a minimization procedure for a multi-label probabilistic framework that takes advantage of min-cut/max-flow Graph-Cut (GC) minimization for inferring the appropriate label from a set of tissue labels for all the pixels within the target image. The image is divided into contiguous regions so that all the pixels belonging to a particular region would share the same label by the end of the process. From a training image dataset stochastic models are build in order to infer a label for each region of the image. The main advantage of the proposed framework is that it splits the problem of segmenting the tissues present in US the images into subtasks that can be taken care of individually.

Resum

Amb 1,4 milions de casos anuals i comptabilitzant el 10,9% del total de diagnòstics, el càncer de mama és el segon càncer més comú darrere del càncer de pulmó, seguit del càncer de colon, d'estómac, de pròstata i de fetge. En termes de mortalitat en tota la població, el càncer de pit és la cinquena causa de mortalitat. Si només es té en compte la població femenina, el càncer de mama lidera la mortalitat per càncer tant en països desenvolupats com en països en desenvolupament.

La imatge mèdica juga un paper crucial a l'hora de combatre la mortalitat per càncer de mama, i en facilita, entre d'altres, les tasques de detecció precoç, diagnosis, biòpsies guiades per imatge o seguiment de l'evolució de les lesions. Tot i que la Mamografia Digital (MD) segueix essent la principal modalitat d'imatge, les imatges d'ultrasò s'han convertit en una valuosa modalitat d'imatge per complementar les exploracions mèdiques. La seva principal vàlua és que aquestes imatges aporten informació que permet determinar la benignitat o malignitat de les lesions sòlides, que no es poden determinar només amb MD. Com a conseqüència de complementar MD amb imatges d'ultrasó, s'estima que entre un 65% i un 85% de les biòpsies prescrites es podrien evitar, tot canviant-les per un seguiment periòdic basat en imatges d'ultrasò.

Malgrat la utilitat mèdica de les imatges d'ultrasó, aquest tipus d'imatges són molt sorolloses i pateixen artefactes que comprometen les capacitats de diagnosis per part dels radiòlegs que han d'interpretar l'estat de salut del pacient a partir d'aquestes imatges. Els errors de diagnosi basats en la lectura d'imatges mèdiques són similars als de qualsevol altra tasca que requereixi inspecció visual i es troben subjectes a la qualitat de les imatges, així com a les habilitats dels radiòlegs per interpretar-les correctament.

Per aquestes raons, dins la comunitat que investiga imatge mèdica de forma general, així com en el cas particular del càncer de mama, s'intenta desenvolupar tant maquinària i/o processos que millorin la qualitat de les imatges, com metodologies per millorar-ne i sistematitzar-ne la lectura i

interpretació. A fi de millorar la interpretació de les imatges, la comunitat mèdica ha desenvolupat un lèxic comú per reduir inconsistències entre les lectures dels radiòlegs. S'ha demostrat que la utilització d'aquest tipus d'eines, consistents en un conjunt d'atributs concrets (lèxic) que són assignats a les imatges per tal de descriure-les, millora el percentatge de lesions correctament diagnosticades, fet que les ha convertit en l'estàndard a l'hora de llegir imatges per part dels radiòlegs.

El fet d'utilitzar un lèxic comú permet comparar múltiples lectures de diversos radiòlegs per millorar, així, la diagnosi final. Tot i que dur a terme aquest tipus de lectures múltiples és d'una pràctica habitual, no deixa de ser molt costosa, ja que diversos especialistes han d'analitzar les imatges. Per aquesta raó, dins el camp mèdic s'han introduït els sistemes CAD d'assistència computaritzada per la diagnosi per obtenir una segona opinió. S'ha demostrat que la diagnosi final produïda per un radiòleg utilitzant un sistema CAD és equiparable a la decisió consensuada per múltiples radiòlegs, fet que permet elleugerir el volum de tasques dels radiolegs. El principal problema en el desenvolupament de sistemes CAD acurats rau en què aquest lèxic depén d'una delineació fidel de les lesions, que un lector expert pot dur a terme de forma intuïtiva i natural però que un sistema CAD necessita d'un proces que realitzi aquesta tasca. D'aquí la importància de desenvolupar sistemes acurats de delineació de lesions en imatges de mama en ultrasò.

En aquest treball, es proposa un sistema automàtic per generar delineacions acurades de les lesions de mama en imatges d'ultrasò. El sistema proposat planteja el problema de trobar la delineació corresponent a la minimització d'un sistema probabilístic multiclasse mitjançant el tall de mínim cost del graf que representa la imatge. El sistema representa la imatge com un conjunt de regions i infereix una classe per cada una d'aquestes regions a partir d'uns models estadístics obtinguts d'unes imatges d'entrenament. El principal avantatge del sistema és que divideix la tasca en subtasques més fàcils d'adreçar i després soluciona el problema de forma global.

Resumen

Con 1,4 millones de casos anuales que contabilizan el 10,9% del total de diagnósticos, el cáncer de mama es el segundo cáncer más común detrás del cáncer de pulmón, seguido por el cáncer de colon, de estómago, de próstata y del cáncer de hígado. En términos de mortalidad respecto toda la población, el cáncer de mama es la quinta causa de mortalidad. Considerando solamente la población femenina, el cáncer de mama lidera la mortalidad por cáncer en países desarrollados y también en países en vías de desarrollo.

La imagen médica es crucial para combatir la mortalidad por cáncer de mama ya que facilita su detección precoz, diagnosis, biopsias guiadas o seguimiento de la evolución de las lesiones. Aunque la Mamografía Digital (MD) sigue siendo la principal modalidad de imagen médica para la visualización de la mama, las imagenes de ultrasonido se han convertido en una valiosa modalidad de imagen para complementar dichas exploraciones médicas. Su principal valua es que las imágenes de ultrasonido aportan información que permite determinar la benignidad o malignidad de las lesiones sólidas, que no se puede determinar usando únicamente MD. Como consecuencia de complementar MD con imágenes de ultrasonida, se estima que entre un 65% y un 85% de las biopsias prescritas se podrían evitar, cambiandolas por un seguimiento periódico basado en imágenes de ultrasonido.

A pesar de la valua médica de las imágenes de ultrasonido, este tipo de imagenes padecen de mucho ruido y artefactos que comprometen las capacidades de diagnóstico por parte de leso radiologos. Los errores de diagnosis debidos a una mala lectura de las imágenes médicas son similares a los errores producidos en cualquier otra tarea que requiera inspección visual. Dichos errores están sujetos a la calidad de las imagenes y a las habilidades de los radiólogos en interpretarlas.

Por las razones mencionadas, en la comunidad que investiga la imagen médica de forma general, así como para el caso particular del cáncer de mama, intenta desarrollar maquinaria y/o procesos que mejoren la calidad de las imagenes, como metodologías para mejorar y sistematizar la lectura

e interpretación de las imagenes. Con el fin de mejorar la interpretación de las imagenes, la comunidad médica ha desarrollado un léxico común para reducir inconsistencias entre lecturas de radiólogos. Está demostrado que la utilización de este tipo de herramientas, que consisten en un conjunto de atributos concretos (léxico) que debe ser asignado a las imagenes a modo de descripción, mejora el porcentaje de lesiones correctamente diagnosticadas. Hecho que ha convertido estas herramientas en el procedimiento estándar de lectura de las imagenes por parte de los radiólogos.

La utilización de un léxico común permite comparar las lecturas de varios radiólogos permitiendo mejorar el diagnosis final. Aunque la práctica de lecturas múltiples es una práctica habitual, no deja de ser muy costosa, ya que varios especialistas deben analizar las imagenes. Por esta razón, se han introducido los sistemas de asistencia computarizada a la diagnosis (CAD) que facilitan una segunda opinión al radiólogo. Esta demostrado que el diagnosis final producido por un radiólogo utilizando un sistema CAD es equiparable al diagnosis consensuando lecturas de múltiples radiólogos, hecho que permite reducir la carga de trabajo de los radiólogos. El principal problema al desarrollar sistemas CAD fiables radica en que dichos léxicos dependen de una correcta delineación de las lesiones. Un lector experto es capaz de visualizar dichas delineaciones de una forma natural e intuitiva, pero un sistema CAD necessita de procesos computarizados para realizar una delineación acurada. De ahí la importancia de desarrollar sistemas fiables para la delineación acurada de lesiones en imagenes ultrasonicas de mama.

En el trabajo aquí presentado, se propone un sistema automático para generar delineaciones acuradas de las lesiones de mama en imágenes de ultrasonido. El sistema propuesto plantea el problema de la delineación como la minimización de un sistema probabilístico multiclase mediante el corte de coste mínimo del graf representando la imagen. El sistema representa la imagen como un conjunto de regiones y infiere una clase para cada una de las regiones presentes en base a unos modelos estadísticos obtenidos durante un proceso de entrenamiento. La principal ventaja del sistema propuesto es que divide el problema en subtareas más fáciles de solventar y finalmente soluciona la segmentación de forma global.

Résumé

Le cancer du sein est le type de cancer le plus répandu (1,4 millions de cas par an, 10,9% des cancers diagnostiqués) après le cancer du poumon. Il est suivi par le cancer du colon, le cancer de l'estomac, celui de la prostate et le cancer du foie . Bien que parmi les cas mortels, le cancer du sein soi classé cinquième type de cancer le plus meurtrier, il reste néanmoins la cause principale de mortalité chez les femmes aussi bien dans les pays occidentaux que dans les pays en voie de développement .

L'imagerie médicale joue un rôle clef dans la réduction de la mortalité du cancer du sein, en facilitant sa première détection par le dépistage, le diagnostic, la biopsie guidée par l'image et le suivi de traitement et des procédures de ce genre.

Bien que la Mammographie Numérique (DM) reste la référence pour les méthodes d'examen existantes, les échographies ont prouvé leur place en tant que modalité complementaire. Les images de cette dernière fournissent des informations permettant de différencier le caratère bénin ou malin des lésions solides, ce qui ne peut être détecté par MD. On estime que 65 à 85% des biopsies prescrites pourraient être évitées par la mise en place d'un suivi régulier basé sur des images échographiques. Malgré leur utilité clinique, ces images sont bruitées et la présence d'artefacts compromet les diagnostiques des radiologues interprètant l'état de santé du patient à partir de celles ci. Les erreurs de diagnostic basées sur la lecture des images médicales sont similaires à toute autre tâche qui exige une inspection visuelle et sont soumises à la qualité des images ainsi qu'aux compétences des radiologistes. C'est pourquoi un des objectifs premiers des chercheurs d'imagerie médicale a été de fournir une meilleure instrumentation dans le but d'améliorer la qualité d'image et des méthodologies permettant d'améliorer et de systématiser la lecture et l'interprétation de ces images. Pour améliorer l'interprétation des images, la communauté médicale a mis au point un lexique commun réduisant les incohérences entre radiologues.

Il a été démontré que l'utilisation de ces outils, composé d'un ensem-

ble spécifique de caractéristiques (lexique) qui sont affectés à des images pour les décrire, en améliorant le pourcentage de lésions correctement diagnostiquées [15], est devenu la norme lors de la lecture des images par les radiologues.

L'utilisation d'un lexique commun permet de comparer plusieurs lectures de différents radiologues afin d'améliorer le diagnostic. Un telle pratique est énormément couteuse en temps. Etant donné qu'il a été prouvé que l'utilisation de Computer Aided Diagnosis CAD en tant que deuxième observateur permet l'obtention de résultats comparables, ces système sont donc utilisés pour améliorer l'exactitude des diagnostics.

Si pour un lecteur qualifié, la délineation fidèle des lésions peut être éffectuée de manière intuitive et naturelle, le CAD nécessite le développement d'un système de délimitation précis pour l'utilisation du lexique.

Le problème principal dans le développement d'un CAD précis vient du fait que ce lexique dépend d'une délineation fidèle des lésions qui, même si pour un lecteur qualifié peut être effectuée de manière intuitive et naturelle. D'où l'importance du développement de systèmes de délimitation précise des lésions dans les images de l'échographie du sein.

La méthode proposée considere le processus de segmentation comme la minimisation d'une structure probabilistique multi-label utilisant un algorithme de minimisation du Max-Flow/Min-Cut pour associer le label adéquat parmi un ensemble de labels figurant des types de tissus, et ce, pour tout les pixels de l'image. Cette dernière est divisée en régions adjacentes afin que tous les pixels d'une même régions soient labelisés de la même manière en fin du processus. Des modèles stochastiques pour la labellisation sont crées à partir d'une base d'apprentissage de données. L'avantage principal de la méthodologie proposée est le découpage de l'opération de segmentation de tissu en sous tâches indépendentes les unes des autres.

Chapter 1

Introduction

> The soul cannot think without a picture
>
> Aristotle

1.1 Breast cancer

Breast cancer is the second most common cancer (1.4 million cases per year, 10.9% of diagnosed cancers), after lung cancer and followed by colorectal, stomach, prostate and liver cancers [1]. In terms of mortality, breast cancer is the fifth most common cause of cancer death. However, it is the leading cause of cancer death among females both in western countries and in economically developing countries [2].

In general, breast cancer incidence rates are higher in western countries not only because of incidence factors like reproductive patterns, such as late age at first birth and hormone therapies, either contraceptives or prolonged, but also, due to the aging of the population, which raises the overall incidence rates even if the age-specific rates remain constant [17], [18].

In contrast to the rising incidence rate of breast cancer over the last two decades in western countries, studies such as Autier et al [19] report that breast cancer mortality has been declining in many countries. This decrease is attributed to the combined effects of breast screening, which allows the detection of the cancer at its early stages, and to the improvements made in breast cancer treatment.

1.2 Image diagnostic techniques applied to breast cancer

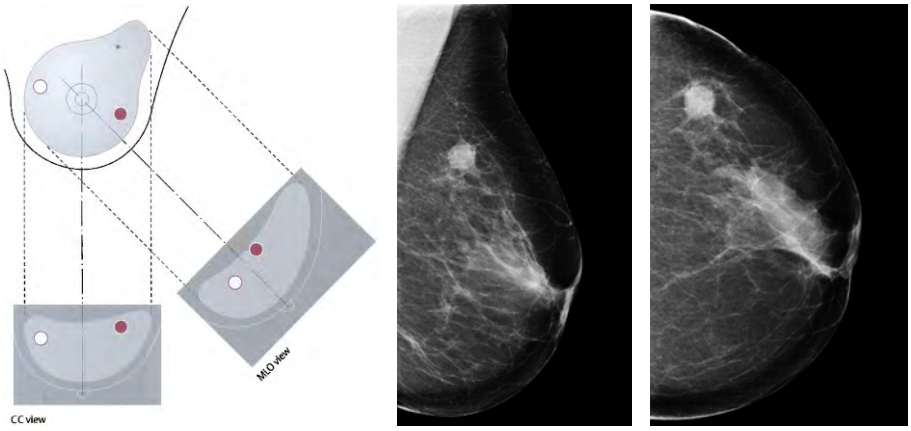
Medical imaging refers to the techniques and processes used to create images depicting physical properties of the human body or animals (or parts thereof) in order to infer health state for clinical purposes or medical therapy. In an editorial by Angell et al. published in the *New England Journal of Medicine* [20], the medical imagine discipline is qualified as one of the most important medical developments of the past thousand years since medical imaging provides physicians with in vivo images describing physiology and functionality of organs.

Without exception, medical imaging plays the most important role in breast cancer mortality reduction, contributing to its early detection through screening, diagnosis, image-guided biopsy, treatment follow-up and suchlike procedures [3].

Digital Mammography (DM) is, and remains, the preferred screening technique for early detection and diagnosis of breast cancer [21]. It is estimated that a 15 to 35% reduction in mortality in breast cancer deaths is due to the wide implementation of screening mammography. However, almost 25% of cancers still go undetected under mammography screening [22], typically in nonfatty breasts where the dense tissue shields the lesions. This is an important limitation in mammography screening, since about 40% of the female population have some dense breast tissue, and dense tissue is a risk factor for developing breast cancer. Patients with dense tissue in 75% or more of the breast have a four to six times higher probability of developing breast cancer compared to patients with dense tissue in 10% or less of the breast [23]. In addition, a large number of mammographic abnormalities (between $65 \sim 85\%$) turn out to be benign after biopsy [6].

Therefore, it is recommended to use other image modalities like US and Magnetic Resonance Image (MRI) screening as complementary images since they are more sensitive than mammography in a dense breast scenario [4]. In some cases these techniques also offer higher specificity than mammography allowing doctors to distinguish benignant and malignant signs which can then be used to reduce the amount of unnecessary biopsies [3], [5], [24].

In spite of these mammography screening drawbacks, mammography remains the gold standard screening technique due to the greater ability mammography has over US or MRI imagery in depicting small non-palpable lesions (always in a non-dense breast scenario) [25]. Also, the fact that microcalsifications, which are a clear sign of malignancy, are usually mistaken



(a) (b) (c)

Figure 1.1: Mammography Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC) view points: (a) illustrates the projection of the two most used view points (image from [27]), which produces images like the Medio-Lateral Oblique (MLO) in (b) and the Cranio-Caudal (CC) in (c). Notice the presence of the pectoral muscle in the upper-left corner of the MLO example (b).

as artifacts in US or MRI imaginary [26]; or the fact that most ductal carcinoma in situ are missed under sonography [11] plays in favor of mammography screening.

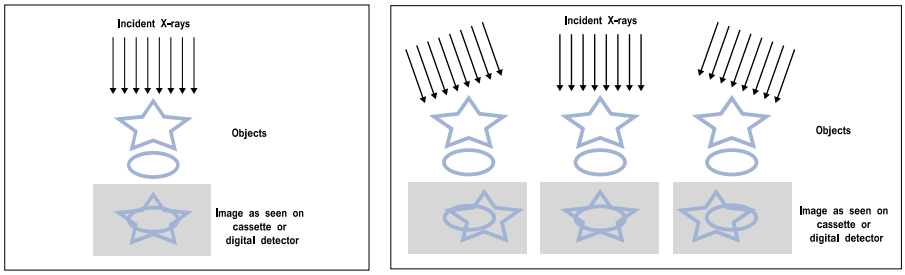
However, combining clinical examination with multiple modality imaging is more sensitive than any individual image modality [4].

1.2.1 X-ray screening, Mammography and Tomosynthesis

Full-Field Digital Mammography and Screen-Film Mammography

Mammography is a two-dimensional image modality that captures electromagnetic waves of an X-ray band passing through a compressed breast. Depending on the compression deformation of the breast, the images are classified into different categories. Figure 1.1 shows the two most used viewpoints for extracting mammograms: the Medio-Lateral Oblique (MLO) view and Cranio-Caudal (CC) view. Figure 1.1(a) illustrates the projection of the breast into the views and fig. 1.1(b,c) show an example of each mammography view of the same breast with a visible mass.

DM is the natural evaluation of screening the breast using X-rays and has



(a) (b)

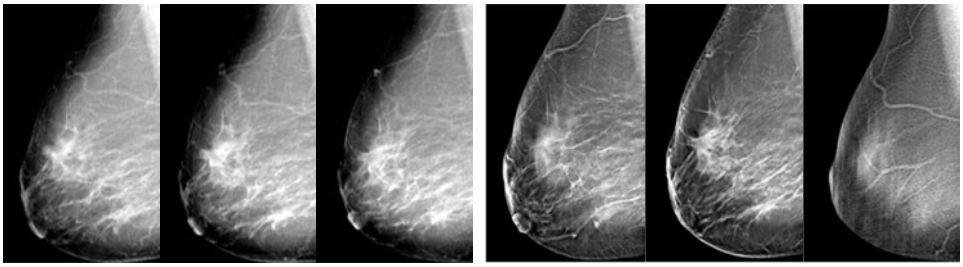
Figure 1.2: Mammography and Tomosynthesis image takes. (a) A mammography single image take illustrating the tissue overlap problem that shows that breast cancer can be shielded by dense normal breast tissue. (b) A multiple image take for tomosynthesis showing how the relative position between two targets vary depending on the X-ray's illumination angle. The views in (b) can be used to unfold the tissue overlap by composing a 3D-volume from the multiple views. The images illustrating this figure are taken from Smith et al. [26].

become the image screening of reference when diagnosing breast cancer [21], [28]. DM can either be digitized Screen-Film Mammography (SFM) when the image is obtained as the digitization of an analogical film or Full-Field Digital Mammography (FFDM) when the image is directly generated in a digital sensor instead of a sensible film.

Although no difference in cancer detection rates between FFDM and SFM [29] have been yet observed, FFDM has become the standard mammography screening due to its obvious advantages in a digitized environment.

Advances in X-ray screening of the breast, Breast Tomosynthesis

This technique tries to overcome the effect of tissue overlap present in regular mammograms. The screening technique is similar to mammography, the breast is compressed between two plates and X-ray attenuation is measured. The difference is that instead of using a single viewpoint, multiple images of the breast are taken at different angles and further combined to reconstruct them into cross-sectional slices. Figure 1.2 illustrates the effect of taking images at different angles, and figure 1.3 shows an example of taking different images of the same breast (fig. 1.3(a-c)) and the resultant cross-sectional slices from synthesizing the 3D-volume (fig. 1.3(d-f)).



(a) (b) (c) (d) (e) (f)

Figure 1.3: Tomosynthesis image acquisition and reconstruction example. Images (a-c) correspond to the X-ray images at different angles of the same take, and images (d-f) correspond to different cross-sectional slices of the reconstructed 3D-volume of the same breast. The images illustrating this figure are taken from A. Smith [30].

1.2.2 Sonography

Ultra-Sound (US) imaging uses high-frequency mechanical waves (sound waves typically within the $1 \sim 20Mhz$ range) in order to insonify the area to inspect and capture the waves reflected at boundaries between tissues with different acoustic properties [9][1]. The most common sonography screening technique applied to breast cancer screening is the hand-held realtime B-mode US imaging system.

B-mode imaging equipment generates two-dimensional images by means of a beam that travels through the tissue. The amplitude of the reflection caused by tissue interfaces is represented as brightness. The depth of the depicted boundaries is proportional to the interval of the reflection arrivals.

Despite the advantages that US screening offers, images lack in quality and suffer from severe artifacts. Another inconvenience of US screening is that regular equipment uses a hand-held probe run over the breast surface by the physician in order to take an arbitrary slice of the breast. This approach strongly relates the acquisition to the ability of the user. Further discussion of these topics can be found throughout Section 1.3 of this document.

[1] We refer the reader to Ensminger and Stulen [9] for a deeper understanding of US physics and image formation.

Real time spatial compound imaging (or sonoCT)

In order to improve the image quality, real time spatial compound imaging, or sonoCT, at every acquisition deflects the US beam and takes three to nine samples at different angles instead of a single take (see fig.1.4a,b) [31]. The sonoCT acquisition procedure of taking multiple views somehow recalls the acquisition process carried out in tomosynthesis. The difference is that sonoCT does not use the extra information to synthetise a 3D-volume, but uses the data redundancy for reducing the artifacts and noise, and to obtain an improved overall image, providing better tissue differentiation [32]. Its main drawback is the bluring effect caused by scene changes between takes. These scene changes can be caused by unintentional movements of the acquisition probe in a hand-held US device or due to movement by the patient. Figure 1.4 intuitively compares the sonoCT acquisition process with regular US imaging and also shows the outcome difference. For further details on this technology, the reader is referred to the works of Entrekin et al. [31], [33].

Automated whole Breast Ultra-Sound (ABUS)

Other advances in US acquisition address the dependency of the physician's skills for taking proper images. In Automated whole Breast Ultra-Sound (ABUS) a much larger transducer is used for exhaustive-scanning of the breast in an automatic manner with no dependency on the user. Then all the acquired slices are combined to generate a three-dimensional volume of the breast, overcoming the limitation of scanning only the focal area of concern as happens in hand-held US screening [34]. Figure 1.5 illustrates both hand-held US and ABUS acquisition systems to intuitively understand the differences between both systems.

Doppler Imaging

Sonographic Doppler imaging or the M-mode sonogram uses the well known Doppler shift effect. When the radiating energy cuts through a moving object, the received signal shifts its frequency depending on the relative velocity between the moving object and the moving observer.

The frequency shift captured by the Doppler effect is displayed as a color overlay in a B-mode image. Doppler imaging supposes a functional image used to visualize the blood flow which is representative of the lesion's metabolism.

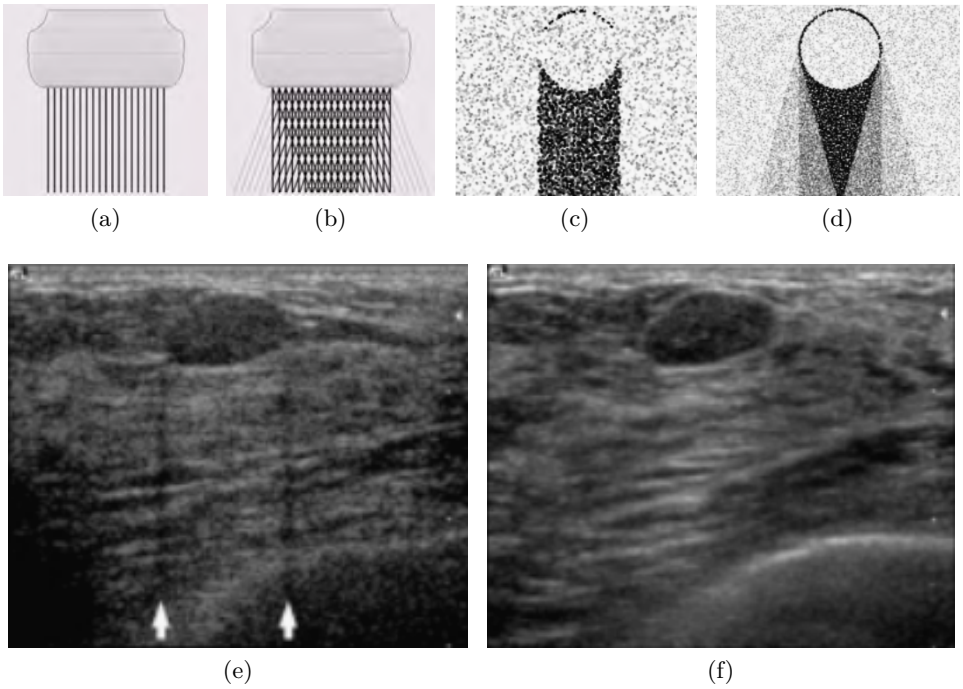


Figure 1.4: Comparison between conventional B-mode US imaging and real time spatial compound US imaging (sonoCT). (a,b) linear transducer comparison: in the conventional acquisition (illustrated in a) a single beam is used, whereas for compound imaging (b) several beams, at different angles, are used. (c,d) illustrates the insonifying advantages of conventional US (c) and sonoCT (d). Finally, (e) and (f) are examples of the same fibroadenoma using conventional screening and sonoCT. Notice that the lateral shadows caused by the fibroadenoma in (e) disappear in (f). Also, a proper hyperechoic boundary in the fibroadenoma's upper left hand corner appears in (f), depicting high reflection at the interface between the regular adipose tissue and the lesion which can not be appreciated in (e). The overall image quality of (f) is far superior to (e), supporting the foundings in [32]. All the images used in this figure are taken from Entrekin et al. [33]

(a) (b)

Figure 1.5: Conventional hand-held US and ABUS acquisition devices comparison. (a) Conventional hand-held US imaging acquisition device. (b) ABUS acquisition device.

Sonoelastography

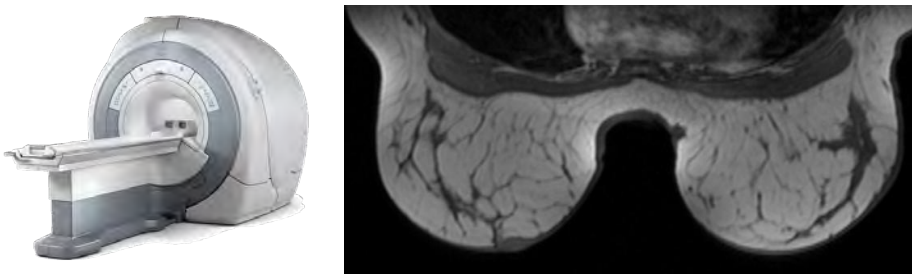
Sonoelastography can be seen as a highly sensitive ultrasonic palpation coloring the stiffness of the tissues over B-Mode sonogram [9]. In order to generate the data, pressure is applied over the tissue through mechanical vibrations (sound wave $< 10Hz$). Then the Doppler effect is used to measure the movement of the tissues. The stiffer the tissue, the lesser the vibration present compared to softer tissues.

1.2.3 Magnetic Resonance Image (MRI)

Although early efforts of using Magnetic Resonance Image (MRI) imaging to screen breasts were discouraging due to low spatial resolution [35], further studies combined with the use of contrast agents proved MRI to be an effective screening technique to assess breast lesions [4].

MRI screening technologies expose the tissue to a strong magnetic field to excite and align the nuclear particles within the tissue. Then the decay signal of the polarization state of each particle is recorded to generate a three-dimensional image. According to the tissue type, the decay signal shows different characteristics allowing technicians to distinguish the tissue type. Figure 1.6 exemplifies an MRI take of a patient.

The main advantage of using MRI is its capability of capturing functional behavior of the breast using a contrast agent to highlight areas containing a dense blood vessel network (known as angiogenesis areas), a typical characteristic of tumor structures.



(a) (b)

Figure 1.6: Magnetic Resonance Image (MRI) example. (a) generic General Electric healthcare resonance unit (image taken from their catalog) (b) transverse MRI image slice from a patient's chest, in which the breast and its structures can be clearly identified.

1.2.4 Other breast imaging techniques

In spite of DM being the principal screening technique for breast cancer and both B-mode US imaging and MRI are considered a beneficial and complementary adjunct to mammography, these modalities are far from perfect. Although the use of Full-Field DM has many advantages and commodities [29], its functioning principles are the same as the first proposals of Screen-Film Mammography in the 1960s [36]. In addition, US and MRI have their own limitations, otherwise mammography wouldn't remain the preferred breast screening modality.

Therefore, improving the current imaging technologies and exploring new imaging modalities is being investigated [21], [26]. Here, some of these modalities are named.

Bioelectric Imaging

This is based on the different electrical properties between normal and malignant breast tissue. These differences are measured with a probe capturing the low level electricity patterns applied to the breast's surface.

Breast Thermography

An infrared camera is used to identify areas of angiogenesis by tracking the temperature of the blood as it flows into the breast.

Near Infrared Optical Imaging

This technique measures the transmission of near infrared light through the breast so that areas of vascular development (angiogenesis) and/or areas saturated with hemoglobin and oxygen (hyper-metabolism) are highlighted.

Contrasts Developing

Contrast agents are being developed to produce contrast-enhanced mammographies and functional MRI, where areas with a particular behavior are highlighted during the screening.

Positron Emission Tomography (PET)

This technique is a nuclear imaging technique in the same category as scintimammography used to restating and evaluating recurrent breast cancer. In Position Emission Tomography (PET), a radioactive glucose, usually 18-fluoro-2deoxyglucose (FDG), is injected into the patient and areas of high tracer uptake are visualized with a gamma camera. A number of breast specific PET scanners are currently in development and being tested in clinical trails to demonstrate their efficiency. However, PET examinations are extremely expensive and are not widely available [26].

Scintimammoraphy

This technique is also a nuclear imaging technique which uses a gamma camera to visualize a radioactive tracer. Although recent advances have been made in high-resolution cameras designed specially for breast imaging, the resolution of scintimammography is still low compared to PET [26].

1.3 Ultra-Sound imaging and its role in Breast Cancer

Although US applied to breast cancer screening was expected to surpass mammography since its initial studies in the early 50s carried out by Wild and Reid [37], and the variety of advances that sonography has undergone [38], Digital Mammography (DM) is, and remains, the preferred screening technique when diagnosing breast lesions [21], [39]–[41]. However, it is widely accepted that extensive mammography density is strongly associated with the risk of breast cancer and that mammographic scanning has

a low specificity in such a scenario [22], [23]. Therefore, the convenience of using alternative screening techniques (US, MRI, PET, and suchlike) is obvious, since there is the urgent need to increase the detection of unnoticeable cancers during physical examination [3], [4], [6], [42]–[45]. Although there is great controversy in using alternatives to mammography as a primary screening tool since in retrospective review lesion signs can be found in the mammography screenings [39]; it is easily understood that multi-modality readings are more sensitive than any individual test alone [3], [4], [6].

Despite the fact that some studies report that other modalities, such as MRI, have higher sensitivity compared to US [4], sonography is the most common image modality adjunct to mammography because it is widely available and inexpensive to perform [3], [34], [46]. Moreover, US, apart from its detection capabilities, has the ability to discern the typology of solid lesions [5], [11]–[13], [40], which can be used to reduce the number of unnecessary biopsies prescribed by DM [7], [8], which are estimated to be between $65 \sim 85\%$ [6]. Eventhough data suggest unnecessary biopsies can be replaced by short-term US screening follow-up, further studies are needed in order to determine whether this conclusion holds [47].

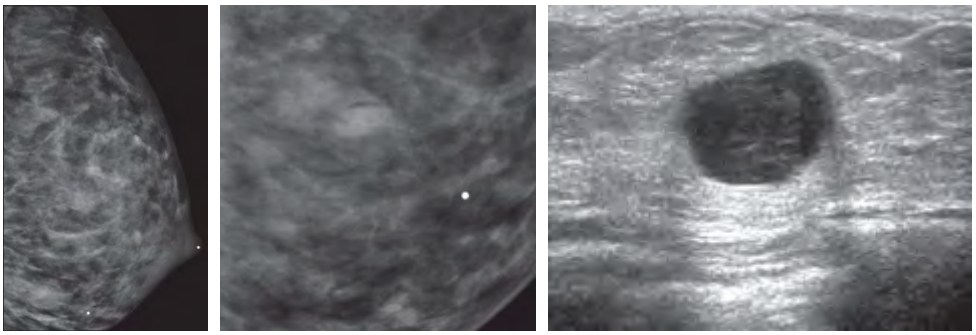
Figure 1.7 illustrates a case taken from Hines et al. [48] where DM and US images of a lactating patient who presented a palpable lesion were taken. In the MLO DM image (fig. 1.7a) and its magnified lateral view (fig. 1.7b), it is hard to spot the lesion, while the lesion is clearly visible when using US screening (fig. 1.7c). The findings in the US data reveal a complicated cyst, which is nothing more than a benign lesion. The patient was declined for aspiration.

1.3.1 Screening of the breast using Ultra-Sound images

The most common US screening technique used for depicting the breast is Hand-Held 2D B-Mode ultrasound imaging. A manually driven transducer (see fig:1.5a) emits high-frequency mechanical waves and captures the reflection of the tissue interfaces to compose a 2D image where the brightness of each spot represents the amount of reflection for that particular position [9].

However, understanding such images is not easy. Therefore, operators and readers must have a thorough knowledge of normal breast anatomy and architecture, which has considerable variability, in order to perform an accurate diagnosis of abnormalities, since the appearance of the lesions are not specific [5], [40], [46].

Since the transducer is driven by the technician, any arbitrary slice plane of the breast can be screened. Figure 1.8 roughly illustrates the topology



(a) (b) (c)

Figure 1.7: Example of lesion shield under DM screening and distinguishable under US screening taken from Hines et al. [48]. Image (a) corresponds to a Medio-Lateral Oblique (MLO) Digital Mammography (DM), (b) is a magnification and (c) corresponds to a Breast Ultra-Sound (BUS) image.

of a breast, indicates a possible slice, and shows two US acquisitions of two healthy breasts to illustrate the structures present within the image. As can be observed in figure 1.8, several structures within the breast can be revealed when screening: skin layers, adipose tissue, fibro-glandular tissue, fibrous tissue, muscle tissue and the chest-wall to name the most important.

The specific appearance of the breast structures depend on physiological particularities of the breast depicted, as well as the acquisition instrument and its configuration, which is readjusted for every patient/image to obtain a clear view in order to perform a diagnosis through visual assessment of the images [49]. With this pretext, US systems manufacturers incorporate image processing techniques to improve the visualization for better visual reading. However, such image modifications might compromise the computerized analysis, since the image modifications are unknown and some of the operations to improve human perception cannot be undone.

Despite the variability in the appearance of breasts, some relationship between tissues hold true, especially the structural ones.

Skin is the most anterior element therefore is depicted at the top of the image, appearing as a bright layer of approximately 3mm or less, often containing a dark central line [40]. The contour and thickness of the skin layer can vary due to inflammation or disease [49].

The chest-wall, when depicted, appears as bright (highly echogenic) arched blobs, which correspond to the anterior part of the ribs and pleura. The chest-wall is the bottom structure in the image, since it corresponds

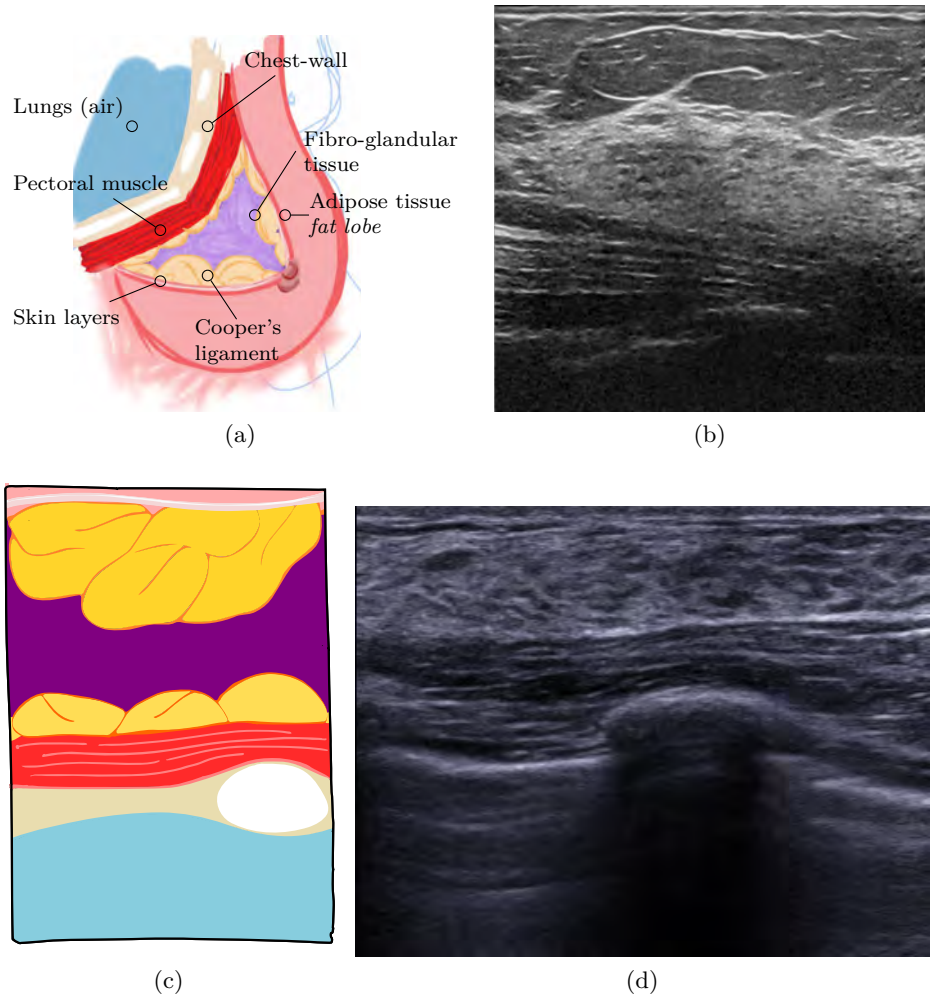


Figure 1.8: Breast structure screening appearance when using ultrasound. The illustration in (a) gives an intuitive idea of the structures present in a breast and their disposition, while illustration (c) represents how those structures are screened by a US device. Images (b) and (d) are two US images taken from healthy breasts to illustrate how the structures present in a breast are seen under US screening.

to most posterior depicted structure when screening. Just above the lungs, which appear as a noisy black area with no structure, as if it were background.

Just above the chest-wall, the pectoral muscle can easily be identified under sonography as bright elongated streams in the direction of the fibers over a dark background parallel to the skin [49].

The area compressed between the skin and the pectoral muscle corresponds to the breast structure, made up of fat lobes (along with the Cooper ligaments) and fibro-glandular tissue in a fairly variable relative amount. The normal appearance of the breast might vary from a completely fatty breast with only a few fibro-glandular structures, to a completely fibro-glandular breast with little or no fat. When a mixture of adipose and fibro-glandular tissue is present in a US screening, they normally appear in a layered fashion and adipose tissue can be found anterior (above) to fibro-glandular tissue (see fig. 1.8). It is also normal that the glandular tissue of the breast contains variable amounts of adipose infiltrations.

Figure 1.9 illustrates several breast topologies which are rated accordingly to the American College of Radiology (ACR) density rates from one to four; one being a completely fatty breast and four a completely dense breast.

When analyzing US images, the terms black, white, dark or bright are not used. Instead, terms like anechoic, hypo-echoic, iso-echoic, hyper-echoic or highly echoic are preferred. Anechoic areas are black areas with no texture due to the lack of scatterers within the tissue. As example, cystic structures show anechoic appearance, since the presence of homogeneous liquid produces no scattering (see fig.. 1.8a,b). As echogenicity reference, adipose tissue (fat) is used so that the structures depicted are denominated hypo-, iso- or hyper-echoic according to their appearance relative to normal breast adipose tissue, since adipose tissue appears near to the middle of the echogenicity spectrum. Although there are other tissues in the middle of the echogenicity spectrum, like periductal elastic tissue or terminal ductal-lobular units, adipose tissue is chosen as a reference because fat lobes are uniformly present in the population and can clearly be identified.

It is worth mentioning here the recommendation of setting the acquisition parameters of the sonographic devices so that adipose tissue appears gray rather than black. Otherwise there is not enough dynamic range to distinguish structures from tissues with a lower echoenicity response such as structures present within some solid nodules resulting in a cyst-like appearance [5].

Fat lobes appear as soft uniform, scattered textured blobs usually grayish

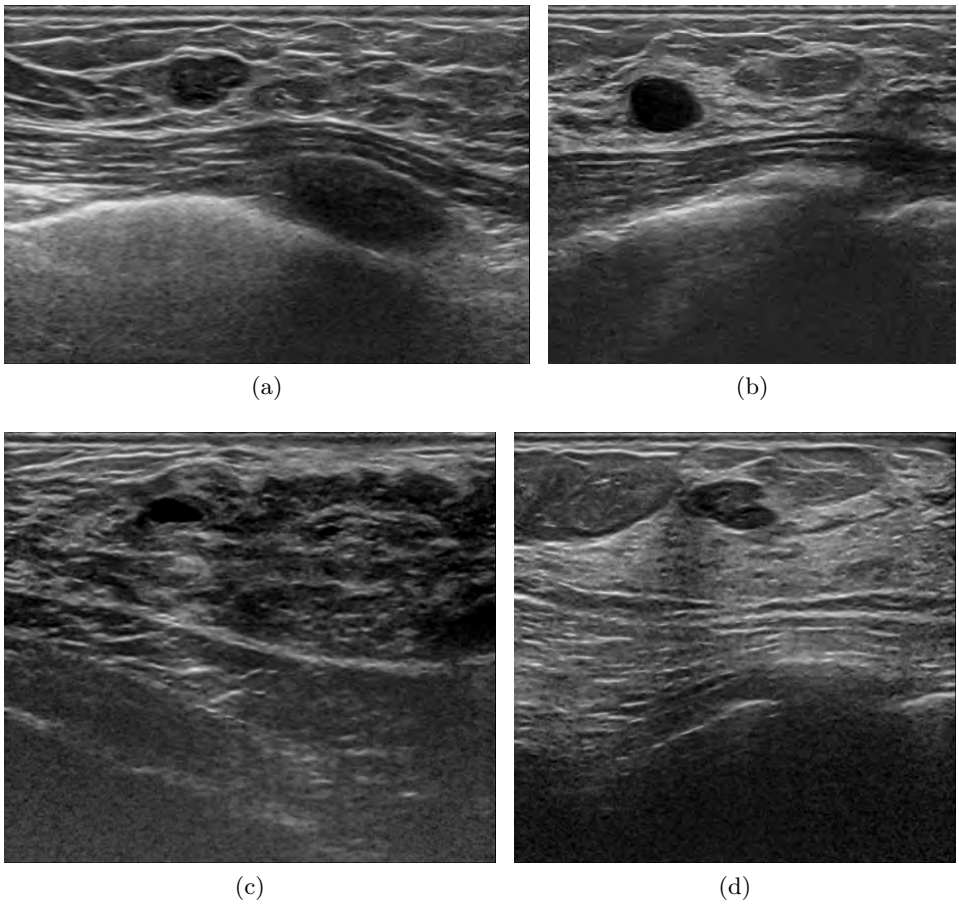


Figure 1.9: Breast Ultra-Sound (BUS) image examples of different adipose and fibro-glandular topologies with the presence of lesions. Image (a) shows a fatty breast rated as class 1 where the fat lobes are present from the skin layer all the way down to the pectoral muscle. In this image, a carcinoma intra ductal is spotted as a hypo-echogenic breast region between the skin and the pectoral muscle. The oval shaped dark area below the pectoral muscle corresponds to a rib. Image (b) illustrates a breast rated as class 2. In the image, the subcutaneous fat and fibro-glandular area beneath it can be clearly identified. An anechoic mass can be found within the fibro-glandular tissue, consistent with a cyst. In image (c), the proportion of subcutaneous fat over fibro-glandular tissue is very little. However, the darkness and uneven aspect of the fibro-glandular tissue indicates infiltrated fat combined with the fibro-glandular tissue giving an overall class 3 of breast density. Notice that within the fibro-glandular tissue, there is a completely anechoic oval spot producing slightly posterior enhancement, corresponds to a cyst. Image (d), rated as a class 4, shows a dense and homogeneous fibro-glandular pattern despite the presence of subcutaneous fat. The hypo-echoic region, with an appearance similar to an isolated fat lobe, corresponds to a fibroadenoma.

in color, (since adipose must be set as the center of the spectrum), suspended from the skin by Cooper's ligaments, which are imaged as highly-echogenic curvilinear lines extending from the breast tissue to the superficial fascial layer [40].

Fibro-glandular tissue has more scatterers, which are distributed in more locally uniform fashion compared to adipose tissue, appearing as a denser hyper-echoic textured region posterior to (under) the fat lobes. The denser the fiber, the higher the presence of scatterers within the tissue, hence the denser and brighter the texture becomes. When screened in US, fibro-glandular tissues have no apparent distribution filling the empty space between the fat lobes, or the lobes and the pectoral muscle.

1.3.2 Elements degrading Breast Ultra-Sound (BUS) images

Regardless of the clinical use of these images, they suffer from various inconveniences such as poor quality and imaging artifacts. This section tries to familiarize the reader with the elements degrading US images by commenting on their presence within example images.

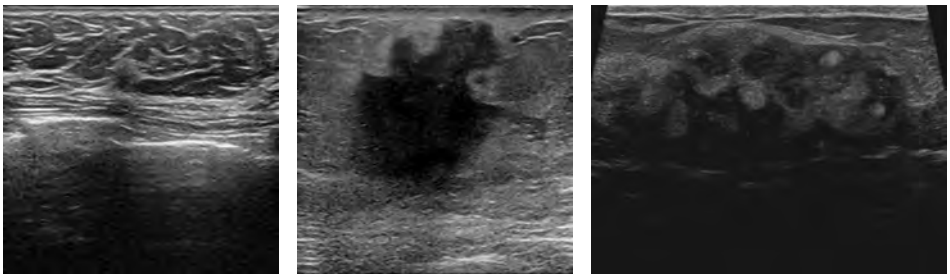
The first thing that needs to be taken into account is that these images are taken by an expert user, usually a radiologist. Therefore, the objects of interest are present and some enhancement procedures have already been applied to the image by the acquisition machinery to obtain a better visualization. All preprocessing image transformations are unknown and differ between acquisition equipment since they are proprietary.

Field of View and Zooming

The structures depicted in a US breast image are quite variable, mainly due to breast topology differences between individuals, and also due to the capabilities of sonographers to focus and zoom in on different areas. Figure 1.10 represents different BUS images where, apart from pathology diversity, the structural elements visualized in the images vary, giving a totally different images.

Weak Edges

Weak edges are produced when adjacent tissues have similar acoustic properties. An insufficient difference between speed propagation of the sound waves in two adjacent tissues yields a feebly back-reflected echo at the tissue interface, degrading the edges of US images.



(a) (b) (c)

Figure 1.10: A partial view of the structural elements of the breast. (a) shows a fatty breast with all the structural elements and an intraductal carcinoma seen as a spicular hypo-echoic region surrounded by fibrous tissue, which appears hyper echoic, producing a slightly posterior shadow in the center of the image. (b) corresponds to a zooming in on a ductal infiltrating carcinoma. Although some Cooper ligaments can be seen in the image showing that the cancer is placed in the subcutaneous fat, there's no breast structure revealed in the image. (c) shows a large hematoma with internal structure preventing the depiction of any other breast structure.

Illumination Inconsistency (shadowing and posterior enhancement artifacts)

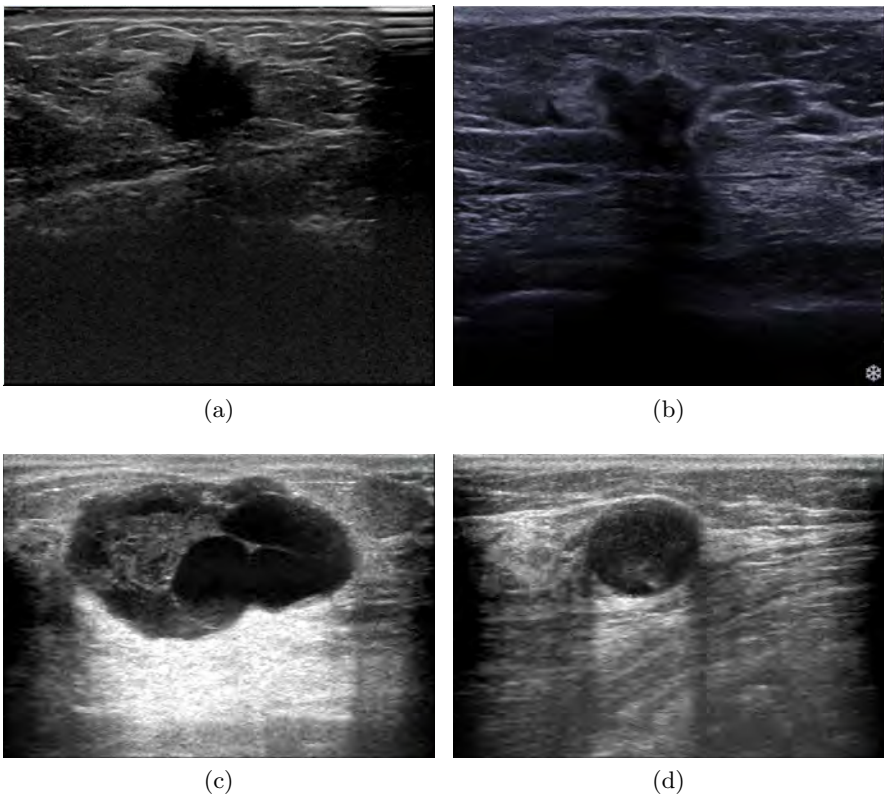
Low dynamic is the consequence of the US wave attenuation by the tissue media. As the mechanical wave travels along the tissue, the dynamic range resolution decreases, producing a lack of contrast as wave energy is dissipated. Shadowing effects occur when the signal has not got enough power to depict any further tissue due to severe attenuation. Nodules with curved surfaces may give rise to lateral refractive edge shadowing. This is seen at the edge of the lesion, not posterior to the mass [40].

Posterior acoustic enhancement has the opposite effect where posterior structures appear brighter mainly due to coherent scattering produced by fairly uniform cellularity structures or cystic lesions.

Figure 1.11 illustrates some posterior acoustic artifacts.

Speckle

Speckle is an unwanted collateral artifact coming from a coherent interface of scatterers located throughout the tissue, so that, even in uniform tissue, speckle appears as a granular structure superimposed on the image. Speckle is an artifact degrading target visibility and limits the ability to detect lower



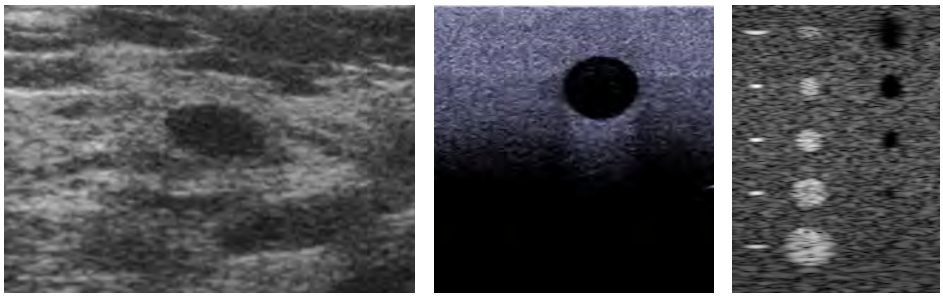
(a) (b)

(c) (d)

Figure 1.11: Illumination inhomogeneities. (a) Shadow artifact (located on the right of the image) produced by inadequate contact of the hand-held probe with the breast. (b) Posterior shadow produced by a solid mass. (c) Posterior enhancement example. (d) Combined pattern of posterior enhancement and refractive edge shadow produced by a round cyst.
In the following other image examples that qualified for the same categories can be found.
Solid mass shadow as in (b): 1.16d.
Posterior enhancment as in (c): 1.13b, 1.19b,c, & 1.15d.
Combined pattern as in (d): 1.16b,c.
No posterior pattern, neither posterior shadow nor enhancement, can be found in: 1.10a-c, 1.13a,c,d, 1.19a, 1.14a,b, 1.16a,e, 1.17a,b, 1.15a-c,e

contrast lesions in US images.

In order to illustrate speckle, figure 1.12 shows a breast screening, a physical phantom screening and a synthetic phantom image in order to show that this unwanted granular texture called speckle is characteristic of US images.



(a) (b) (c)

Figure 1.12: Speckle noise characteristic of Ultra-Sound (US) images. (a) A breast screening image. (b) Screening of a physical phantom of a clean simple cyst. (c) Synthetic phantom computed using Field II ultrasound simulator [50] (image taken from tool documentation).

Observe that when shadow is present, due to a solid lesion for instance, most often there is no presence of speckle beneath the total signal attenuation, making it impossible to determine the real extension of the lesion but at the same time, reveals physical information that the absence of speckle can be used for diagnosis (see fig.1.11b).

1.3.3 Breast lesion assessment based on Ultra-Sound imaging

One of the problems of interpreting medical images is that they are subject to subjectivity that lead to inconsistent diagnoses due to the lack of uniformity among the readings (intra- and inter-observer variability) [14]. Therefore, efforts were made to build up a set of lexicon tools [11]–[13] which are standardized descriptors that set up a common framework facilitating BUS image interpretation and allowing easy comparison and interpretation by experts. Although some indeterminate categories still persist, the development of these interpretive criteria has improved the ability to differentiate benign from malignant masses to the point that these lexicons are considered one of the most important advances in breast US [40].

Stavros et al. [5] collected the features describing the lesions that had been used previously and proposed a preliminary lexicon to describe the lesions and set the bases to perform diagnosis of solid lesions rather than just discriminate between cyst and solid lesion. In order to increase the consistency and reproducibility when assessing breast lesions using US screening, the ACR society published the US Breast Imaging-Reporting and Data System (BI-RADS) [12] lexicon as an extension of the existing and widely accepted BI-RADS standard descriptors for mammography screening.

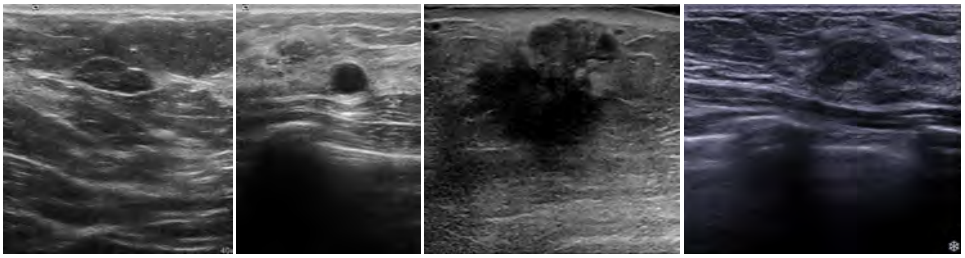
The diagnosis criteria was designed using primary signs referred to characteristics of the mass itself, and secondary signs referring to produced changes in the tissues surrounding the mass [11].

Another example is the work carried out by Hong et al. [51] studying the correctness of the diagnosis based on the lexicon descriptors proposed in [13] and [12] and comparing both lexicons in terms of PPV and NPV which represent the percentage of properly diagnosed cases based on a particular test (lexicon descriptions in this case) [15]. In the experiment, 403 images with single lesions were analyzed by one of the three experts participating in the experiment, using both lexicons to describe the images in order to compare the lexicons. The results proved the usefulness of using these lexicons for assessing solid masses and also reported the highly predictive value when using BI-RADS descriptors for assessing solid lesions. The results supporting the usefulness of the lexicon are a consensus from the medical community [52], [53].

Once the BUS imagery power of diagnosis was established, along with the development of reliable lexicons that facilitate the diagnosis, recent studies, such as Calas et al.[54], analyzed the repeatability and inter-obeserver variability in the diagnosis. In this paper, a set of 40 images is reviewed by 14 expert radiologists with 4 to 23 years experience who have all been using the BI-RADS lexicon since 2005. This study corroborates the utility and stability in the assessment of using these descriptiors for describing lesions to perform a diagnosis. However, the study reveals the increasing disagreement among the experts when the lesion size is small since it is more difficult to properly describe the lesion in the lexicon terms; an issue that would need to be addressed by reviewing and improving the lexicon in the future. The study also confirms how challenging it is to perform a diagnosis based only on a single US image. They found that some experts (8 out of 14), for a particular image sample, miss-classified a meullary carnicoma as benign, since this type of carcinoma is characterized by a partially circumscribed contour and a discrete posterior acoustic enhancement that can be confused with a complicated cyst.

Figure 1.18 illustrates the BI-RADS lexicon proposed by ACR and how those findings are distributed across different lesion types used as examples. For each feature, a single attribute must be chosen; the one which best describes the scenario. Figures 1.13-1.15, try to familiarize the reader with how similar the interpretative features of the lexicon are. These features can force the reader to analyze primary signs (those characterizing the mass itself) and secondary signs which describe the tissues surrounding the lesions.

As a primary sign, the shape, orientation and internal echo-patterns of the mass are analyzed along with the interface between the mass and the surrounding tissue. Figure 1.13 illustrates the mass shape criteria, where an oval indicates elliptical or egg-shaped lesions, A round shape indicates spherical, ball-like, circular or globular lesions. A lobular shape indicates that the lesion has two or three ondulations and an irregular shape is for any lesion that can not be classified in the previous categories.



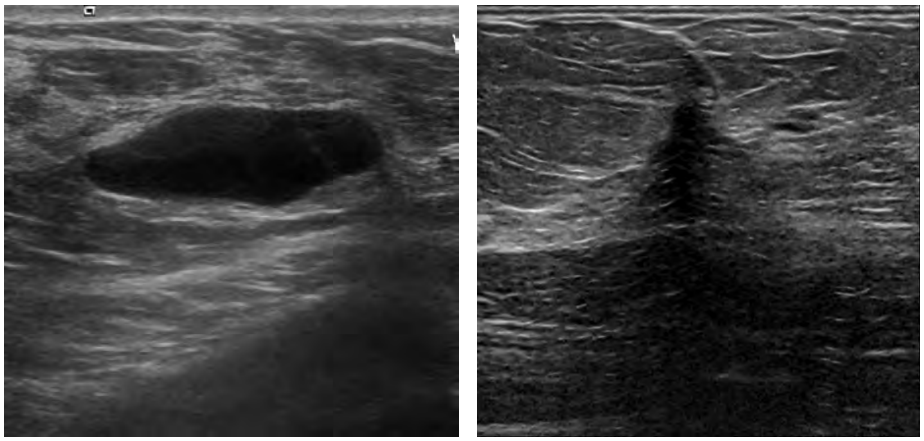
(a) (b) (c) (d)

Figure 1.13: Mass shape examples: (a) Oval shaped lesion. (b) Round masses. (c) Irregular shaped masses. (d) Lobular masses.
In the following other image examples that qualified for the same categories can be found.
Oval shaped lesion (a): 1.10c, 1.19c, 1.16a, 1.17a, 1.15a,c,e.
Round masses (b): 1.11d, 1.10a, 1.13b, 1.19a,b, 1.17b & 1.15b.
Irregular shaped masses (c): 1.11a, 1.10b, 1.14b, &1.16b,c,e.
Lobular masses (d): 1.11c, 1.14a, 1.16d & 1.15d.

Figure 1.14 illustrates mass orientation which can be parallel when the long axis of the lesion keeps the same orientation of the fibers so that the lesion doesn't cross tissue layers ("wider than tall" criteria). Non-parallel ("taller than wide") indicates a growth across the tissue layers.

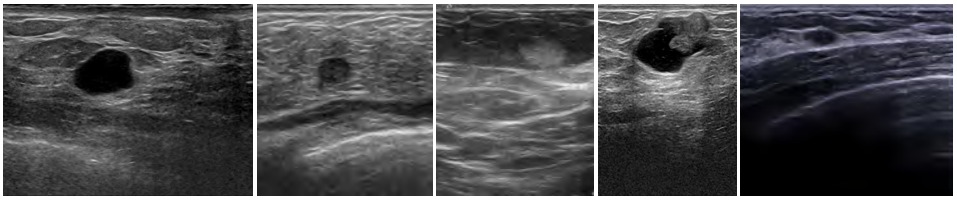
Figure 1.15 illustrates the internal echo pattern criteria, which describes the mass echogenity with respect to fat.

Figure 1.16 illustrates the mass margin criteria, describing the shape of



(a) (b)

Figure 1.14: Mass orientation: (a) Parallel to the skin. (b) Non-parallel to the skin.
In the following other image examples that qualified for the same categories can be found.
Parallel to the skin (a): 1.11c 1.10c, 1.13a,d, 1.19a-c, 1.14a, 1.16a,d, 1.17a, & 1.15a-c,e.
Non-parallel to the skin (b): 1.11a,d 1.10a,b, 1.13b,c, 1.14b, 1.16b,c,e, 1.17b, & 1.15d also qualify as non-parallel oriented lesions.



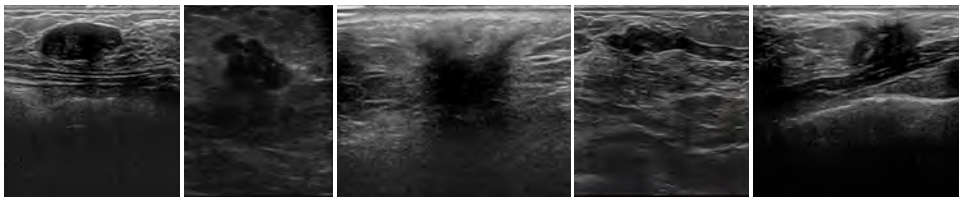
(a) (b) (c) (d) (e)

Figure 1.15: Interior echo-pattern of the mass: (a) Anechoic. (b) Hypo-echoic. (c) Hyper-echoic. (d) Complex. (e) Iso-echoic.
In the following other image examples that qualified for the same categories can be found.
Anechoic (a): 1.13b, 1.19a-c, & 1.14a also qualify as anechoic lesions.
Hypo-echoic (b):1.11a,d 1.10a,b, 1.13c,d, 1.14b, 1.16a-e, 1.17a,b, & 1.15b also qualify as lesions with an abrupt interface.
Complex (d): 1.11c, 1.10c & 1.15d also qualify as masses with complex internal echopattern.
Iso-echoic (e): 1.13a & 1.15e also qualify as masses with iso-echoic internal echopattern.

the interface between the lesion and the tissue which can be circumscribed when the interface is smooth and distinguishable, even if the rim is thick, thin or non-perceptible. Indistinct is used in cases where delineating a proper boundary would be difficult since the lesion fades within the surrounding tissue. Angular is when part of the margin is formed by linear intersections that form acute angles. Microlobulated is when the margin is characterized by more than 3 small ondulations. Spiculated is applied when the margin is characterized by sharp projecting lines.

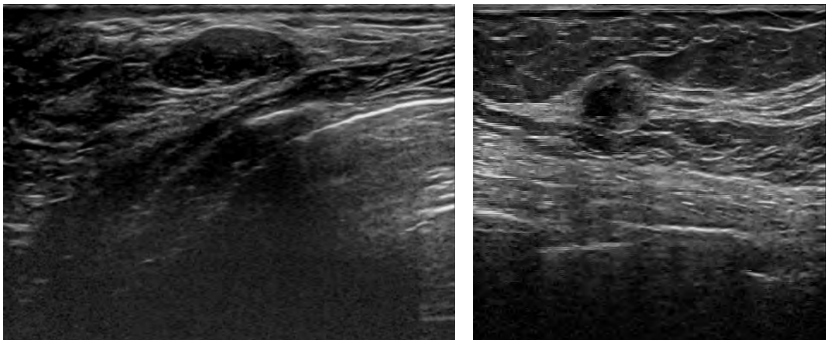
Figure 1.17 illustrates the lesion boundary criteria describing the transition between the mass and the surrounding tissue. Abrupt is used when there is a sudden change in contraposition of the echogenic halo which happens when the lesions develop a fibrous layer covering them.

The secondary signs describing the surrounding tissue are composed by the background echo-texture and the posterior acoustic pattern (see fig. 1.19 and 1.11).



(a) (b) (c) (d) (e)

Figure 1.16: Mass Margin description: (a) Circumscribed. (b) Indistinct. (c) Angular. (d) Microlobulated. (d) Sipiculated.
In the following other image examples that qualified for the same categories can be found.
Circumscribed (a): 1.11c,d 1.10c, 1.13a,b, 1.19a-c, 1.14a, 1.17a, & 1.15a-c,e also qualify as circumscribed lesions.
Angular (c): 1.10b, 1.13c, & 1.14b also qualify as lesions with an angular margin.
Microlobulated (d):1.10a, 1.13d, 1.17b, & 1.15d also qualify as microlobulated lesions.
Spiculated (e): 1.11a & 1.16e also qualify as spiculated lesions.



(a) (b)

Figure 1.17: Lesion Boundary: (a) Abrupt interface. (b) Echogenic halo.
In the following other image examples that qualified for the same categories can be found.
Abrupt interface (a): 1.11c,d 1.10c, 1.13a,b,d, 1.19a-c, 1.14a, 1.16a,b,d, 1.17a, & 1.15a-e also qualify as lesions with an abrupt interface.
Echogenic halo (b): 1.11a, 1.10a,b, 1.13c, 1.14b, 1.16c,e, & 1.17b also qualify as lesions surrounded by an echogenic halo.

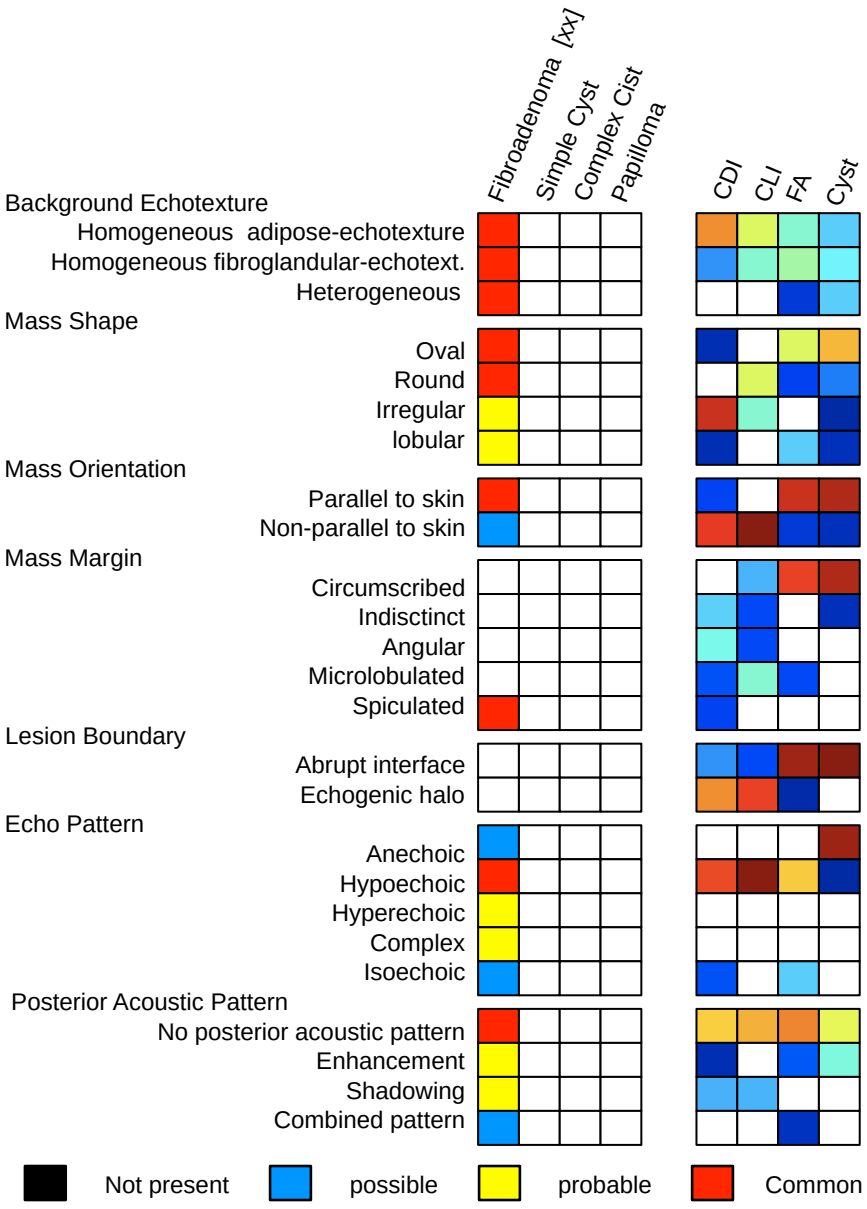
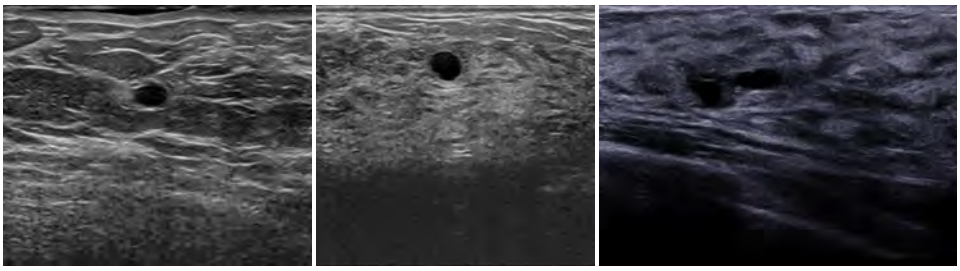


Figure 1.18: Breast Imaging-Reporting and Data System (BI-RADS) descriptors for assessing breast lesions in US images and their occurrences across several lesion types.



(a) (b) (c)

Figure 1.19: Background echo-texture: (a) Homogeneous adipose-echotexture. (b) Homogeneous fibro-glandular-echotexture. (c) Heterogenous echo-texture.
In the following other image examples that qualified for the same categories can be found.
Homogeneous adipose-echotexture (a): 1.11a,c, 1.10a-c, 1.13a,c, 1.14b, 1.16a,b, & 1.15b,c,d' also qualify as masses surrounded by homogeneous adipose echotexture.
Homogeneous fibro-glandular-echotexture (b): 1.11d, 1.13d, 1.14a, 1.16c-e, 1.17b, & 1.15a,e also qulify as masses surrounded by homogenoeus figro-glandular echotexture.
Heterogenous echo-texture (c): 1.13b, 1.19c, & 1.17a also qualify as masses in a heterogeneous background.

1.4 Computer Aided Diagnosis (CAD)

Radiologists infer the patients' state of health based on visual inspection of images depicting the existing conditions of the patient captured with a screening technique such as X-Ray radiography, Ultra-Sound (US), MRI, etc. Radiologic diagnosis error rates are similar to those found in any other task requiring human visual inspection, and such errors are subject to the quality of the images and the ability of the reader to interpret the physical properties depicted in them[10].

Providing better instrumentation in order to improve the quality of the images as well as methodologies and procedures in order to improve the interpretation of the readings have been the major goal of researchers and developers in the medical imaging field. Although the idea of using computer systems to analyze radiographic abnormalities has been around since the mid-1950s [55], the development of such ideas is still undergoing and unsolved due to technological limitations in computational power since the volume of the data within the images and the nature of the procedures for analyzing the data are in some cases intractable. Studies such as Chan et al. [56] support those thesis that state that the use of a computer, in this case for spotting microcalsification clusters in mammography images, produces statistically significant improvement in radiologists' performance. Since the goal of medical imaging is to provide information to the radiologists that reduces the diagnosis uncertainty by either reducing the errors when searching abnormalities, reducing interpretation errors or reducing the variation among observers.

Anything that helps the radiologists to perform a diagnosis can be considered as CAD, from a data visualization system to a fully integrated system that, from an input, image outputs a final diagnosis that can be taken as a second reading. Despite the wide coverage of CAD, such techniques and systems can be broadly categorized into two types: Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx) [16].

CADe implies that radiologists use computer outputs of the locations of suspect regions, leaving the characterization, diagnosis, and patient management to be done manually.

CADx extends the computer analyses to yield output on the characterization of a region or lesion, initially located by either a human or a CADe system.

1.4.1 Image segmentation applied to BUS segmentation for CADx applications

As stated earlier, the lexicon descriptors proposed in [13] and [12] have proven to be a useful framework for radiologists when analyzing BUS images. The PPV and NPV when describing lesions with these tools turned them into the standard for human reading and diagnosis based on BUS images.

One of the advantages of using CAD systems is that computerized systems can take advantage of other low-level information that is usually hidden to a human reader. Although there are some designs based only on low-level features, such as the approach proposed by Liu et al. [57], most of them combine both low- and high-level features. High-level cognitive features, like lexicons, are subject to an accurate delineation of the lesions so that features can be extracted. Moreover, the use of high-level features based on segmentations similar to lexicons brings the CAD system closer to the radiologist routines, facilitating the decision making which is the final goal of a CAD system.

Therefore, segmentation is a key step for CAD systems that might be seen as a CADe procedure or as an intermediate step between CADe and CADx if this segmentation is somehow guided by the user. However, segmentation is not an easy task to perform.

Image segmentation is the process of partitioning an image into multiple meaningful segments which simplifies the further analysis of the image. Any segmentation procedure needs to address two aspects: targeting the structures that one wants to identify, and dealing with the noise present in the image. In our case, we are aiming for an accurate delineation of lesions with a low false positive rate without mistaking similar structures as a lesions.

1.5 Thesis Objectives

Summing up, US imagery automatized analysis is challenging in general, and in particular for breast lesion assessment since it is one of the most difficult tasks to perform due to all the aforesaid drawbacks. However, the clinical value of assessing breast lesions in US data [5], [12], [13], [51], [54], justifies the growing interest within the medial imaging field of addressing BUS-CAD systems. Moreover, the lexicon tools developed to improve the understanding among radiologists have proven to be useful for assessing breast lesions. However, these descriptors are subject to an accurate delineation of the lesion which when read by an expert radiologist is instantly understood.

Our goal is to propose a fully automatic segmentation procedure able to delineate the lesions as well as fully partition tissues of interest within the image so that high-cognitive features can be extracted for driving CADx systems. Although various projects have addressed the problem of breast lesion segmentation in US data, such as as [58]–[61], the segmentation task remains unsolved.

1.6 Thesis Organization

This thesis is structured as follows: This first chapter introduces US imaging modality for assessing breast lesions, the importance of CAD systems for accurate readings of breast ultrasound imagery, and the role of segmentation in order to obtain high-level information that can be used to develop more accurate CAD systems. A description of the objectives of this thesis and this organization summary can also be found in the first chapter. Chapter 2 analyses the state-of-the-art of image segmentation techniques applied to automatic breast lesion delineation in ultrasound data. Chapter 3 proposes an easy to modify framework not only to delineate the lesions but also to delineate other structures of interest present in the images. The proposed framework, consists of building up an objective function that is further minimized. This chapter covers all the parts of the proposed framework as well as reporting the experiments carried out and a discussion of the outcome. Finally, the thesis ends with some conclusions wrapping up the work exposed here and proposes research lines for further work.

Chapter 2

A review of current methodologies for segmenting breast lesions in Ultra-Sound images

B. Watterson

US imaging has proven to be a successful adjunct image modality for breast cancer screening [3], [4], especially in view of the discriminative capabilities that US offers for differentiating between benign or malignant solid lesions [7]. As a result, the number of unnecessary biopsies, which is estimated to be between $65 \sim 85\%$ of all prescribed biopsies [6], can be reduced [7] with the added advantage of a close follow-up with sequential scans [8].

However, the noisy nature of the US image modality and the presence of strong artifacts, both degrading the overall image quality [9], raise diagnosis error rates as would happen in any other human visual inspection task [10]. Therefore, uniform terms in order to reduce diagnosis inconsistencies among

readers [14] characterizing, describing and reporting the lesions have been developed [5], [11]–[13] so that double readings can be performed and a more accurate diagnosis achieved. The main inconvenience of double readings is cost, justifying the use of CAD systems, which have also proven to improve diagnosis accuracy [16].

BUS CADx, as mentioned earlier, can take advantage of either low-level features, high-level features or both [62]. However, in order to take advantage of high-level features or descriptors similar to the lexicon descriptors proposed in [12], [13], an accurate segmentation is needed (see section 1.3.3).

2.1 The role of segmentation within a Breast ultrasound Computer Aided Diagnosis (CAD) system

Segmentation is a fundamental procedure for a CAD system. Figure 2.1 illustrates the idea that procedures for segmentating breast lesions in US data can be found within a CAD system workflow as part of CADe, as part of CADx or as a stand alone step using detection information and providing further information that can be used for conducting a diagnosis.

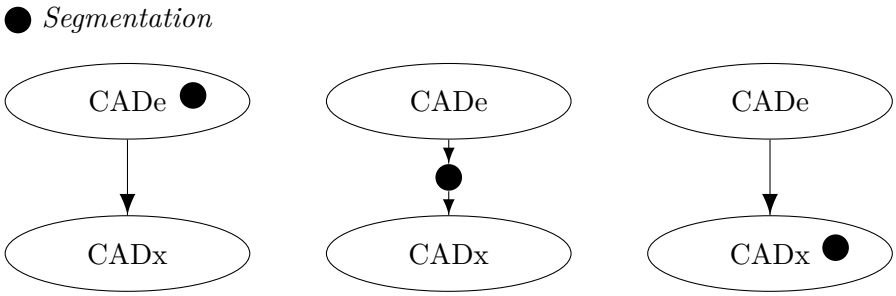


Figure 2.1: Illustrative idea of the role of segmentation within a CAD framework showing that it can either be a separate process between a CADe and a CADx or it can belong to any of the two CAD typologies: CADe, CADx

Segmentation procedures integrated within CAD systems can either be manual, interactive or automatic depending on the amount of effort or data supplied by the user. CADx systems needing high-level descriptors supplied by a user or a non-aided manual delineation also fall into the manual category and therefore, are not extensively reviewed. As an example of this

category, we cite the work presented by Hong et al. [51], which describes a system working on BI-RADS descriptors supplied by an expert based on the reading of images.

Figure 2.2 compiles methodologies of interest and categorizes them according to the following groups and subgroups:

Interactive Segmentation: methodologies requiring any kind of user interaction to drive the segmentation.

- *Fully-Guided* are those methodologies where the user is asked to accompany the method through the desired delineation.
- *Semi-Automatic* are those methodologies where the segmentation is conditioned by the user by means of labeling the regions instead of the delineation path.

Automatic Segmentation: methodologies with no user interaction.

- *Auto-Guided* are an evolution of Semi-Automatic methodologies so that user interaction has been substituted by an automatic procedure (usually as an automatic initialization of the original Semi-Automatic procedure).
- *Fully-Automatic* are ad-hoc automatic procedures designed in such a manner that no user interaction can be incorporated.

2.1.1 Interactive Segmentation

While fully automatic segmentation still remains unsolved, it is obvious that manual delineations are unacceptably laborious and the results suffer from huge inter- and intra-user variability, which reveals its inherent inaccuracy. Thus, interactive segmentation is rising as a popular alternative alleviating the inherent problems in fully automatic or manual segmentation by taking advantage of the user to assist the segmentation procedure. Interactive methodologies are mainly designed as general purpose techniques since the segmentation is controlled by a skilled user who supplies the knowledge regarding the application domain. Depending on the typology of information the user provides the system in order to govern the segmentation, two distinct strategies can be differentiated: *fully-guided* and *semi-automatic*.

For a fully-guided strategy, the user indicates the boundary of the desired segmentation and accompanies the procedure along the whole path. Some

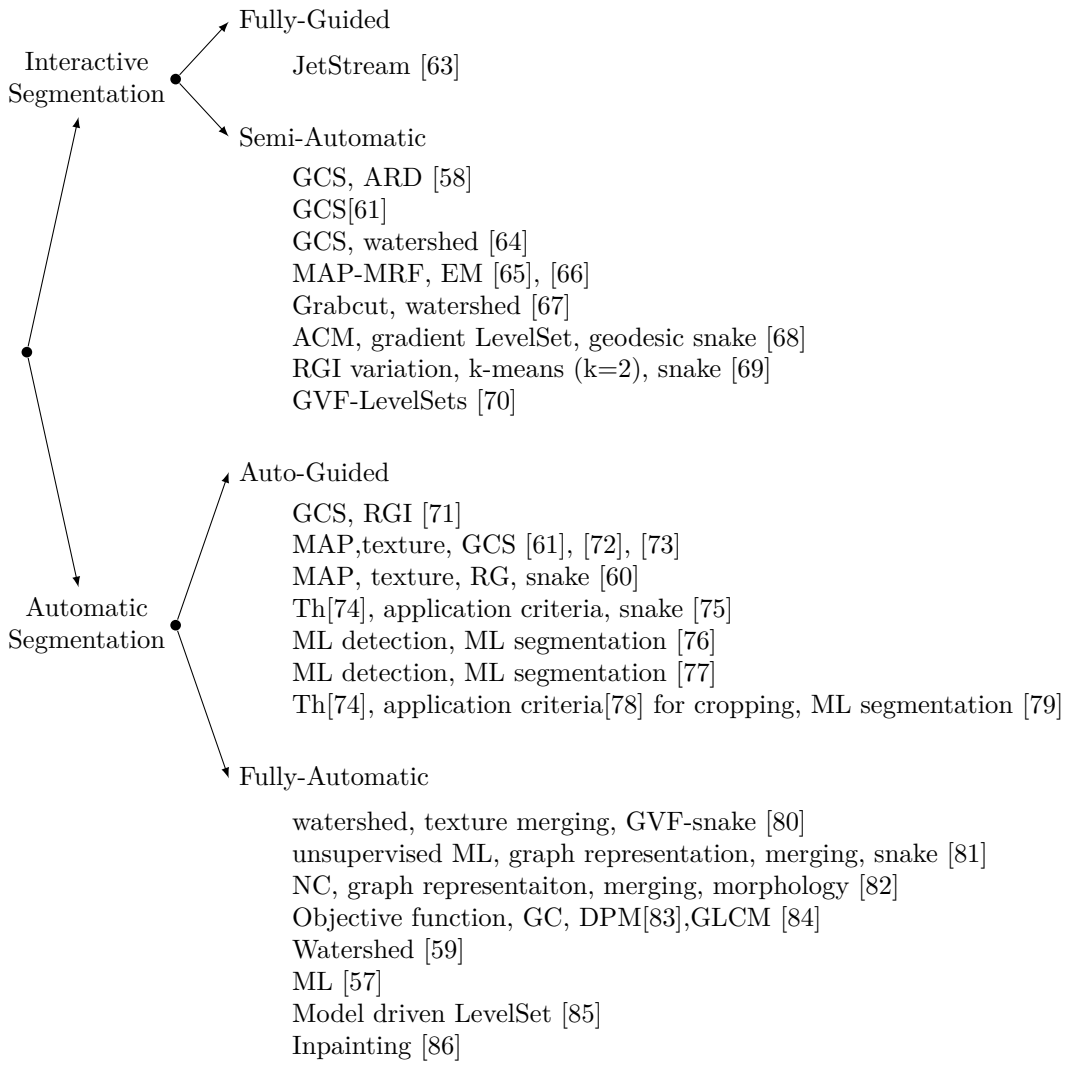


Figure 2.2: List of breast lesion segmentation methodologies and their highlights. The methodologies are groped in two categories: interactive and automatic; with four subcategories: Fully-Guided, Semi-Automatic, Auto-Guided and Fully-Automatic.

successful general purpose techniques that require this kind of user interaction, and just to name a couple, are: *intelligent-scissors* [87], or *Jetstream* segmentation [88], both deriving from the *live-wire* technique [89], which requires the user to indicate roughly the path of the desired boundary and the segmentation procedure automatically adjusts to the underlying desired partition in an interactive manner.

For a semi-automatic strategy, the user constrains or initializes the segmentation procedure by indicating parts or elements belonging to each object to be segmented (i.e. foreground/background). The segmentation procedure generates the final delineation from this information. Two popular general purpose interactive segmentation techniques falling in this category are: *lazy snapping* [90] and *grabcut* [91] both based on the work proposed by Boykov and Jolly [92] which takes advantage of GC and a naive indication of the elements present within the image to find a proper delineation of the object of interest.

Although interactive segmentation procedures are designed in a general manner, due to the difficulties present in US images, some interactive segmentation procedures especially designed for delineating breast lesions in US data have been developed. The remainder of this section compiles these procedures in terms of the aforementioned fully-guided and semi-automatic terms.

Fully-guided interactive segmentation applied to Breast Ultrasound images

Due to the quantity of knowledge extracted from the user when segmenting with a fully-guided interactive procedure, it is rare to find a fully-guided segmentation designed for a particular application. However, Angelova and Mihaylova [63], [93] implemented a jetstream [88] especially designed to be applied to segment breast lesions in US data images.

It can be argued that their proposal is not a fully-guided procedure as the authors have limited the user interactivity since it is not allowed to condition the segmentation along the whole path. The method is initialized by four point locations indicating the center of the lesion, an inner bound, an outer bound, and a point lying within the desired boundary. These four locations drive the whole segmentation that takes advantage of intensity and position information. In this sense the methodology can be categorized as semi-automatic. However, it has been considered fully-guided since it is based on a fully-guided procedure, namely jet stream. Implementation of multiple reinitialization of the boundary location in order to achieve fully-

guidance is straight forward despite not being covered in the original work.

The evaluation of the method is done in a qualitative manner using a dataset of 20 images. No quantitative results are reported.

Semi-automatic segmentation applied to Breast Ultrasound images

In this section we consider semi-automatic segmentation methods; those methods requiring the user to impose certain hard constraints like indicating that certain pixels (seeds) belong to a particular object (either lesion or background).

Horsch et al. [58] propose using a Gaussian Constraining Segmentation (GCS) consisting of combining a Gaussian shape totally or partially defined by the user with an intensity dependent function. The final segmentation consists of finding the contour resulting from thresholding the Gaussian constrained function that maximizes the Average Radial Derivative (ARD) measure. The maximization is done in an exhaustive manner. The segmentation performance was tested on a 400 image dataset achieving a mean Area Overlap (AOV) of 0.73 when compared to manual delineation by an expert radiologist. Massich et al. [61] proposed a methodology inspired by GCS with different user interactability levels that fall into the interactive and semi-automatic procedures category when manually initialized with a single click. The difference between this work and the original GCS methodology lies in the intensity dependent function and the manner in which the final threshold is chosen since a disparity measure is minimized instead of maximizing the ARD coefficient. In this proposal, the intensity dependent function used is robust to the thresholding so that if, instead of dynamically choosing a thresholding based on the error measure or ARD, a fixed threshold (properly tuned for the dataset) is preferred, the segmentation results are consistent. Although a slightly lower performance in terms of mean is reported, 0.66 compared to 0.73 obtained by the original GCS methodology, there is no difference statistically when comparing the result distribution in a common dataset [61], and the methodology proposed by Massich et al. demands less user interaction. Another work based on GCS [58] is the work proposed by Gomez et al. [64] where watershed transform is used to condition the intensity dependent function. As in the original GCS proposal, ARD maximization is used in order to find the adequate threshold that leads to the final segmentation. Although a larger dataset should be used in order to corroborate the improvement and the fact that the multivariate Gaussian is determined by 4 points supplied by the user, a mean overlap of 0.85 is

reported using a 20 image dataset.

In Xiao et al. [65], the user is required to determine different Regions Of Interest (ROIs) placed inside and outside the lesion in order to extract the intensity distribution of both. Then, these distributions are used to drive an Expectation Maximization (EM) procedure over the intensity spectrum of the image incorporating a Markov Random Field (MRF) used for both smoothing the segmentation and estimating the distortion field. Although in [65] the method is only qualitatively evaluated in a reduced set of synthetic and real data, further studies reducing the user interaction from different ROIs to a single click [66] reported results using two larger datasets of 212 and 140 images obtaining an AOV of 0.508 for the original method and 0.55 for the less interactive proposal, and a Dice Similarity Coefficient (DSC) score of 0.61 and 0.66 respectively.

Other examples of semi-automatic procedures addressing segmentation of breast lesions in US images are: the implementation of the grab-cut methodology proposed by Chiang et al. [67] or the various manually initialized implementations of the popular Active Contour Models (ACMs) technique [68]–[70]. These ACM methodologies reported really good results achieving a mean AOV of 0.883 for the implementation presented in [68]. Within the group of methodologies using ACM, Alemán-Flores et al. [68] connected two completely different ACM procedures in a daisy-chain manner. First, the image is simplified by applying a modified Anisotropic Diffusion Filter (ADF) that takes texture into account, using the Gabor filter responses to drive the amount of diffusion. Then, a manual seed is used to initialize a gradient regularized LevelSet method as if it were a region growing procedure growing in the simplified image. Finally, the pre-segmentation[1] obtained is used to initialize a geodesic snake ACM that evolves using intensity information from the inner and outer parts. In a similar way, Cui et al. [69] evolves two ACMs in a daisy chain manner. However, in this case the ACMs are identical, differing only in their initialization. Finally, the best solution from the two ACMs is selected. A mean AOV of 0.74 was reported on a large dataset of 488 images. Gao et al. [70] tested on a small dataset of 20 images the use of a GVF-based LevelSet ACM that also took into account the phase congruency texture [94] along with the gradient information, achieving a mean AOV of 0.863.

[1]The segmentation obtained from the first ACM procedure.

2.1.2 Automatic Segmentation

Although automatic segmentation of breast lesions in ultrasound images remains unsolved, huge efforts to obtain lesion delineations with no user interaction have been made in the last few years. In order to categorize the automatic segmentation methodologies, two distinct strategies when designing the methodologies have been adopted for classification: methodologies automatizing semi-automatic procedures so that no user interaction is required, and ad-hoc methodologies designed in a manner that no element can be substituted by user supplied information.

The former has been named *auto-guided* procedures since for this case the information supplied by the user has been substituted by an automatic methodology that guides the semi-automatic segmentation, while the latter have been identified as *fully automatic* procedures.

Notice that for this work, only methodologies outputting a segmentation are reviewed. Therefore, CADe procedures that can be used to initialize a semi-automatic procedure are out of the study unless there is explicitly paired work such as in (Drukker et al. [71] , Horsch et al. [58]) or (Shan et al. [78], (Shan et al. [79]).

Auto-guided Segmentation

Listed here are segmentation methodologies that consist of automatizing semi-automatic procedures or methodologies conceived as a two step problem: lesion detection and further segmentation of any detected lesions; methodologies that in some sense can be seen as a decoupled CADe and further segmentation.

A clear example of this group is the work proposed by Drukker et al. [71] where an automatic detection procedure is added to the original GCS segmentation [58] eliminating user interaction.

In order to properly detect the lesion to successfully delineate it using GCS, several rough GCS segmentations are performed in a sparse regular grid. Every position on the grid is constrained (one at a time) with a constant bivariate Gaussian function. The resulting Gaussian constrained image depending function is thresholded at several levels in order to generate a set of delineations. The Radial Gradient Index (RGI)[2] is calculated for all the delineations of every delineation set. The maximum RGI reward of every delineation set is used to generate a low resolution image which is

[2]This differs from the GCS procedure used for the final delineation since ARD index is used.

thresholded to determine an approximation of the lesion's boundaries. This approximation is used to determine a seed point in order to control the final segmentation as proposed in [58]. The method was evaluated solely as a detection in a 757 image dataset achieving a TPR of 0.87 and a FPR of 0.76.

Massich et al. [61] also proposed a methodology based on GCS as [71] with several levels of user interaction contemplating the no user interaction scenario. The method consists of a 4 step procedure: seed placement procedure (CADe), a fuzzy region growing, a multivariate gaussian determination, and finally, a GCS. The seed placement produces an initial region that is further expanded. Once expanded, the final region is used to determine a multivariate Gaussian which can have any orientation. This is an improvement with respect to the original GCS formulation in [58] allowing better description of oblique lesions since, in the original work, only Gaussian functions orthogonal to the image axis were considered. Similar to the original work, this constraining Gaussian function is used to constrain an intensity dependent function that is thresholded in order to obtain the final delineation. The intensity dependent function and the manner of determining the most appropriate threshold differ in the two proposals. The method is evaluated using a dataset of 25 images with multiple Ground Truth (GT) annotations. For evaluation purposes, the multiple annotations are combined using Simultaneous Truth and Performance Level Estimation (STAPLE) [95] in order to obtain the Hidden Ground Truth (HGT). Then the methodology is assessed in terms of area overlap with the merging of the delineations weighted by the HGT saliency, achieving a reward coefficient of 0.64 with no user interaction. Those results are comparable to the results achieved by [58] since segmentations obtained from missed or wrongly detected lesions were also taken into account to produce the assessing results. Further details on the exact seed placement algorithm can be found in [72], [73]. This seed placement is based on a multi-feature Bayesian Machine Learning (ML) framework to determine whether a particular pixel in the image is a lesion or not. From the learning step, a Maximum A Posteriori (MAP) probability plane of the target image is obtained and thresholded with certain confidence (0.8 as reported in [73]). Then the largest area is selected as the candidate region for further expansion. Due to the sparseness of the data within the feature space, Independent and Identically Distributed (IID) is assumed so that MAP can be calculated from the marginals of each feature, a fact that does not always hold indicates that more complex models are needed.

Madabhushi and Metaxas [60] proposed using the *Stavros Criteria* [13]

to determine which pixels are most likely to be part of a lesion. The *Stavros Criteria* integrate the posterior probability of intensity and texture (also assuming IID) constraining it with a heuristic taking into account the position of the pixel. The best scoring pixel is used to initialize a region growing procedure outputting a preliminary segmentation of the lesion. This preliminary delineation is then sampled for initializing an ACM procedure that takes into account the gradient information of the image to deform the preliminary segmentation into the final segmentation. A dataset of 42 images is used in order to evaluate the methodology in terms of boundary error and area overlap. The average mean boundary error between the automated and the GT is reported to be 6.6 pixels. Meanwhile, the area overlap is reported in terms of False Positive (FP) area (0.209), False Negative (FN) area (0.25) and True Positive (TP) area (0.75) which can be used to calculate an area overlap coefficient of 0.621 in order to compare with the other methodologies. As an alternative, Huang et al. [75] proposed using a LevelSet ACM using a rather heuristic initialization and also evolving using intensity gradient. The initialization is obtained by simplifying the image using Modified Curvature Diffusion Equation (MCDE), which has been demonstrated to be more aggressive than ADF, then the Otsu automatic thresholding procedure [74] is used to generate candidate blobs with the bounding box ROI of the selected one is used as initialization for the LevelSet procedure. The selection of the best blob is done by taking into account application domain information such as preference for larger areas not in contact with the image borders similar to the recall measure proposed by Shan et al. [78]. A DSC of 0.876 is reported using a dataset of 118 images.

Zhang et al. [76] and Jiang et al. [77] proposed using a two step ML procedure. The first step is a database driven supervised ML procedure for lesion detection. Detected regions with high confidence of being lesion and non-lesion are further used to learn the appearance model of the lesion within the target image. The second step consists of a supervised ML segmentation procedure trained on the target image using the previously detected regions. Both methods fall into the category of auto-guided procedures because the first ML step is used to substitute the detection information which can be directly exchanged by a user interaction. Under this hypothesis of exchanging lesion detection by user interaction, the resulting methodologies reassemble to the semi-automatic methodology proposed by Xio et al. [65]. In contrast, if the statistical models used to drive the second ML step producing the final segmentation in [76], [77] were inferred from dataset annotations, then both methodologies would be considered fully-guided and would resemble the work proposed by Hao et al. [84] since the first step is usually provided

by user interaction.

If the models for the second step are determined from the database instead of the image, then the possibility of obtaining such information from the user would not exist and the methods would no longer belong tho the auto-guided category.

Unlike all previous works, Shan et al. [79] proposed using the detection just to simplify the following segmentation procedure. The lesion detection procedure described in [78] is used to crop the image into a subset of the image containing the lesion. Then a database driven supervised ML segmentation procedure is carried out in the sub image to determine a lesion/non-lesion label for all the pixels. The segmentation stage takes advantage of intensity, texture [61], energy-based phase information [96] and distance to the initially detected contour [78] as features. Notice that despite this segmentation algorithm being a database driven ML process, the crop procedure is needed to reduce the variability of labeling and such cropping can be performed by a user. Therefore the method proposed by Shan et al. [79] has been considered auto-guided, but it could be argued to be a fully automatic procedure since the distance to the initial contour is needed as a feature for the segmentation process.

In general, *auto-guided* procedures have been considered those automatic segmentation procedures that, at some point, could be substituted by a process involving the user. These methodologies are usually designed in two steps where lesions are detected and further segmented.

Fully Automatic

In opposition to *auto-guided* methodologies, *fully automatic* methodologies are considered those methods such that, at no point, can be substituted by some user interaction.

Huang and Cheng [80] proposed using an ACM to perform the final segmentation [97] operating on the gradient image. In order to initialize an ACM, a preliminary segmentation is obtained, over-segmenting the image and merging similar regions. The watershed transform [98], [99] is applied to the image intensities to obtain an over-segmentation of the image, and then, the regions are merged, depending on the region intensities and texture features extracted from Gray-Level Co-occurrence Matrix (GLCM). Although the work does not cover how to select the proper segment to use as an initial segmentation among the segments resulting after the merging, any kind of machine learning to elect the best candidate can be assumed. Similarly, Huang et al. [81] and Liu et al. [82] also split the image into regions or seg-

ments as a first step for further analysis. To determine the image segments, Huang et al. [81] use unsupervised learning and Liu et al. [82] use normalized cuts [100] in order to achieve an image over-segmentation as that obtained when applying the watershed transform in [80]. The difference between the three works lies in how the segments are managed once determined since both [81], [82] utilize a graph representation to merge similar regions. In this graph, each node represents a segment, and the edges connecting contiguous segments are defined according to some similitude criteria in the contiguous segments. Finally, the weaker edges are merged forming larger regions in an iterative manner. Notice that even when using a graph representation, the operation performed is not a graph cut minimization [92]. The graph is only a representation used to keep track the merging schedule.

Further ideas using image segments as building blocks were explored for general image understanding applications [101] and have also been applied to breast lesion segmentation in US data [84]. The most common form for such approaches consists of an objective function minimization framework where the basic atomic element representing the images are those image segments which receive the name of superpixels and the goal is to assign them either a lesion or a non-lesion label in order to perform the segmentation. The most common form of objective function usually takes into account the datamodel driving the segmentation as the output of an ML stage and combines them with regularization (or smoothing) term which imposes labeling constrains in the form of Conditional Random Field (CRF) or MRF.

In this research line, Hao et al. [84] proposed automatically segmenting breast lesions using an objective function combining Deformable Part Model (DPM) [83] detection with intensity histograms, a GLCM based texture descriptor and position information using a Graph-Cut minimization tool and normalized cuts [100] as image segments. The proposed methodology reported an average AOV of 0.75 of a 480 image database.

In contrast, Huang and Chen [59] only performed the spliting of the image using watershed transform, while Liu et al. [57] only classified image patches arguing that inaccurate delineations of the lesions also lead to good diagnosis results when using appropriated low-level features.

Liu et al. [85] incorporated a learnt model of the lesions' appearance to drive a region based LevelSet formulation. The model is obtained by fitting a Rayleigh distribution to training lesion samples and the LevelSet evolves to fit the model into the target image. The LevelSet initialization corresponds to a centered rectangle with a size of one third of the target image. Despite its naive initialization, the reported average AOV using a dataset of 76 images is 0.88. The correctness of use Rayleigh distribution in

order to model the data can be argued regardless of its popularity and the results achieved. J.A. Noble [102] questions the usage of Rayleigh models to characterize tissue in US data images since, in the final images provided by US equipment, the Rayleigh distribution of the data no longer holds.

A completely different approach is proposed by Yeh et al. [86], where a method for inpainting degraded characters is adapted to segment breast lesions in US images. The idea consists of performing local thresholding and produces a binary image and reconstructs the larger blobs as if they were degraded. Despite the originality of the method and having been tested in a rather small dataset (6 images), the reported results achieve results of AOV[3] 0.73.

2.2 Segmentation methodologies and features

Despite interaction or information constraints needed to drive segmentations, a large variety of segmentation algorithms have been proposed for general image segmentation including the particular application of breast lesion segmentation in US data. As Cremers et al. [103] pointed out, earlier segmentation approaches were often based on a set of rather heuristic processing, while optimization methods became established as straighter and more transparent methods where segmentations of a given image are obtained by standardized methods minimizing appropriate cost functionals [103]. Although the chronological difference cannot be appreciated for breast lesion segmentation since early applications such as Xio et al. [65] were already taking advantage of optimization methods. A tendency to move towards optimization methodologies, as can be seen [77], in lieu of methodologies driven by obscure heuristics in a full manner such as in [58], [61], [71] or partially like [60].

Within the optimization methods, *spatially discrete* and *spatially continuous* categories can be found. For the discrete case, the segmentation problem is formulated as a labeling problem where a set of observations (usually pixels) and labels are given, and the goal is to designate a proper label for all the observations. These problems are usually formulated as *metric labeling* problems [104] so that smoothing regularizations can be imposed to encourage neighboring elements to have similar labels. Further information in segmentation procedures posted as a labeling problem can be found in Delong et al. [105] as a continuation of the work started by Boykov et al. [104] in their seminal paper of Graph-Cut (GC).

[3]this value has been calculated from the TP, FN and FP values reported in [86]

In spatially continuous approaches, the segmentation of the image is considered an infinite-dimensional optimization problem and is solved by means of variational methods. These methods became popular with the seminal paper on *Snakes* by Kass et al. [106] where finding boundaries becomes an optimization process. *Snakes* consists of a propagating contour defined as a set of control points (explicit formulation) that evolves in accordance with the gradient of an arbitrary energy function. These functions are formulated as a set of Partial Differential Equations (PDEs) specifically designed for each application to bound an object of interest, ensuring a smooth delineation.

The same problem can also be formulated in an implicit manner where the evolving contour or surface is defined as the zero level set of a one dimension expanded function [107]. This new formulation (named *LevelSet*) overcomes limitations of *Snakes* such as naturally handling topological changes and initialization relaxation. Extension to other segmentation criteria rather than just using an intensity gradient such as color, texture or motion, which was not straight-forward in *Snakes* formulation, can easily be done.

Both formulations of the spatially continuous approaches LevelSets and Snakes compose the segmentation procedures called ACM. Although Snakes and LevelSets are intended to work with gradient information, there are geodesic extensions allowing the contour evolution to depend on region information instead of gradients [85].

Figure 2.3 maps the methodologies presented in section 2.1 (see fig. 2.2) regarding its usage of ML, ACM, and other strategies.

2.2.1 Active Contour Models (ACMs)

ACM segmentation techniques are widely applied in US applications such as organ delineation [108] or breast lesion segmentation [60], [68]–[70], [75], [80], [81], [85]. Notice in figures 2.2 and 2.3 that most of the ACM methodologies correspond to the gradient driven ACM techniques (7 out of 8). Two of them are formulated as implicit contour (LevelSet), while the remaining are formulated in an explicit manner (snakes). A known limitation of these methodologies is that the results are highly dependent on the initial estimate of the contour. Therefore, ACM has been used as a post processing step that allows an initial segmentation to be attracted towards the boundary and control the smoothness of the curve simultaneously.

Jummat et al. [109] compare some of the multiple strategies to condition and model the evolution of the snakes applied to segment breast lesions in US 3D data. In this comparison, Ballon-snakes [110] reported better

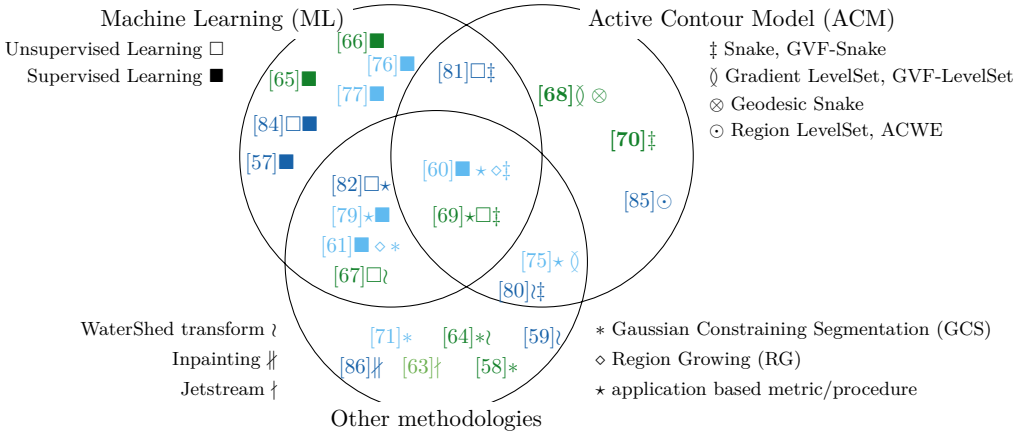


Figure 2.3: Conceptual map of the segmentation strategy used in the methodologies reported in figure 2.2. The methods have been grouped according to the segmentation methodology: ML,ACM or others. Each circle has its own iconography representing the sub-strategies that can be found in each class. The color here is used to represent user interactability being: fully guided (dark-green), semi-automatic (light-green), auto-guided(light-Blue), and fully automatic(dark-blue).

performance than GVF-Snakes [111].

However, taking everything into consideration, the segmentation results when using ACM are highly dependent on the correctness of the contour initialization. In contrast, Liu et al. [85] proposed using a model driven LevelSet approach which can use an arbitrary initialization. In this case, the initial contour is a centered arbitrary rectangle. The contour evolves, forcing the intensity distribution of the pixels of the inner part of the contour to fit a model Probability Density Function (PDF) obtained from a training step. Since it uses region information, a rather naive initialization can be used.

2.2.2 The role of Machine Learning (ML) in breast lesion segmentation

When addressing the lesion segmentation problem, two subproblems arise: a) properly detecting the lesions; and b) properly delineating the lesion. In the literature, ML has proven to be a useful and reliable tool, widely used to address either one of those two subproblems or both (either in a daisy-chain manner or at once). ML uses elements with a provided ground truth

(i.e. lesion/non-lesion) to build up a model for predicting or inferring the nature of elements with no ground truth provided within the models. The stochastic models built up from a training procedure can be used to drive optimization frameworks for segmenting.

ML techniques, strategies and features applied to image processing, image analysis or image segmentation are countless even when restricting them to breast lesion segmentation. Therefore, a deep discussion on this topic is beyond the scope of this work, since any ML proposal is valid regardless of its particular advantages and disadvantages. However, it is our interest to analyze the nature of the training data used to build the stochastic models and is our goal since it conditions the nature of the overall segmentation.

When segmenting a target image using ML, two training strategies arise in order to build the stochastic models:

- use relevant information obtained from annotated images to drive the segmentation of the target image [79], [84].
- use information from the target image itself to drive the segmentation [76], [77].

Notice that in order to drive the segmentation from information from the target image itself, this information must be supplied by the user leading to an interactive procedure [65], [66]; or the information must be provided by another automatic procedure leading to an auto-guided procedure such as [76]. However, for detection application, only information from other images with accompanying GTs are used [60], [72], [73], since user interaction would already solve the detection problem. Taking this into account, figure 2.4 illustrates the 5 possible scenarios.

Database Trained Detection: generates statistic models from a training dataset to detect lesions in a target image using any sort of ML and features [60], [61], [72], [73], [76], [77], [84].

Image Trained Segmentation: from information supplied by the user, an ML procedure is trained from the target image in order to produce a segmentation [65], [66].

Database Trained Segmentation: the statistic models generated from the dataset are not used for localizing the lesion but rather to perform the segmentation itself. These methodologies produce image segmentation with no user interaction [57], [79]. In such a scenario, the features for constructing the models need to be robust to significative differences between the images.

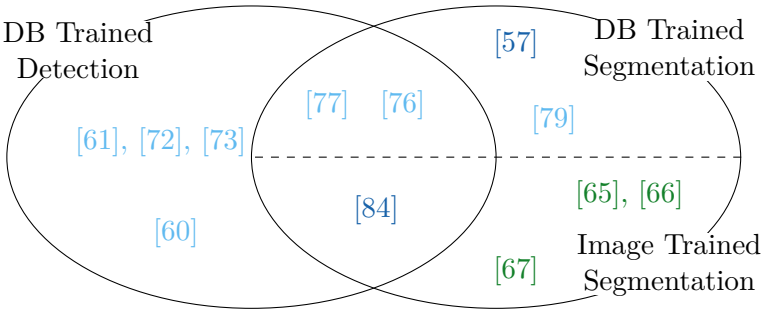


Figure 2.4: Supervised Machine Learning (ML) training and goals, ending up with a combination of 5 different strategies. The references are colored indicating the user interaction: semi-automatic (light-green), auto-guided(light-Blue), and fully automatic(dark-blue).

Database Trained Detection and Image Trained Segmentation: detection and segmentation are performed in a daisy chain manner like the models from a training dataset facilitate the detection of lesions within a target image. Once the suspicious areas are detected, they are used to train another ML procedure within the target image to drive the final segmentation. Although the errors in the detection step are propagated, this approach has the advantage that the statistical model driving the final segmentation has been specially built for every target image. The main drawback is that building this statistical model involves a training stage which is computationally very expensive [76], [77].

Integrated Methodology: trying to take advantage of the detection without building a specific model for the target image. Since there is no need to make the final detection decision whether there is a lesion or not, the posterior probability of the decision process can be used as another feature like a filter response of the image and integrated with the ML procedure [84].

2.2.3 Others

Here are listed other methods or parts of methods that are neither explicitly ACM nor ML procedures, nor are they basic image processing or image analysis techniques such as thresholding or region growing. In this sense, three main groups can be identified:

- Gaussian Constraining Segmentation (GCS) based methods
- unsupervised learning and over segmentation
- disk expansion for image inpainting

Methods using GCS for segmenting breast lesions in US data [58], [61], [64], [71] are inspired by the work of Kupinski et al. [112] which was initially adapted to US data by Horsch et al.[113]. They are based on constraining a multivariate Gaussian function with an image dependent function so that, when the resulting function is thresholded, a possible delineation is generated. Although these methodologies are not posted in the ACM form, they are equivalent to a fast marching LevelSet procedure [114]. Thresholding can be seen as a contour propagation, while the Gaussian constraining forces the direction of the propagation to be constant.

Some methods split the image or over-segment them for further operations like contour initialization [80], [81] or higher level features extraction from a coherent area so that it can be used in ML procedures [67], [84]. In order to carry out such an operation from a ML point of view, several unsupervised learning techniques have to be used in order to group the pixels: fuzzy C-means, K-means [69], and robust graph based clustering [81]. From an image analysis point of view, the grouping of similar contiguous pixels is equivalent to performing an over-segmentation of the image. Watershed transform [59], [67], [80] and Normalized Cuts (NC) [82], [84], [100] are popular techniques used to obtain an over-segmentation, also known as super pixels [115].

Finally, Yeh et al. [86] proposed a totally different approach for breast lesion segmentation based on inpainting of degraded typology. The image is transformed into a binary image using local thresholding and then the largest object within the binary image is reconstructed as the final segmentation.

2.2.4 Features

Intensity remains the most used feature within the methods analyzed. A feasible explanation might be found in the difficulty of incorporating other features rather than intensity or its gradient in the ACM procedures. A way to incorporate features other than intensity, such as texture, within the process is proposed by Aleman-Flores et al. [68]. The segmentation is carried out as two ACMs connected in a daisy chain manner. The second ACM evolves through the target image, whereas the first ACM used to obtain a

preliminary segmentation evolves using a generated image encoding the texture. This image is obtained by processing the target image using a modified anisotropic smoothing driven by texture features. The ACM evolves towards the gradient of this generated image already encoding texture information.

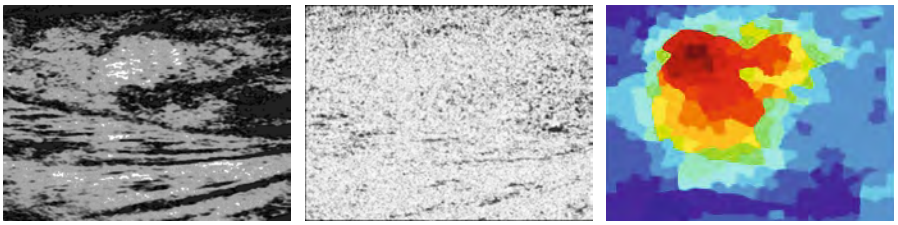
Texture descriptors have been more widely explored for methodologies incorporating ML since these methodologies naturally deal with multiple features. However, texture description is highly dependent on the scale of the features and seeing speckle as image texture is arguable since speckle is an unwanted effect that depends on the characteristics of the screening tissue, the acquisition device and its configuration [9]. However, images does look like a combination of texture granularities depending on the tissue which has encouraged the exploration of texture descriptors [60], [61], [72], [73], [80], [84], [116]. However, the use of a naive descriptor, like the one used in [60], [61], [72], cannot represent the large variability in texture present throughout the images. This can be qualitatively observed by comparing the MAP of the intensity and texture features, as shown in figure 2.5, where the latent information contained in the texture (fig. 2.5b) is less than that contained in the intensity feature (fig. 2.5a). A solution to cope with such texture variability consists of exploring multiple texture descriptors at multiple scales at the expense of handling larger feature sets resulting in a higher computation complexity and data sparsity that need to be handled.

On the other hand, texture can be seen as a filter response, so it performs the posterior of a classification process. Therefore, more sophisticated textures can be seen as the outcome of an ML process. Hao et al. [84] propose synthesizing texture from a lesion detection process (DPM) that takes advantage of Histogram of Gradients (HOG) taken at different scales. Figure 2.5c illustrates the feature plane inferred from the DPM process.

2.3 Segmentation assessment

Comparing all the methodologies reviewed in section 2.1 is rather cumbersome. The lack of a common framework for assessing the methodologies remains unaddressed, especially due to the absence of a public image dataset despite its being highly demanded by the scientific community [62], [102], [108]. However, the lack of a common dataset is not the only aspect complicating the comparisons. Here is a list of some of the feasible aspects complicating direct comparison of the works reviewed.

- Uncommon database



(a) (b) (c)

Figure 2.5: Qualitative assessment of feature planes: (a) Maximum A Posteriori (MAP) of intensity feature, (b) MAP of texture feature used in [60], [61] and (c) quantized DPM feature [84](image taken from the original work in [84]).

- Uncommon assessing of criteria and metrics
- Different degrees of user interaction
- Inability to quantify the user effort when interacting with a method
- Correctness of the GT used when assessing
- Uncommon treatment of missegmentation due to improper detection

The dificulty of comparing the methodologies using distinct datasets, distinct assessing criteria and distinct metrics is clear. Section 2.3.1 analyzes the criteria and metrics used to analyze the different methodology proposals. In order to conduct a discussion comparing the methodologies in section 2.4, when enough information is available, the reported results are set to a common framework for comparison purposes despite being assessed with different datasets. The assessment regarding user interaction is not further analyzed other than the already described interactive and automatic classification along with their respective subcategories (see section 2.1 and fig. 2.2). The correctness of the GT for assessing the segmentations refers to the huge variability of the delineations found when analyzing intra expert and inter expert variability on the segmentations [66]. In this regard, later in this chapter (see section: 2.3.2), a short discussion about the work that took intra and inter-observer delineation variability into account for assessing segmentation proposals can be found. Finally, the frontier between segmentation errors and errors due to the detection process is unclear and a proper criteria is not set. Massich et al. [61] take all the segmentations into account even if the segmentation has been wrongly initialized by the automatic detection procedure. Meanwhile, Zhang et al. [76] only use 90% of the

best segmentations to perform the segmentation assessment, arguing that the remaining segmentations suffered poor detection and that segmentation result assessment should not be subject to wrong initializations.

The rest of this section describes different area and boundary metrics collected from the works cited above, comments on the correctness of the assessing GT, based on intra- and inter-observer GT, variability and discusses the results reported.

2.3.1 Evaluation criteria

Although multiple criteria arise when assessing segmentations, this criteria can be grouped into two families depending on whether they are area or distance based metrics as illustrated in figure 2.6. Area based metrics assess the amount of area shared (Area Overlap (AOV)) between the obtained segmentation and the reference. On the other hand, distance based metrics quantify the displacement or deformation between the obtained and the desired delineations.

For the sake of simplicity, the name of the reported similarity indexes has been unified.

Area based segmentation assessment metrics

When analyzing the areas described by the segmented region to be assessed, A and the manually delineated reference region M (see fig. 2.6b), 4 areas become evident: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN); corresponding to the regions of the confusion matrix in figure 2.6a.

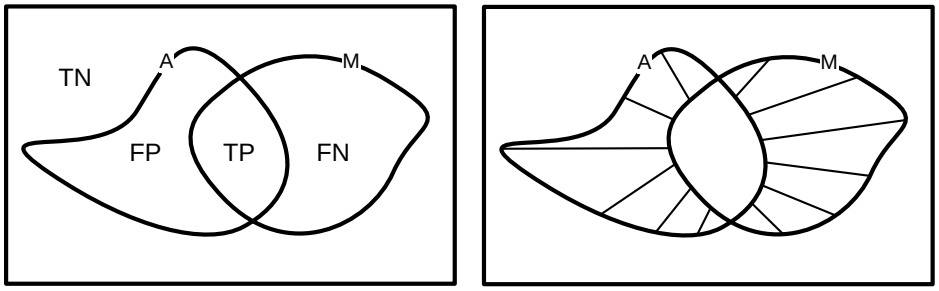
True Positive (TP) is found as the area in common ($A \wedge M$) between the two delineations A, M. The TP area corresponds to the correctly segmented areas belonging to the lesion.

True Negative (TN) is found as the area ($\overline{A} \wedge \overline{M}$) not belonging to either of the delineations A nor M . The TN area corresponds to the correctly segmented areas belonging to the background of the image.

False Positive (FP) is found as the area ($A \wedge \overline{M}$) belonging to the assessing segmentation A and not as a part of the reference delineation M. FP corresponds to the area wrongly labeled as a lesion since this area does not belong to the reference delineation.

		Segmentation Ground Truth (GT) (reference)	
		Positive	Negative
Segmentation Outcome (prediction)	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

(a)



(b) (c)

Figure 2.6: Methodology evaluation. (a) Statistical hypothesis test errors confusion matrix. (b) Graphic representation of the statistical hypothesis test errors for assessing the performance in terms of area. (c) Graphical representation of the boundary distance performance measures.

False Negative (FN) is found as the area $(\overline{A} \wedge M)$ corresponding to the reference delineation M but not as a part of the assessing segmentation A. FN corresponds to the areas of the true segmentation that have been missed by the segmentation under assessment.

Area metrics (or indexes) for assessing the segmentation are defined as a dimensionless quotient relating the 4 regions (TP, FP, FN and TN) described by the segmentation outcome being assessed (denoted A in fig:2.6a) and the reference GT segmentation (denoted M). Most of the indexes are defined within the interval $[0, 1]$ and some works report their results as a percentage.

Area Overlap (AOV), also known as overlap ratio, the Jaccard Similarity Coefficient (JSC) [70] or Similarity Index (SI) [79][4], is a common similarity index representing the percentage or amount of area common to the assessed delineation A and the reference delineation M according to equation 2.1. The AOV metric has been used to assess the following works: [58], [61], [64], [68], [69], [79], [84], [85]

$$AOV = \frac{TP}{TP + FP + FN} = \frac{|A \wedge M|}{|A \vee M|} \qquad \in [0, 1] \tag{2.1}$$

Dice Similarity Coefficient (DSC), also found under the name of SI [75], [80][5], is another widely used overlap metric similar to AOV. The difference between DSC and AOV is that DSC takes into account the TP area twice, one for each delineation. The DSC index is given by equation 2.2 and the relation between AOV or JSC and the DSC similarity indexes is expressed by equation 2.3. Notice that the DSC similiarity index is expected to be greater than the AOV index [66]. The DSC metric has been used to assess the following works:[66], [75], [76], [80]

$$DSC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = \frac{2|A \wedge M|}{|A| + |M|} \qquad \in [0, 1] \tag{2.2}$$

$$DSC = \frac{2 \cdot AOV}{1 + AOV} \tag{2.3}$$

[4]Notice that Similarity Index (SI) is also used formulated as the Dice Similarity Coefficient (DSC) in [75], [80] which differs from the SI definition in [79].

[5]Notice that Similarity Index (SI) is also used formulated as the Area Overlap (AOV) in [79] which differs from the SI definition in [75], [80].

True-Positive Ratio (TPR), also known as the recall rate, sensitivity (at pixel level) [66], [77] or Overlap Fraction (OF) [75], quantifies the amount of properly labeled pixels as lesion with respect to the amount of lesion pixels from the reference delineation (eq: 2.4). Notice that like the DSC, this value always remains greater than AOV (or equal when the delineations are identical). The TPR metric has been used to assess the following works: [60], [75], [77], [79]–[81], [85], [86]

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{|M|} = \frac{|A \wedge M|}{|M|} \qquad \in [0, 1] \tag{2.4}$$

Positive Predictive Value (PPV) corresponds to the probability that the pixel is properly labeled when restricted to those with positive test. It differentiates from TPR since here the TP area is regularized by the assessing delineation and not the reference, as can be seen in equation 2.5. PPV is also greater than AOV. The PPV metric is also used to assess the work in [66].

$$PPV = \frac{TP}{FP + TP} = \frac{TP}{|A|} = \frac{|A \wedge M|}{|A|} \qquad \in [0, 1] \tag{2.5}$$

Normalized Residual Value (NRV), also found as the Precision Ratio(PR) [59], corresponds to the area of disagreement between the two delineations regularized by the size of the reference delineation, as described in equation: 2.6. Notice that the NRV coefficient differs from $1 - \text{AOV}$ since it is regularized by the reference delineation and not the size of the union of both delineations. The NRV metric has been used to assess the following works: [59], [64], [82].

$$NRV = \frac{|A \oplus M|}{|M|} \qquad \in \left[0, 1 + \frac{A}{|M|}\right] \tag{2.6}$$

False-Positive Ratio' (FPR'), as reported in the presented work, is the amount of pixels wrongly labeled as lesion with respect to the area of the lesion reference, as expressed in equation 2.7. The FPR' metric has been used to assess the following works:[60], [79], [81], [85], [86] The FPR' has also been found in its complementary form $1 - TPR$ under the name of Match Rate (MR) [59].

$$FPR' = \frac{FP}{TP + FN} = \frac{FP}{|M|} = \frac{|A \vee M - M|}{|M|} \qquad \in \left[0, \frac{A}{|M|}\right] \tag{2.7}$$

Notice that the FPR' calculated in equation 2.7 differs from the classic False-Positive Ratio (FPR) obtained from the table in figure 2.6a, which corresponds to the ratio between FP and its column marginal ($FP + TN$), as indicated in equation 2.8. The FPR, when calculated according to equation 2.8, corresponds to the complement of specificity (described below).

$$FPR = \frac{FP}{FP + TN} = 1 - SPC \qquad \in [0, 1] \tag{2.8}$$

False-Negative Ratio (FNR) corresponds to the amount of pixels belonging to the reference delineation that are wrongly labeled as background, as expressed in equation 2.9. Notice that it also corresponds to the complement of the TPR since $TP \cup FN = M$. The FNR metric has been used to assess the following works: [60], [81], [86]

$$FNR = \frac{FN}{|M|} = \frac{|A \vee M - A|}{|M|} = 1 - TPR \qquad \in [0, 1] \tag{2.9}$$

Specificity corresponds to the amount of background correctly labeled. Specificity is described in equation 2.10 and is usually given as complementary information on the sensitivity (TPR). Specificity corresponds to the complementary of the FPR when calculated according to equation 2.8. The specificity index is also used to assess the work in [66], [77].

$$SPC = \frac{TN}{TN + FP} = \frac{|\overline{A} \wedge \overline{M}|}{|\overline{M}|} = 1 - FPR \qquad \in [0, 1] \tag{2.10}$$

Boundary based segmentation assessment metrics

Although the boundary assessment of the segmentations is less common than area assessment, it is present in the following works: [60], [64], [68], [70], [76], [79], [81]. Like when assessing the segmentations in terms of area, the criteria for assessing disagreement between outlines are also heterogeneous which makes the comparison between works difficult. Unlike the area indexes, with the exception of the further introduced Average Radial Error (ARE) coefficient, which is also a dimensionless quotient, the rest of

the boundary indexes or metrics are physical quantitative error measures and are assumed to be reported in pixels. Although some of the reported measures are normalized, they are not bounded by any means.

Zhang et al. [76] propose using average contour-to-contour distance (E_{cc}) for assessing their work. However, no definition or reference is found on it. Huang et al. [81] propose using ARE, defined in equation 2.11, where a set of n radial rays are generated from the center of the reference delineation C_0 intersecting both delineations. The ARE index consists of averaging the ratio between the distance of the two outlines $|C_s(i)-C_r(i)|$ and the distance between the reference outline and its center $|C_r(i) - C_0|$.

$$ARE = \frac{1}{n}\sum_{i=1}^{n}\frac{|C_s(i) - C_r(i)|}{|C_r(i) - C_0|} \tag{2.11}$$

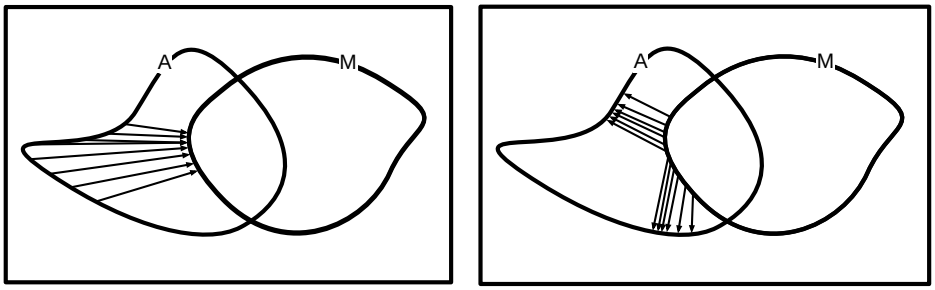
The rest of the works base their similitude indexes on the analysis of the Minimum Distance (MD) coefficients. The MD is defined in equation 2.12 and corresponds to the minimum distance between a particular point a_i within the contour A (so that $a_i \in A$) and any other point within the delineation M.

$$\mathrm{MD}(a_i, M) = \min_{m_j \in M} \|a_i - m_j\| \tag{2.12}$$

Hausdorff Distance (HD), or Hausdorff error, measures the worst possible discrepancy between the two delineations A and M as defined in 2.13. Notice that it is calculated as the maximum of the worst discrepancy between (A, M) and (M, A) since MD is not a symmetric measure, as can be observed in figure 2.7. The HD as defined in equation 2.13 has been used for assessing the segmentation results in Gao et al. [70]. Meanwhile, Madabhushi and Metaxas [60] and Shan et al. [79] only take into account the discrepancy between the assessed delineation A with reference delineation M, here denoted as HD' (see eq. 2.14). In [60], [79], the HD' is also reported in a normalized form $\frac{\mathrm{HD}'}{\eta}$, where η is the length of the contour of reference M.

$$\mathrm{HD}(A, M) = \max\left\{\max_{a_i \in A}\mathrm{MD}(a_i, M), \max_{m_i \in M}\mathrm{MD}(m_i, A)\right\} \tag{2.13}$$

$$\mathrm{HD'}(A, M) = \max_{a_i \in A}\mathrm{MD}(a_i, M) \tag{2.14}$$



(a) (b)

Figure 2.7: Illustration of the non-symmetry property of the Minimum Distance (MD) metric. (a) $\mathrm{MD}(a_i, M)$, (b) $\mathrm{MD}(m_i, A)$

Average Minimum Euclidian Distance (AMED), defined in equation 2.15, is the average MD between the two outlines. [70]. Similar to the case of the HD' distance, Madabhushi and Metaxas [60] and Shan et al. [79] only take into account the discrepancy between the assessed delineation A with reference to the delineation M to calculate the AMED' index (see eq. 2.16). The AMED index can be found under the name of Mean Error (ME) in [60] and Mean absolute Distance (MD) in. [79].

$$\mathrm{AMED}(A, M) = \frac{1}{2} \cdot \left[\frac{\sum_{a_i \in A} \mathrm{MD}(a_i, M)}{|A|} + \frac{\sum_{m_i \in M} \mathrm{MD}(m_i, A)}{|M|} \right] \quad (2.15)$$

$$\mathrm{AMED'}(A, M) = \frac{\sum_{a_i \in A} \mathrm{MD}(a_i, M)}{|A|} \quad (2.16)$$

Proportional Distance (PD), used in [64], [68], takes into account the AMED regularized with the area of the reference delineation according to equation 2.17

$$\mathrm{PD}(A, M) = \frac{1}{2\sqrt{\frac{Area(M)}{\pi}}} \cdot \left[\frac{\sum_{a_i \in A} \mathrm{MD}(a_i, M)}{|A|} + \frac{\sum_{m_i \in M} \mathrm{MD}(m_i, A)}{|M|} \right] * 100 \quad (2.17)$$

2.3.2 Multiple grader delineations (Study of inter- and intra-observer segmentation variability)

Assessing the true performance of a medical imaging segmentation procedure is, at least, difficult. Although method comparison can be achieved by assessing the methodologies with a common dataset and metric, true conclusions about the performance of the segmentation are questionable. Assessing segmentations of medical images is challenging because of the difficulty of obtaining or estimating a known true segmentation for clinical data. Although physical and digital phantoms can be constructed so that reliable GT are known, such phantoms do not fully reflect clinical imaging data. An attractive alternative is to compare the segmentations to a collection of segmentations generated by expert raters.

Pons et al. [66] analyzed the inter- and intra-observer variability of manual segmentations of breast lesions in US images. In the experiment, a subset of 50 images is segmented by an expert radiologist and 5 expert biomedical engineers with deep knowledge of a breast lesion appearance in US data. The experiment reported an AOV rate between 0.8 and 0.852 for the 6 actors. This demonstrates the large variability between GT delineations; a fact that needs to be taken into account in order to draw proper conclusions about the performance of a segmentation methodology. However, having multiple GT delineations to better assess the segmentations performance is not always possible. When possible, several strategies have been used to incorporate such information.

Cui et al. [69] tested the segmentation outcome against 488 images with two delineations provided by two different radiologists. The dataset is treated as two different datasets and the performance on both is reported. Yeh et al. [86] used a reduced dataset of 6 images with 10 different delineations accompanying each image. The performance for each image was studied in terms of reward average and variation of the 10 reference delineations. Aleman-Flores et al. [68], where a dataset of 32 image dataset with 4 GT delineations provided by 2 radiologists (2 each) was available, assessed the segmentation method as if there were 128 (32×4) images.

A more elaborate idea to estimate the underlying true GT is proposed by Massich et al. [61] and Pons et al. [66]. Both works propose the use of STAPLE in order to determine the underlying GT from the multiple expert delineations. STAPLE states that the ground truth and performance levels of the experts can be estimated by formulating the scenario as a missing-data problem, which can be subsequently solved using an EM algorithm. The EM algorithm, after convergence, provides the Hidden Ground Truth (HGT) es-

timation that has been inferred from the segmentations provided by the experts as a probability map. Massich et al. [61] propose to assess the segmentation against a thresholded HGT and weight the AOV index with the HGT. The authors in [61] argued that apart from comparing the segmentation resulting from binarizing the greaders segmentation agreement, the amount of agreement the needs to be taken into account. This way, properly classifying a pixel with large variability within the graders produces less reward and miss classifying a pixel with great consensus penalizes.

2.4 Discussion

As has been said all along in section 2.3, accurate comparison of the segmentation methodologies from their proposal works is not feasible. The major inconveniencies are uncommon assessing datasets and inhomogeneous assessing criteria, but the fact that all the indexes for assessing segmentations seen in section 2.3 are made at the image level can also be added. Therefore, the statistics used for reporting the performance of segmentation methodologies at the dataset level might vary as well. Most of the works report their dataset performance as an average of the image assessment reward. Some works complement such information with minimal and maximal value [64], the standard deviation [68], [69], [76], [81], [84], [85], or median [68], [84]. Some other works prefer to report the distribution of their results graphically [61], [70], [86]. Finally, in [75], [79], it is not specified which statistic has been used, although mean is assumed.

Despite all the mentioned inconveniences, information regarding performance of all the works presented here is gathered in table 2.1 and graphically displayed in figure 2.8 in order to analyze some trends. In table 2.1, the works presented are grouped depending on the user interaction according to the 4 categories described in section 2.1: interactive segmentation (fully-guided and semi-automatic) and automatic segmentation (auto-guided and fully-automatic). For each method the size of the dataset, the number of different GT delineations per image used to assess the methodology and the results in the original work are reported. If the assessment index is found under another name rather than the name used in section 2.3, the name used here as a reference appears in brackets to homogenize the nomenclature in order to facilitate comparison. Finally, when enough information is available, an inferred AOV value, also to facilitate comparing the works is shown in the last column of the table.

Figure 2.8 displays only those methods where AOV was available or

could be inferred from the reported data. These representations synthesize the methods' performance and the datasets used for the assessment in a single view. The different works are radially placed according to different criteria and the references are colored in terms of the user interaction categories defined in section 2.1.The AOV appears in blue in percentage as well as graphically within a score circle. In this score circle, there is also presented the intra- and inter-observer variability segmentation results reported in [66] as a blue colored swatch within two dashed circles that represent the minimum and the maximum disagreement reported in the experiment. The size of the dataset used for assessing the segmentation performance appears in red. In the center of the radial illustration, a 3 class categorization of the size of the dataset has been carried out. The 3 classes correspond to small (less than 50 images), medium (between 50 and 250 images) and large (more than 250 images).

Figure 2.8a arranges the works presented according to the categories shown in figure 2.3; ACM, ML, others, and their combination. This representation in sectors facilitates ascribing the importance of a particular segmentation type at a glance, since combinations of these are placed contiguous to the unaccompanied type. For readability purposes, methodologies combining aspects of these three categories ([60], [69]) have been chosen to belong to the combination of the two categories best describing the method. So, Madabhushi and Metaxas [60] is treated as a combination of ML and ACM, and Cui et al. [69] as an ACM and other methodology combinations. Figure 2.8b arranges the presented works according to the user interaction. Figure 2.8c only takes into account the presented works that make use of ML and are arranged according to the criteria exposed in section 2.2.2 (see fig:2.4) plus the unsupervised methods. Finally, Figure 2.8d represents the methodologies belonging to the ACM class, arranged by type (see fig:2.3 and section 2.2.1).

When analyzing the figures, an already stated observation arises while comparing the methodologies against the swatch representing the inter- and intra-observer variability: some works surpass the performance of trained human observers. A feasible explanation is that the complexity of the datasets used for assessing the methodologies and the dataset used for assessing the observers variability differ. This would also explain the unfavorable results of the methodology proposed by Xio et al. [65] when quantitatively assessed in [66], using the same dataset used for assessing the inter- and intra-observer variability. This observation corroborates the need of a public dataset of breast US images with annotated information.

Despite the fact that any conclusion will be biased due to uncommon

assessing datasets, some observations can still be made. Although ACM methodologies have been tested mostly in rather small datasets, a trend to achieve better results when using ACM methodologies can be seen in figure 2.8a and corroborated when comparing the areas of the plots in figures 2.8b and 2.8c. This shows that the combining image information with structural regularizing forces produce accurate results. Although more methodologies implementing similar technologies are needed to draw proper conclusions, a tendency to obtain lower results when using the Snakes ACM formulation can be seen in figure 2.8d. Such a tendency is explained by the influence that initialization has when using Snakes.

The segmentation performance reported for methodologies based on ML varies from the most unsatisfactory results to results comparable to human performance, as can be seen in figure 2.8. This figure also indicates that these methodologies have been tested mainly in large datasets. Of the methods within this category, the methodology proposed by Xio et al. [65] reports the most unsatisfactory results. Despite the difficulties due to a challenging dataset aside, other reflections can be done based on the reported results and the nature of the methodology. Such a bad performance is surprising from the point of view of the classification, since the proposed ML procedure is trained using information supplied by a user from the same target image. In it, a combination of EM and MRF procedures fit two model lesion/non-lesion extracted from several ROIs specified by the user in order to perform the segmentation. The results obtained indicate that there is a strong overlapping in appearance between lesions and non lesion areas in the image, which for the application of breast screening in US images is true. This indicates that more elaborate features than intensity at pixel level are needed. This hypothesis is supported by the results obtained in [76], [79] where more elaborate features are used, producing results which are within the range of a human observer.

Methodologies categorized as other methodologies perform within the range of the state-of-the-art. As an observation, Gomez et al. [64] proposed a methodology based on the popular GCS [58], which has been reported to obtain the best results within the other methodologies category achieving an AOV of 85.0%. On the other hand, Massich et al. [61] proposed a methodology also based on GCS reporting the most unsatisfactory results (64.0%) but with the advantage of allowing less user interaction.

Notice that similar to the fact of using an uncommon image dataset, distinct consideration of the detection errors also bias the comparison. For instance, the AOV of 84.0% reported in [76] is obtained once the worst 10% of the segmentations are discarded arguing that such bad results are not

due to the segmentation procedure but due to a wrong detection instead. In contrast, the lower results reported by Madabhushi and Metaxas [60] when comparing them to the rest of the methodologies using ACM can be explained due to wrong initialization of the ACM step.

Despite the bias subject to analyze the segmentation performance of the reviewed methodologies from the results compiled in table 2.1, some of the general trends observed are summarized here. Methodologies using ACM reported good results, although they have been tested mainly in small datasets. Moreover, when using ACM methodologies, the correctness of the results are subject to the initialization of the ACM step with the exception of the LevelSet proposal in [85], since the proposed LevelSet implementation allows a naive initialization. Methodologies using ML have been tested mainly on larger datasets. Methodologies using more sophisticated features produce results comparable to those achieved when using ACM.

Table 2.1: Performance reported with the works presented. In the table, the overall size of dataset used for testing, the number of delineations per image, the results reported and, when possible, the inferred Area Overlap (AOV) coefficient can be found.

work	DB size	GT	Reported Metric	AOV
[63]	20	1	∼	∼
[58]	400	1	AOV 0.73	73.0%
[64]	50	1	AOV 85%, NRV 16%, PD 6.5%	85.0%
[65], [66]	352	6	Sensitivity(TPR) 0.56, Specificity 0.99, PPV 0.73, AOV 0.51, DSC 0.61	50.8%
[66]	352	6	Sensitivity(TPR) 0.61, Specificity 0.99, PPV 0.80, AOV 0.55, DSC 0.66	54.9%
[67]	16	1	∼	∼
[68]	32	4	AOV 0.88, PD 6.86%	88.3%
[69]	488	2	AOV 0.73±0.14 AOV 0.74±0.14	74.5%
[70]	20	1	TPR>0.91, FPR 0.04, JSC(AOV) 0.86, DSC 0.93, AMED 2pix., HD=7pix.	86.3%
[71]	757	1	Results reported as detection	∼
[61]	25	7	AOV 0.64	64.0%
[60]	42	1	FPR 0.20, FNR 0.25, TPR 0.75 ME(AMED') 6.6pix.	62.0%
[75]	118		SI(DSC) 0.88 OF(TPR) 0.86	77.6%
[76]	347		AOV 0.84±0.1, ECC 3.75±2.85pix.	84.0%
[77]	112	1	∼	∼
[79]	120	1	TPR 0.92, FPR 0.12, SI(AOV) 0.83, HD' 22.3pix., MD(AMED') 6pix. (when using SVM classifier) TPR 0.93, FPR 0.12, SI(AOV) 0.83, HD' 22.3pix., MD(AMED') 6pix. (when using ANN classifier)	83.0% 83.1%
[80]	20		SI(DSC) 0.88, OF(TRP) 0.81	78.6%
[81]	20	1	TPR 0.87, FP 0.03, FN 0.13, ARE 9.2% (benign) TPR 0.88, FP 0.02, FN 0.13, ARE 9.2% (malignant)	85.2%
[82]	40	1	NRV 0.96 (benign); NRV 0.92 (malignant)	∼
[84]	480	1	JSC(AOV) 0.75±0.17	75%
[59]	60	1	PR(NRV) 0.82, MR(FPR) 0.95	∼
[57]	112		Diagnosis results reported only	∼
[85]	76	1	TPR 0.94, FPR 0.07, AOV 0.88	88.1%
[86]	6	10	TPR>0.85, FNR<0.15, FP<0.16	73.3%

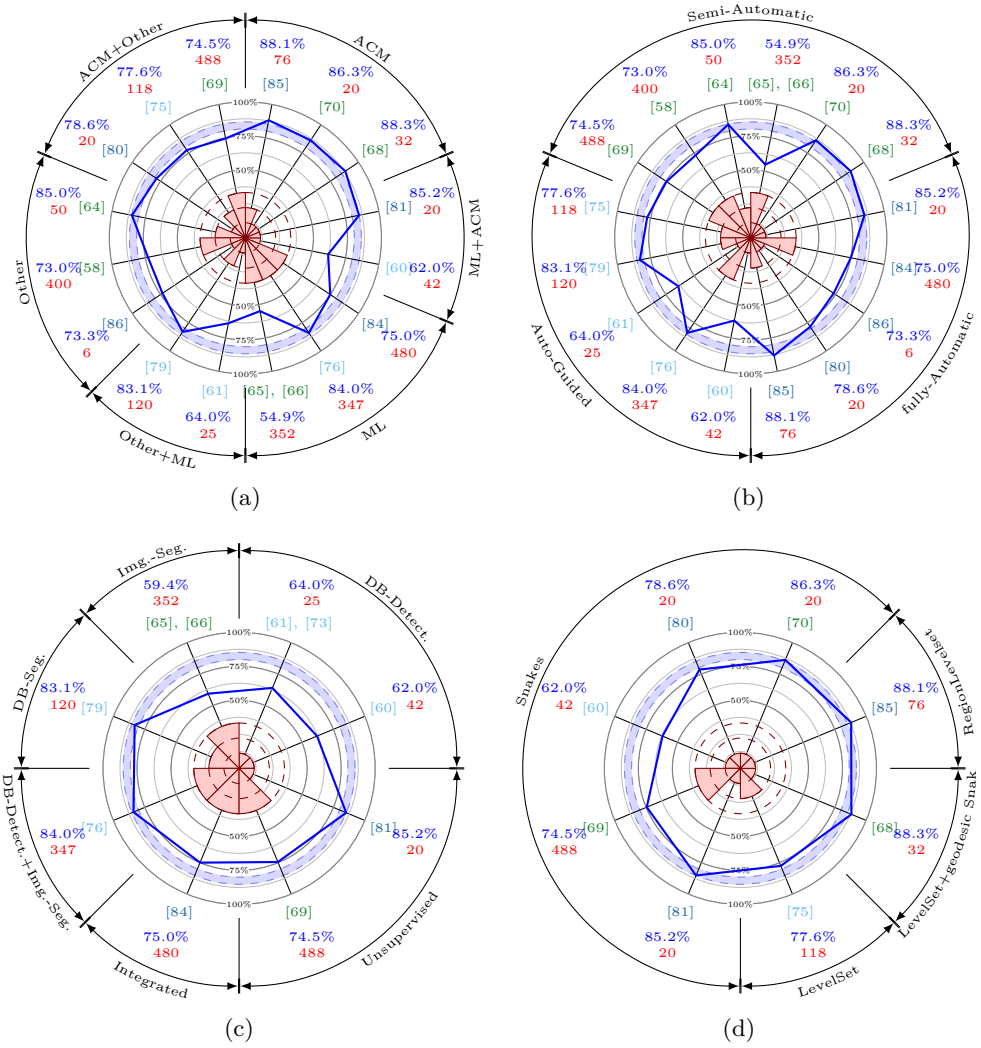


Figure 2.8: Graphical comparison of the methods presented that reported Area Overlap (AOV) or enough data to be inferred. The inner part of the plot illustrates the size of the dataset used in terms of small, medium, large. The blue swatch illustrates the inter- and intra-observer experiment results carried out in [66]. The coloring of the reference indicates the user interacthability: semi-automatic (light-green), auto-guided(light-Blue), and fully automatic(dark-blue).

Chapter 3

Objective Function Optimization Framework for Breast Lesion Segmentation

> Reality is the murder of a beautiful theory by a gang of ugly facts.
>
> R. Glass

3.1 Introduction

Despite the inherent bias in the analysis carried out in section 2.4, good results are reported when using optimization procedures. Optimization methods offer a standardized manner to approach segmentation by minimizing an application-driven cost function [103]. Obviously the segmentation results are subject to the correctness of the cost function design. Although this cost function can be adapted from one optimization scheme to another, some particularities of every optimization framework need to be taken into account since different facilities are offered when modeling depending on the chosen framework. These optimization methodologies include ACM methodologies and ML procedures solving the *metric labeling* problem [104] such as [76], [84].

Table 3.1 summarizes some of the characteristics of the optimization methodology families. Despite the possibility of finding particular instances and implementations contradicting this summary (table 3.1) due to the vast

Table 3.1: Optimization methods characteristics

		ML+MRF	ACM		
			Snakes	LevelSets	ACWE
Spatially	continuous		✓	✓	✓
	discrete	✓			
Contour control or modeling			✓	✓	✓
Data model		✓			
Need of initialization			✓	✓	
Topology changes		✓		✓	✓

and extensive work carried out by the scientific community in this field, the table pretends only to illustrate and highlight some differences in order to present a short discussion of these families.

Both ACM and the combination of ML and MRF address the same problem in different manners: the former using a spatially continuous formulation whereas the latter uses a discrete formulation. ACM offers full control of the forces pulling the contour. These forces deform the contour based on the image information and forces constraining the contour interact to find an equilibrium state. Contour constraining forces allow us to impose shape or smoothness on the delineation. This is not the case for the ML and MRF combination where no restriction of the contour can be easily made. In contrast, it offers great facilities to fit complex models due to the use of ML where high level features can be used to drive the segmentation. The possibility of fitting high level models when using ACM procedures is limited despite some attempts made in that sense. The Active Contour Without Edges (ACWE) [117], where a Levelset segmentation is driven by region information should be mentioned. In this regard, Liu et al. [85] propose to using, as region information, an error measure between the data inside the delineation and a Rayleigh model inferred from training data. Most ACM procedures require an accurate initialization. Differences in ACM formulation allows topology changes within the contour enabling the delineation of multiple objects, which is a desirable capability. However, its downside is that some false positive delineations might arise, as in the case of combining ML and MRF, where no restrictions of any kind are made on the segmentation topology.

This chapter is devoted to the presentation of a discrete optimization framework based on ML and MRF for segmenting breast lesions in US images. However, this chapter also reports some insights to our previous attempts to address the same goal but, in this case, using Gaussian Constraining Segmentation (GCS) instead of a discrete optimization framework. The similarities between GCS and ACM are quite extensive since GCS can be assimilated as a fast-marching procedure [114], a type of LevelSet where the direction of the contour propagation remains constant. Therefore, adding more details of our previous work is used here to illustrate some of the inconveniences reported in table 3.1 when using these methodologies applied to breast lesion segmentation in US images and how this limitations are commonly overcome using our work as illustration case.

3.2 GCS-based segmentation

When applying GCS or ACM to segment breast lesion US images with no user interaction, common strategies to overcome their limitations arise. The main inconveniences are:

- the need of an accurate initialization.
- the lack of a manipulable data model to introduce high level features for the contour to evolve.

The need of an accurate initialization is overcome mainly by generating a preliminary segmentation usually using ML procedures [60], [81]. The use of ML allows us to take advantage of high level features to localize the lesions but, most of the time, there is no need to introducing high level features within the contour evolution when the initialization is close enough to the solution. A practical way to introduce high level features into the evolution of the ACM consists of generating an image from those high level features and let the ACM evolve in the synthesized image [68].

In our GCS-based segmentation proposal, both strategies, generating a preliminary segmentation taking advantage of ML, and letting the segmentation evolve on a synthesized image, are used to perform the segmentation.

3.2.1 General outline of the GCS-based segmentation framework

Figure 3.1 shows the basic operations for the proposed GCS-based segmentation framework: after an initial region $R_0(x, y)$ is determined, it is con-

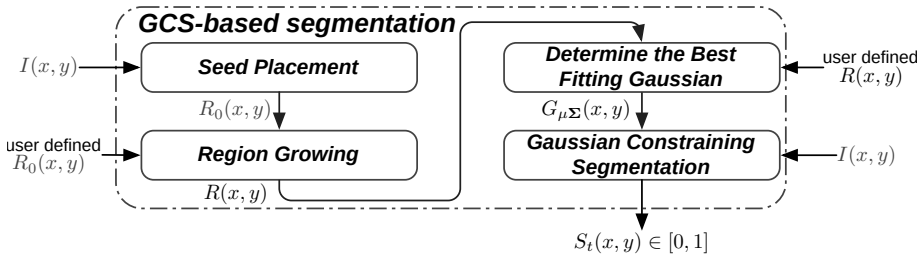


Figure 3.1: Block diagram for the Gaussian Constraining Segmentation (GCS) framework for segmenting breast lesions.

verted into a preliminary lesion delineation $R(x, y)$ by means of a region growing algorithm. This lesion delineation is used to obtain a multivariate Gaussian function describing the shape, position and orientation of the lesion ($G_{\mu\Sigma}(x, y)$). Finally, the Gaussian Constraining Segmentation (GCS) procedure refines the segmentation by thresholding an intensity dependent function $\Psi(x, y)$ constrained by the multivariate Gaussian function describing the lesion.

3.2.2 Seed Placement

In order to obtain an accurate initialization, either information from the user is supplied or ML procedures are used in order to infer this knowledge from annotated data. In this case, the adopted solution has been to use a basic ML to integrate the information in the same manner that the *Stavros Criteria* spots breast lesions. The *Stavros Criteria* state that intensity and texture exhibit high specificity [13] and in [118] this information can be found in combination with a tendency that radiologists have of placing the lesions in the center of the image using an ad-hoc heuristic.

Figure 3.2 outlines the working scheme of our proposal to combine the Intensity Texture and Geometric (ITG) constraints in a more generic manner complying with the Bayesian framework described in equation 3.1 in order to obtain a posterior or total probability plane. This probability plane is then thresholded and the largest area from the foreground is selected as the seed region $R_0(x, y)$. The threshold has been empirically set and kept constant for all the images at 0.8 as a good tradeoff between large foreground regions and low lesion recall.

$$P(Lesion|I,T) = \frac{P(I,T|Lesion) \cdot P(Lesion)}{P(I,T)} \tag{3.1}$$

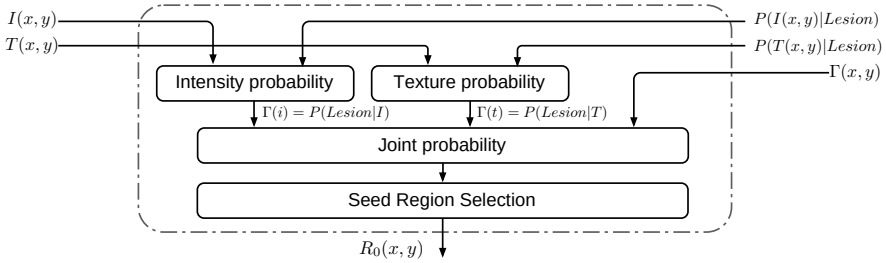


Figure 3.2: Block diagram for the Intensity Texture and Geometric (ITG) based seed placement proposal.

In equation 3.1, notice that the denominator $P(I,T)$, which is difficult to obtain, can be ignored since it is common to the two classes $\{Lesion, \overline{Lesion}\}$ and therefore it cancels out. The term referred to as the prior knowledge of a pixel being a lesion $P(Lesion)$ is assumed to be a centered multivariate Gaussian distribution proportional to the image. Figure 3.3 shows the spatial distribution of lesion pixels from an annotated dataset indicating the validity of assuming that the lesions are placed in a normal manner with respect to the image's center. $P(I,T|Lesion)$ corresponds to the multivariate distribution of pixels being a lesion based on intensity and texture features. The main disadvantage of this term is that the sparsity of the data leads to bad results. Therefore, IID needs to be assumed in order to be able to estimate the multivariate distribution from its marginals at the expense of some inaccuracy (see eq:3.2).

$$P(I,T|Lesion) \overset{iid}{=} P(I|Lesion) \cdot P(T|Lesion) \tag{3.2}$$

With all this in mind, the final posterior probability can be calculated accordingly to equation 3.3, where $P(I|Lesion)$ and $P(T|Lesion)$ are non-parametric estimations of the intensity and texture PDF determined from training data by performing a occurrence quantification, followed by Gaussian smoothing and posterior normalization to guarantee that the total accumulated probability remains equal to 1.

$$P(Lesion|I,T) = P(I|Lesion) \cdot P(T|Lesion) \cdot P(Lesion|x,y) \tag{3.3}$$

The texture measure used here is given by equation 3.4 which corresponds to the difference between the pixel intensity $I(x,y)$ and the mean intensity of its N nearest neighbors. For this implementation an eight pixel neighborhood is used.

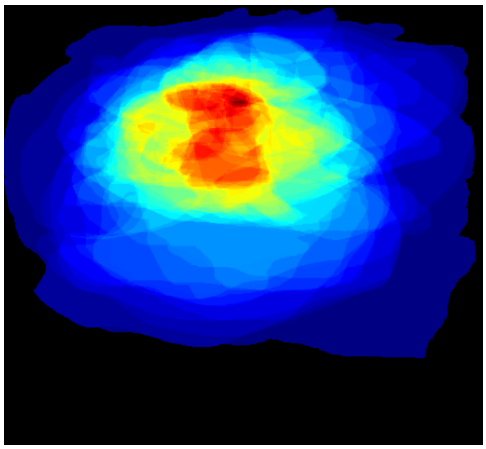


Figure 3.3: Lesion pixel occurrence in a normalized image $P(x, y|Lesion)$ obtained from an annotated dataset

$$T(x,y) = I(x,y) - \frac{1}{N}\sum_{\delta=0}^{N-1} I_\delta(x,y) \tag{3.4}$$

Wrapping up, the seed placement methodology proposed here to initialize the segmentation makes use of five inputs to automatically determine a seed region: the intensity image, the texture image, the intensity and texture Probability Density Functions, and the seed location prior along with a fixed parameter to split the probability plane into foreground and background.

3.2.3 Preliminary lesion delineation using region growing

The growth of the initial region $R_0(x, y)$ into the preliminary lesion delineation $R(x, y)$ is done in an iterative manner where at each iteration, the pixel candidates are tested in order to be aggregated to the next iteration region R_{i+1}.

To generate the candidate pixels, a morphological operation with a structural element consisting of a 3 pixels radius disc is performed. This expansion methodology is preferred over using only the pixels connected to the current region R_i to perform a larger exploration that cope with the noise nature of the images. Conversely, to compensate the loose policy of the exploration step, a restrictive criterion is needed for aggregating a pixel to next iteration's region. The criterion used to aggregate a position is based on the mean $\mu(\cdot)$ and the standard deviation $\sigma(\cdot)$ of the pixels' intensity

from the current region R_i, as described in equation 3.5. Where in order to aggregate a candidate position (x, y) to the next iteration region R_{i+1}, its intensity value must not be further than 0.5σ of the current region mean intensity.

$$I(x,y) \in R_{i+1} \iff I(x,y) \in \mu(I(R_i)) \pm \frac{\sigma(I(R_i))}{2} \tag{3.5}$$

The final region $R(x, y)$ is obtained by applying a dilatation using the same 3 pixels radius disc structural element.

3.2.4 Gaussian Constrain Segmentation (GCS)

In order to integrate image characteristics such as homogeneity, texture, etc., a synthesized image ($\Psi(x, y)$) taking into account all these characteristics is generated to drive the segmentation. This work-around, to incorporate high level features or image constraints, can also be observed when utilizing ACM (see. [68]).

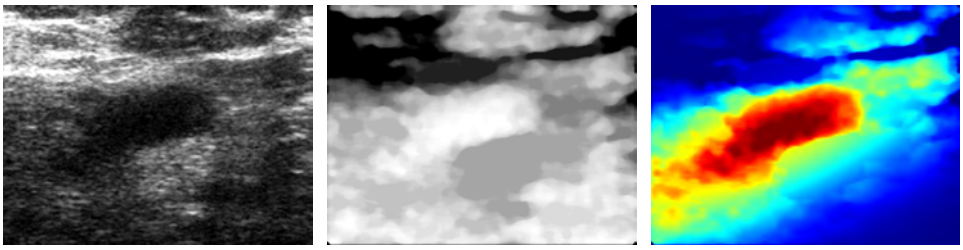
For this particular application synthesized image is obtained from the intensity image by applying in the following order: brightness inversion, three stages of median filtering with a $5x5$ kernel and a gray-scale morphological opening operation.

Figure 3.4 illustrates the image dependent function. Figure 3.4b, calculated from the original image (fig. 3.4a), and fig. 3.4c shows how this function reshapes the original multivariate Gaussian to fit the lesion. The function used in this case inverts the image intensity, performs multiple median filtering stages in order to obtain piecewise constant regions to preserve edges, and uses morphological operations to fill the holes to ensure a constant direction of the propagating contour.

In order to finally generate a segmentation, the function representing the lesion is finally thresholded, as illustrated by the GCS working scheme shown in figure 3.5. In our proposal, this thresholding can be determined in two different manners: by training a dataset tune up or dynamically determining the threshold best suited to each image. For an initial proposal, the space of possible thresholds was sampled and the threshold that minimizes the sum of the variance inside and outside the delineation is selected.

3.2.5 Qualitative results

In order to accompany the quantitative results reported in section 2.4, figure 3.6 illustrates some qualitative results. Here, several overlaid color lines



(a) (b) (c)

Figure 3.4: $\Psi(x, y)$ construction for Gaussian Constraining Segmentation (GCS) segmentation purposes. (a) original image, (b) image dependent function, (c) Gaussian constrained function.

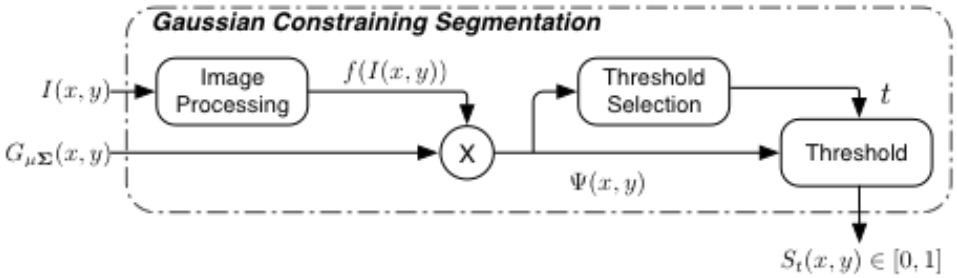
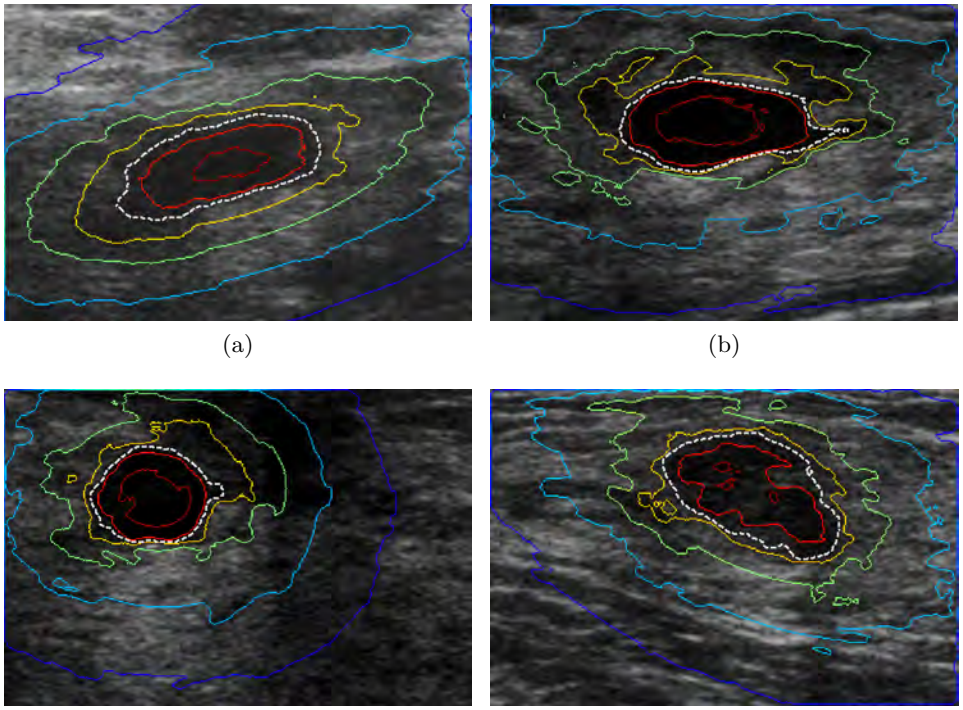


Figure 3.5: Gaussian Constraining Segmentation (GCS) outline.

illustrating possible thresholds at different levels in a color scheme in which-cold colors represent low value thresholds and warm colors represent high value thresholds. The white dashed line illustrates the final lesion delineation produced by the methodology when using the proper threshold. Notice that approximately all the segmentations are found at a similar threshold level. Therefore, the thresholding step can be substituted by a tune up threshold.

3.3 Optimization framework for segmenting breast lesions in Ultra-Sound data

The use of supervised ML has long been prevalent in lesion segmentation [119] to train a classifier using a database of training images with the ground truth provided, so that the segmentation of lesions in test images with no such ground truth may be predicted. During the training phase, features are extracted from some elements within the training images and provided along with information on their ground truth to the training model of the



(a) (b)

(c) (d)

Figure 3.6: Qualitative results achieved using Gaussian Constraining Segmentation (GCS) segmentation as proposed in [61] complementing the quantitative results from chapter 2. Colored lines represent possible thresholds at different levels and the white dashed line outlines the final segmentation.

classifier. This enables a relationship between the features and their class to form. The testing of an image is performed by extracting the same features from compatible elements within the target image and passing them to the trained classifier, which then provides a prediction of the class and/or a probability of the matching to a class.

Optimization frameworks solving the labeling metric problem adds spatial coherence to the class prediction obtained by the classifier according to equation 3.6. Spatial coherence is a standard regularization used in computer vision, imposing the condition that the majority of the signal components forming an image correspond to a low frequency, thus contiguous elements should be encouraged to have similar label.

$$U(\omega) = \sum_{s \in S} D_s(\omega_s) + \sum_{s} \sum_{r \in \mathcal{N}_s} V_{s,r}(\omega_s, \omega_r) \tag{3.6}$$

Just before analyzing equation 3.6, let some terms be defined. Let's define an image as a set of sites, $\mathcal{S}$, and let the goal of labeling be the assignation of a particular label from a defined set of possible labels $\mathcal{L}$ to every site $s \in \mathcal{S}$. Now, let ω be defined as a particular configuration (or labeling) within all the possible label configurations ($\omega \in W$). Thus, ω_s corresponds to the labeling of the current configuration ω for the particular site s, so that $\omega_s = l$, $l \in \mathcal{L}$.

In equation 3.6, $U(\omega)$ corresponds to the cost of a particular configuration ω and is defined as the combination of two independent cost functions. The former term, $D_s(\omega_s)$, is referred to as the *data* term, while the latter, $\sum_{r \in \mathcal{N}_s} V_{s,r}(\omega_s, \omega_r)$, is indistinctly referred to as the *pairwise* or *smoothing* term. The data term is the cost of assigning a particular label l (also denoted ω_s) to the site s based on the image data of s, whereas the pairwise or smoothing term represents the cost of the assignation ω_s taking into account the labels of its neighbor sites, ω_r, $r \in \mathcal{N}_s$.

Figure 3.7 corresponds to a toy example of a 4 site image. The total cost for every possible configuration ω along with the contribution of each of the terms, the data term as well as the smoothing term, is represented.

Notice that function $U(\cdot)$ is defined within the labeling space rather than in the image data. Therefore, the underlying true segmentation of the image must be estimated as $\hat{\omega} = \arg\min_{\omega} U(\omega)$. The image information shapes the functions $D(\cdot)$ and $V(\cdot, \cdot)$, corresponding to the data and the pairwise terms.

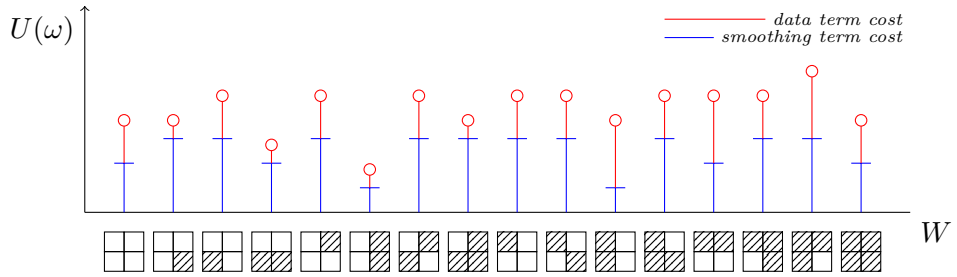


Figure 3.7: Toy example illustrating data and pairwise costs and how the overall minimal segmentation is selected.

3.3.1 System Outline

The basic structure of the algorithm developed to implement the optimization framework described in equation 3.6 is outlined in figure 3.8. Although further details of all the elements composing the figure can be found throughout this chapter, a brief description is given here in order to conduct an overview of the segmentation scheme proposed.

Each image is first converted into a set $\mathcal{S}$ of arbitrary elements denoted sites used here to represent images consisting of groups of contiguous pixels sharing some characteristic. These pixel groups are called superpixels. On one hand, a set of features used to characterize each superpixel are extracted from the images and then used to train a classifier if the superpixels are from the training set images, or for classification if they are from the images to be segmented. From this classification, the class reward of each superpixel is used to build up the data term $D_s(\cdot)$ of every site s.

On the other hand, relationships between the sites in $\mathcal{S}$ are established in order to impose spatial coherence within the vicinity $\mathcal{N}_s$ of a particular site s, so that $V_{s,r}\,|r \in \mathcal{N}_s,\ \forall s \in \mathcal{S}$. Determining a proper criterion to define $\mathcal{N}_s$ is also part of the design process.

Both the data and smoothing terms are combined within the last step where the set of all possible labeling solutions W is explored in order to determine $\omega_s \in \mathcal{L}$ for all $s \in \mathcal{S}$ by simultaneously minimizing both terms.

3.3.2 Pre-processing

As already stated, there are several elements that degrade the quality of US images (see section 1.3.2). Pre-processing procedures are usually used before applying segmentation methodologies in order to minimize the effect of

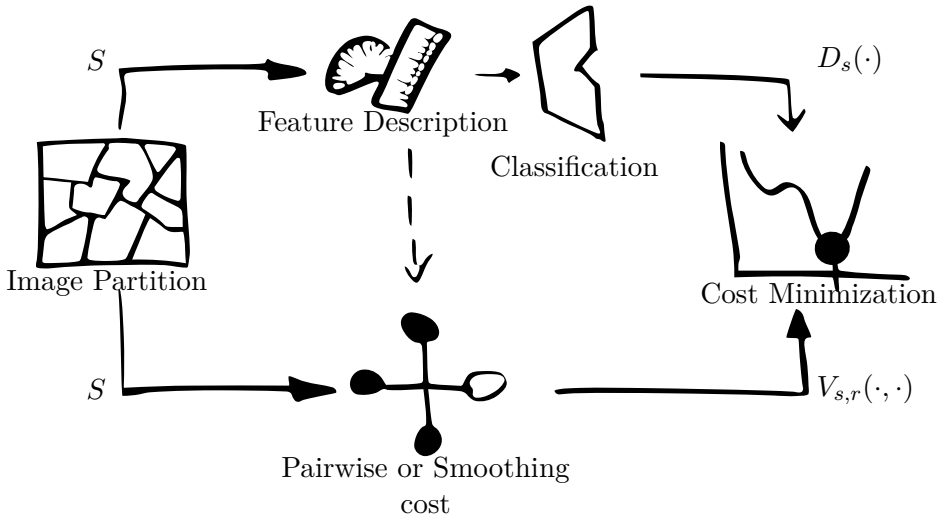


Figure 3.8: Conceptual representation of the optimization framework proposed for segmenting breast lesions in Ultra-Sound (US) data.

these image degrading elements. In order to get an idea of the pre-processing methodologies used for breast ultrasound for further segmentation of lesions, the reader is referred to [62], [108]. The use of a pre-processing step before segmenting may be appropriate for algorithms with a high reliance on the intensity of pixels, such as region growing algorithms [60] or gradient based ACM [80]. However, based on the fact that unsupervised tissue characterization can be done based on speckle signatures [120] indicating that some tissue discriminative information encoded within the speckle which therefore can be used to maximum advantage in the classification process. As such, global pre-processing to remove the speckle noise or other artifacts from the ultrasound images was not implemented. Instead, the task of conditioning the data is done at the feature description stage allowing every descriptor to perform specialized pre-processing operations.

3.3.3 Image Partition

The goal of labeling procedures is to divide the image into elements and use supervised ML in order to infer a label for each element based on a training stage. In this statement, a clear key part is the designation of these elements. Some examples of such elements can be pixels, regular patches or any collection of pixels sharing some characteristics. Actually, virtually anything associated with a set features in order to apply a classification

procedure is a feasible element to be used. The inconvenience of performing supervised ML in order to label a pixel-based representation of the image is that the information from a single element is rather meager, limited to the pixel depth or some filter response for that particular position. We can overcome these limitations by taking advantage of richer information present in more complex structures such as patches [121], sliding windows [122] or pixel clusters with similar spatial and intensity/color information [115] all of which have been used for over a decade in a multitude of Computer Vision (CV) applications, including the top-performing submissions to the multi-class object segmentation PASCAL VOC Challenge [123], [124].

Despite this wide usage, patches and sliding windows have several inconveniences. When using disjunct patches, the images are partitioned into subimages in such a way that there is no pixel belonging to two different subimages. The effect of describing the image as a collection of patches allow high level descriptors to be associated to the elements now forming the image. However, assigning a label to each of these patch elements produces gross results due to the severe discretization of the data. A direct solution consists of producing a finer sampling by reducing the size of the patches, yet the advantage of using patches declines when there are not enough pixels forming the patch to produce valuable information. On the other hand, and in the intent to overcome this drawback, a sliding window is used which extracts patches from a denser grid, allowing those patches to overlap, giving the resulting space of subimages such a large volume of data that it makes the problem, at least, difficult to handle, even if the sliding window is taken in a singular scale [125]. Another inconvenience of using regular patch structures is their inflexibility, which leads to taking into account undesired pixels within the descriptors extracted from those patches that subsequently introduce noise into the classification stage, creating the need of more robust classification techniques able to handle noisy environments. A common work-around is to use only patches and windows fully contained within the objects to ensure a more homogeneous characteristic [122], supposing a loss of possible elements that can be used for training purposes within the object boundaries.

Superpixels overcome some of the limitations mentioned by relaxing the shape of the patches in favor of irregular patches adapted to the underlying characteristics of the images. In this manner, at the expense of an unsupervised learning stage or a conservative over-segmentation of the image to generate such irregular patches, superpixels capture image redundancy and reduce the required complexity of the image processing tasks which follow, and provide appropriate regions for extracting local features from through-

out the image [115].

The underlying idea is that superpixel algorithms grip pixels into perceptually meaningful atomic regions, which can then be used to replace the rigid structure of the pixel grid. To fulfill such a goal, two equivalent approaches can be adopted:

- a conservative over-segmentation of the image [98], [100], [126], [127]
- an unsupervised learning technique taking into account the pixels appearance and its location [115], [128], [129].

Although it is difficult to determine which superpixel generation best suits a particular application, some characteristics of the superpixel methodology, such as boundary fastening, lattice regularization or computational cost, might be hints for a particular application. The best way to understand these characteristics is by visually comparing different superpixel procedure outcomes. Figure 3.9, extracted from Achanta et al. [115], where a review of some of the superpixel techniques based on clustering pixels used in the literature is given, allowing visual comparison of the strength, weakness and compromise trade-off regarding the aforesaid desirable characteristics of several superpixel approaches.

Summing up, our inclination to use superpixel technologies lies in reducing the complexity in computational terms due to the reduction of the sites set S, the possibility of extracting high-level features compared to using pixel elements, and, the fact that the final delineation gets decoupled from the classification step and linked to the partition boundaries of the superpixels. From all the work regarding superpixels or techniques that could lead to the use of superpixels, the following have been highlighted and taken into account in regard to our application of segmenting breast lesions in US data.

Quick-shift

Quick-shift [130] is a mode seeking algorithm equivalent to Mean-Shift [131] but outperforms the Mean-Shift technique in terms of computational cost producing reasonable superpixels. Mode seeking is applied to superpixel extraction to generate arbitrarily shaped clusters with no particular spacial disposition by clustering the paired data $(p, f(p)),\ p \in \Omega$ where $p \in \Omega$ are the image pixels and $f(p)$ their intensity or color coordinates. Both [131] and [130] share the same $f(p)$ representation term.

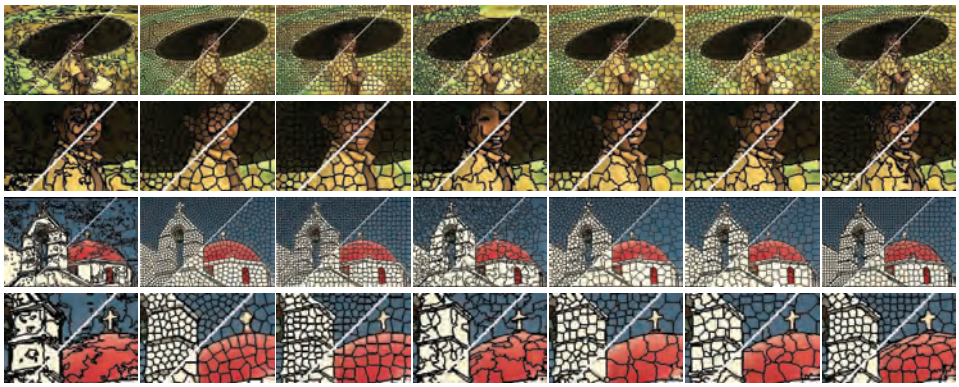


Figure 3.9: Visual comparison of super pixels produced by different methods. From left to right: a graph-based approach proposed by Felzenszwalb and Huttenlocher [127], a normalized cut proposed by Shi and Malik [100], Turbopixels proposed by Levishtein et al. [129], Quick-shift proposed by Vedaldi and Soatto [130], compact global optimization proposed by Veksler et al. [128], constant-intensity global optimization proposed by Veksler et al. [128], and Simple Linear Iterative Clustering (SLIC) proposed by Achanta et al. [115]. In the full view of the images, the overall distribution of the superpixels can be observed, and on the detail, boundary fastening or regularity can be observed. Image obtained from [128].

Figure 3.10 illustrates in a qualitative manner how the quick-shift superpixel technique fastens to the structures present within the images using some image examples.

Simple Linear Iterative Clustering (SLIC)

SLIC [115] also falls into the category of mode seeking algorithms clustering paired data of pixels and their appearance $(p, f(p)),\ p \in \Omega$ similar to Quick-shift. However, in SLIC, the treatment of the pixel and intensity/color information is different in order to generate a fairly homogeneous distribution of superpixels across the image. Despite the fact that SLIC is becoming the reference of the superpixel state-of-the-art, it has been dismissed since SLIC is unable to adhere to weak edges in a noisy environment such as found in US images due to its homogeneity sampling condition that makes it outstanding for natural imaging environments but unsuitable for our application.

Global Probability Boundary (gPb)

gPb [132] is not a superpixel technique per se. Instead, gPb is one of best performing techniques within the state-of-the-art of edge or boundary detection. gPb couples multiscale local brightness, color, and texture cues to a powerful globalization framework using spectral clustering to obtain contours and a significance description in a weighted boundary map form. This boundary detection can be used to extract superpixels, since performing a thresholding procedure at any level leads to an over-segmentation of the image that can be used as a superpixel.

Figure 3.11 illustrates the gPb detection and the superpixel delineations obtained when thresholding the gPb descriptor at different levels.

3.3.4 Feature descriptors

Feature description is a key step to generate the data-model since the correctness of the cost resulting from the classification stage directly depends on the features describing the superpixels. Features are nothing more than measures on the image carried out at each superpixel. The features from superpixels in the images in the training set are later used with their ground truth information to train the classifier on the characteristics of superpixels belonging to any target tissue: lesions, fat, fibroglandular, etc. The same features need to be extracted from the superpixels in the images to be tested, to be passed on to the classifier for a prediction of their probability of belonging to each learned class.

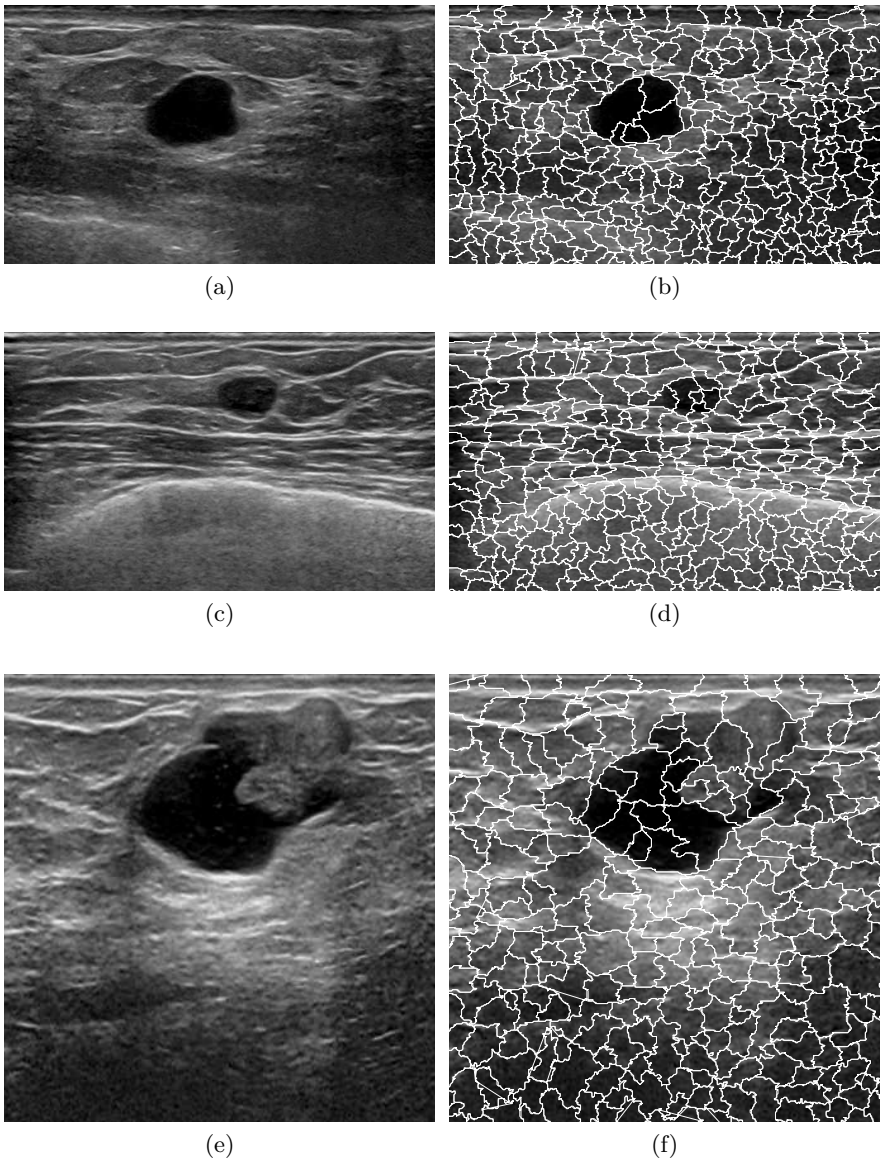
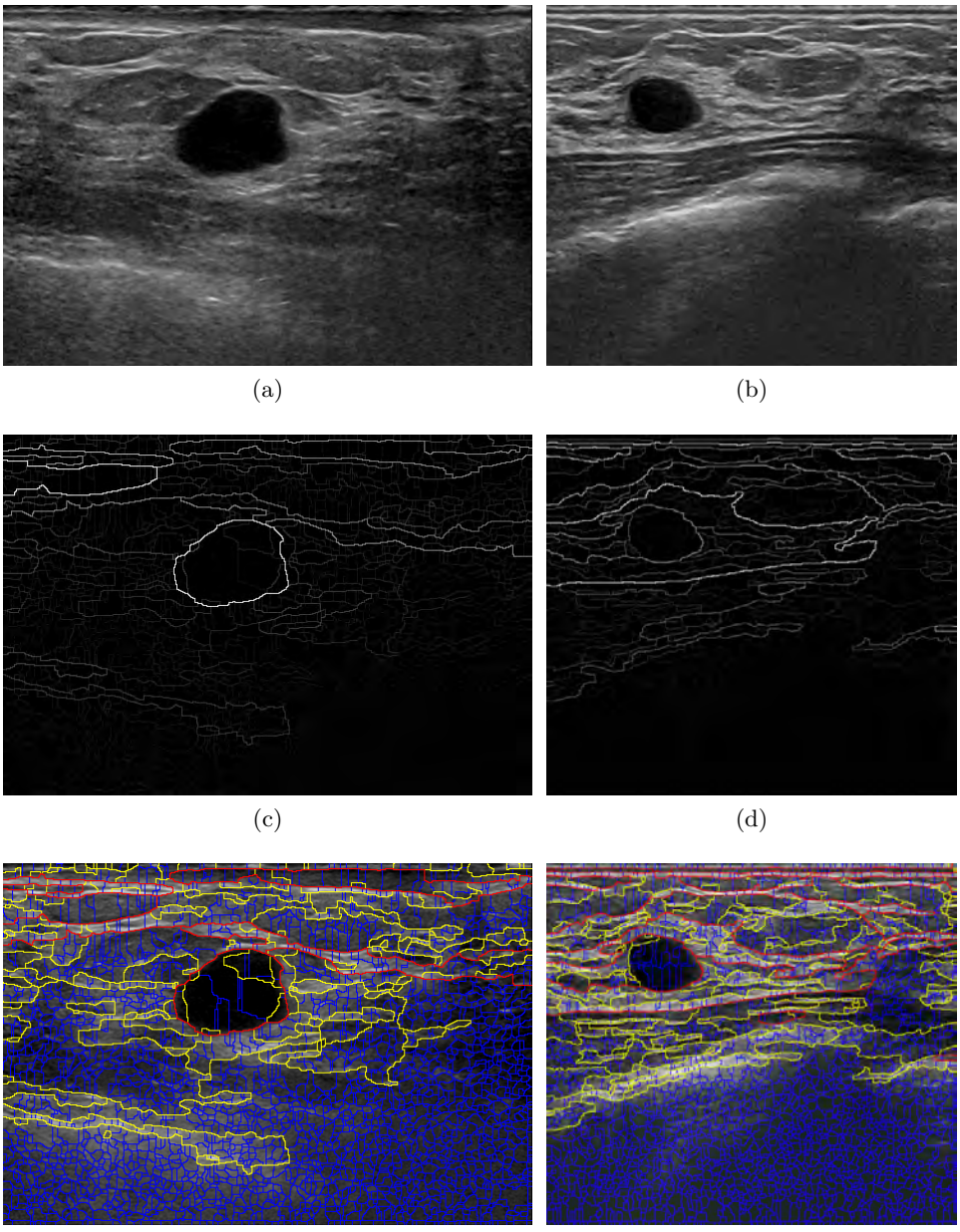


Figure 3.10: Qualitative analysis of Quick-shift [130] based superpixels. Left column represents the original image, while the right column's overlay in white shows the superpixels' boundaries.



(a) (b) (c) (d) (e) (f)

Figure 3.11: Qualitative analysis of using Global Probability Boundary (gPb) [132] as a superpixel. The top row shows the original images, the middle row the gPb value in gray scale and the bottom row illustrates the different superpixel sets obtained using different threshold values as a different color overlay delineation on the original image.

The main advantage of using superpixels to carry out feature description is that high level features can be designed to discriminate between different tissues. Defining the proper feature descriptors is crucial for the final discriminative power of the data model since the better the separability the feature offers, the better the classification performance of any classifier. When designing feature descriptors, two aspects arise: the conceptual idea of what needs to be measured and the actual measure to compare entities. As already mentioned, breast intensity and texture are highly discriminative as compiled by the *Stavros Criteria* [13] and, in one way or another, those characteristics drive all the segmentation procedures mentioned here. To give another reason in favor of these characteristics, three out of seven BI-RADS descriptors are based on quantifying these attributes by a human reader. Apart from appearance features, the tendency in placing the lesions at the center of the image by the radiologists has also proved to be a valuable information either for driving the segmentation procedure [60], [61] or for correcting the obtained segmentations [81], [85].

Here follows the feature descriptors proposed to describe the superpixels in order to segment breast lesions by labeling the superpixels based on these feature descriptors.

Describing the Brightness of the regions

US image brightness of a region is related to two aspects: the acoustic impedance difference of two tissues at their interface producing a back reflexion of the wave and the amount of scatterers within the tissue also reflecting the wave back [9]. When a radiologist assesses a breast US image in BI-RADS terms, there are two terms that refer to the bright appearance of the lesion:

Echo Pattern: Anechoic, Hypoechoic, Hyperechoic, Complex, Isoechoic.

Posterior Acoustic Pattern: No posterior acoustic pattern, Enhancement, Shadowing, Combined pattern.

Therefore, there is information encoded within the global brightness of the regions and, as stated in section 1.3.3, this information is used in relation to the elements present in the image. The amount of brightness of a region has no meaning unless it is compared with the elements present in the image. It is difficult to take advantage of the *posterior acoustic pattern*, since it fulfills more diagnostic purposes. However, trying to encode the *echo pattern* term makes sense since *Anechogenity, Hypoechogenity, Isoechogenity and*

Hyperechogenity characterize the tissues. Usually those terms refer to the bright appearance of the region with respect to a region of adipose tissue which is usually in the middle of the range offering a grayish aspect.

However, using the brightness directly, or categorizing the regions with respect to their position within the possible intensity spectrum does not work, since the brightness representation depends on the imaging system's characteristics and configuration and therefore, is inhomogeneous across the entire dataset. Instead, we propose two possible features, B_{μ} and B_{Md}, to describe the region's brightness as a quadruple that compares a statistic of the superpixel's intensity distribution with four statistics of the intensity distribution of the entire image, as described in equations 3.7, 3.8.

$$\mathrm{B}_{\mu} = abs\Big(\mu\Big(I(s)\Big) - \Big\langle \min\Big(I(\mathcal{S})\Big), \max\Big(I(\mathcal{S})\Big), \mu\Big(I(\mathcal{S})\Big), Md\Big(I(\mathcal{S})\Big)\Big\rangle\Big) \tag{3.7}$$

$$\mathrm{B}_{Md} = abs\Big(Md\Big(I(s)\Big) - \Big\langle \min\Big(I(\mathcal{S})\Big), \max\Big(I(\mathcal{S})\Big), \mu\Big(I(\mathcal{S})\Big), Md\Big(I(\mathcal{S})\Big)\Big\rangle\Big) \tag{3.8}$$

Where $\mu(I(s))$ corresponds to the superpixel's mean intensity value, $Md(I(s))$ corresponds to the superpixel's median intensity value, $\min(I(\mathcal{S}))$ corresponds to the image's minimum intensity value, $\max(I(\mathcal{S}))$ corresponds to the image's maximum intensity value, $\mu(I(\mathcal{S}))$ corresponds to the image's mean intensity value, and $Md(I(\mathcal{S}))$ corresponds to the image's median intensity value. The superpixel's quadruples are normalized by a value such that the furthest element within the features has distance 1. The equation 3.9 illustrates the normalization factor for B_{μ}. The factor normalizing B_{Md} is constructed in a similar manner.

$$\mathrm{B}_{\mu} \text{ normalization factor} = \max_{i\in[1,4];\ s\in\mathcal{S}} \Big\{{\mathrm{B}_{\mu}}^{i}(s)\Big\} \tag{3.9}$$

Figure 3.12b illustrate these features for a set of superpixels, shown in fig. 3.12a, selected to analyze how the descriptor captures the intuitive idea of *echo pattern* in the BI-RADS assessment. Each element of the feature term is represented on an axis in order to easily analyze the feature signatures of some superpixel examples. The two statistics of the superpixel proposed to use here $(\mu(\cdot), Md(\cdot))$ are represented in different colors. Figure 3.12 shows how hypoechoic regions: s_1,s_5; isoechoic regions: s_3,s_4; and hyperechoic regions: s_2,s_7; share distinguishable class signatures. Notice how the signature evolves from an almost trianglular shape pointing downwards corresponding to an anechoic region like s_1, is transformed to a diamond shape for hypoechoic regions like s_5, until it collapses for the isoechoic

regions (s_3, s_4), then opens into a rhombus shape for hypoechoic regions (s_2, s_6, s_7) to end up as an inverted diamond corresponding to highly hyperechoic streams which usually correspond to Cooper's ligaments or the pleura (s_8).

The differences between using $\mu(\cdot)$ and $Md(\cdot)$ as an intensity statistic describing the superpixel (see s_1 and s_2 signatures in fig. 3.12b) indicate that it might be interesting to use both features since there are signature variations between the two, especially in the anechoic case (s_1), where the median fully captures the fact that there are no scatters in the superpixel.

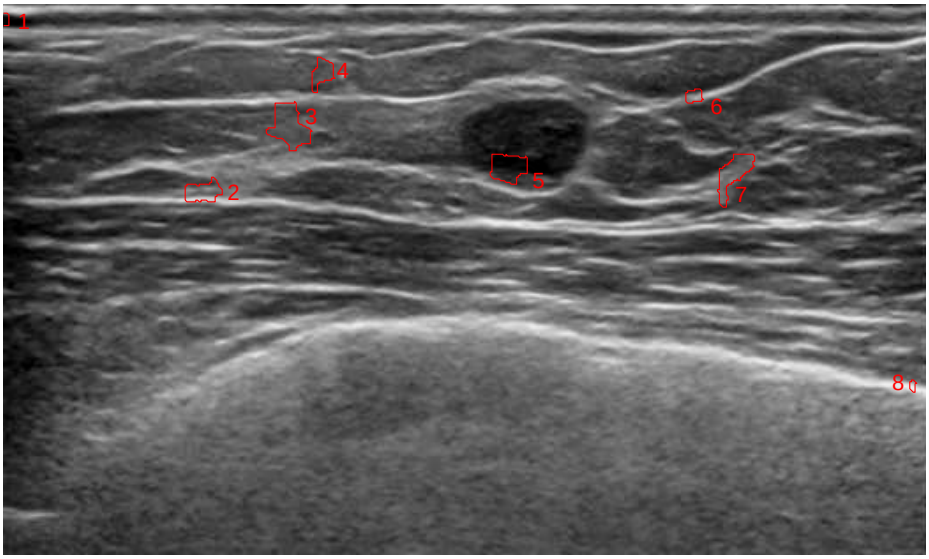
Figure 3.13 shows a qualitative representation of the brightness feature mentioned visualizing every part of the feature as an image. The images are scaled for visualization purposes since the dynamic range of the images corresponding to $|\mu(s) - \mu(I(\mathcal{S}))|$ and $|\mu(s) - Md(I(\mathcal{S}))|$ is smaller than the dynamic range of images corresponding to $|\mu(s) - \min(I(\mathcal{S}))|$ and $|\mu(s) - \max(I(\mathcal{S}))|$. Figures 3.14 and 3.15 replicate the study for a different image example, this time using quick-shift superpixels to show that, despite some differences within the shapes, the pattern remains constant. It is worth saying that every system should be trained and tested based on their superpixels, therefore, slight shape differences within the signatures do not influence one another.

Describing the overall appearance of the regions

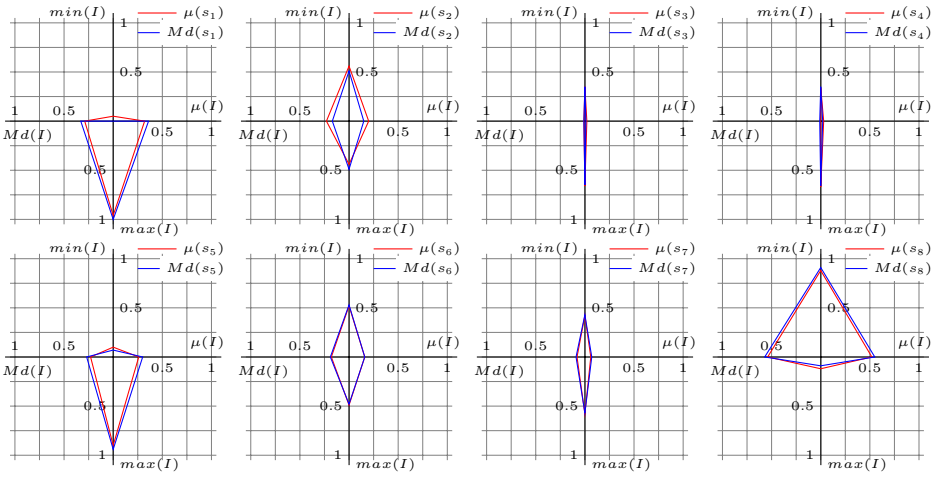
Collapsing all the information present in the brightness spectrum of the superpixels into a quadruple, as is the case of the just described brightness feature, supposes a severe discretization of the brightness information present in each superpixel. Here, it is proposed to analyze the whole brightness spectrum of the superpixel rather than just some statistics in order to build up a feature based on similitudes of the superpixels' brightness distribution. Bear in mind that for ML procedures it is good to keep the dimensionality of the final descriptor as low as possible. Therefore, despite the possibility of using the whole histogram as a feature, this is undesirable due to the dimensionality added to the final feature (even when taking advantage of down-sampling the histogram).

We propose to generate a feature in the form of n-tuple where n is the number of tissue classes present in the GT and each element within the tuple represents the Quadratic-Chi (QC) distance [133] between superpixel's brightness spectrum and the Median Absolute Deviation (MAD) [134] model brightness spectrum build up for each of tissue class.

In order to set all the superpixel's brightness spectrum in a common

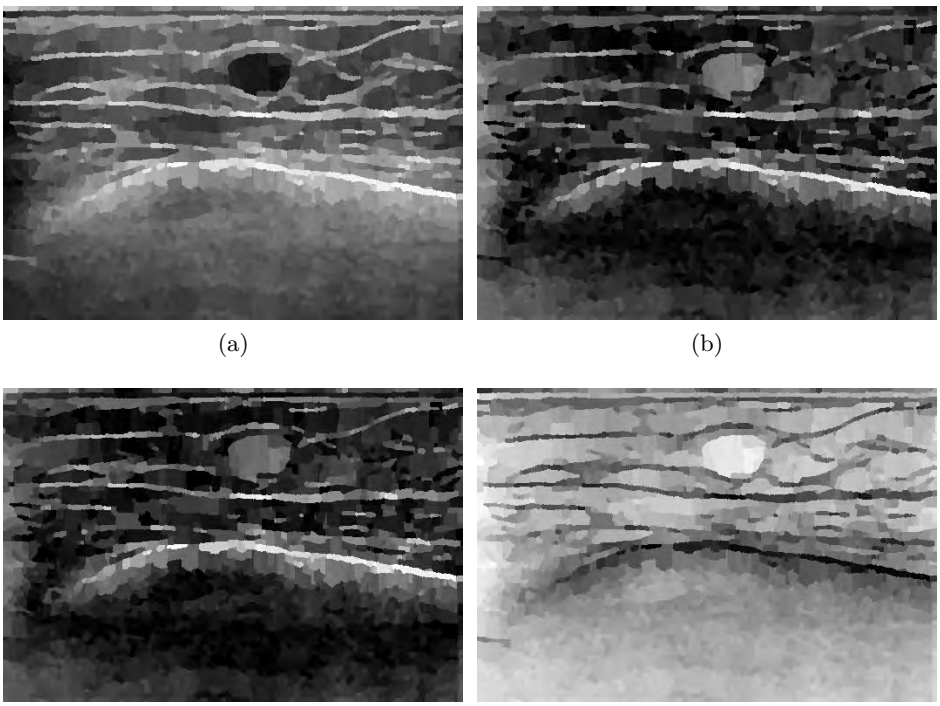


(a)



(b)

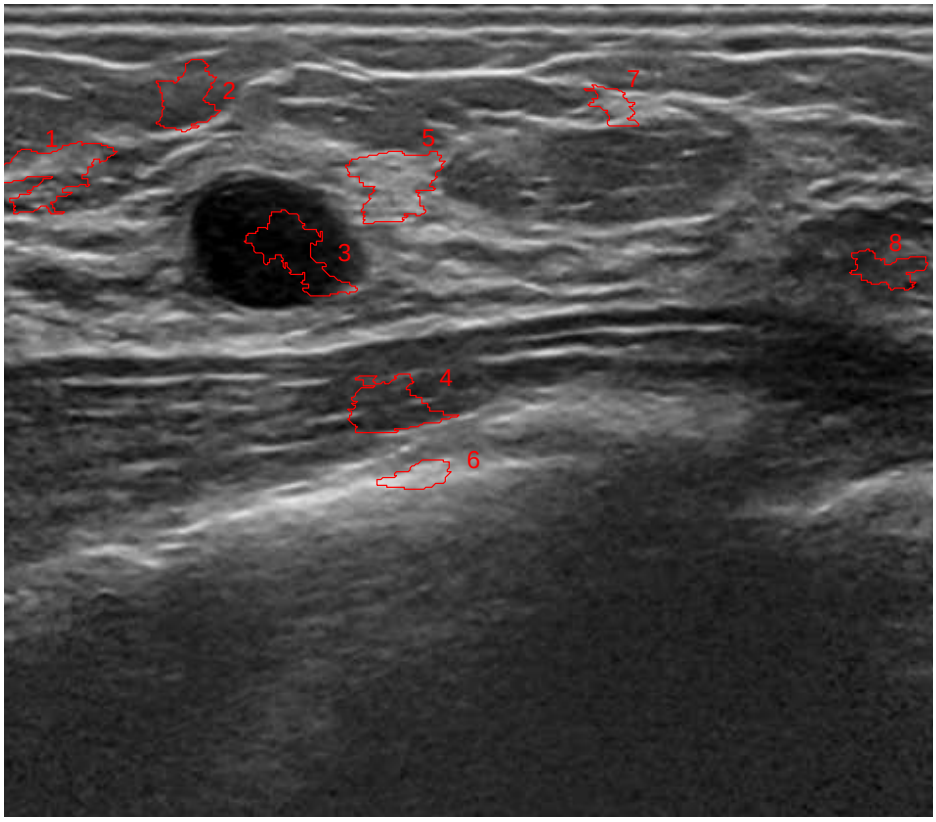
Figure 3.12: Brightness appearance feature based on comparing superpixel and image statistics. (a) example image illustrating the following region types: anechoic (1), hypoechoic (5), isoechoic (3,4) and hyperechoic (2,6,7,8). (b) illustrates the feature's signature of the regions highlighted in (a).



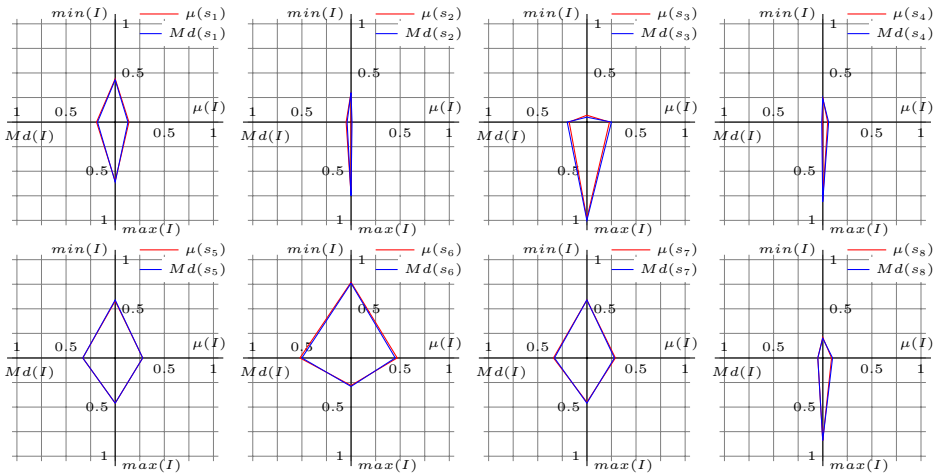
(a) (b)

(c) (d)

Figure 3.13: Qualitative examination of the brightness features of the example image used in fig. 3.12, where (a) corresponds to $|\mu(s) - \min(I(\mathcal{S}))|$, (b) corresponds to $|\mu(s) - \mu(I(\mathcal{S}))|$, (c) corresponds to $|\mu(s) - Md(I(\mathcal{S}))|$, and, (d) corresponds to $|\mu(s) - \max(I(\mathcal{S}))|$. The intensity of the images has been stretched for visualization purposes.

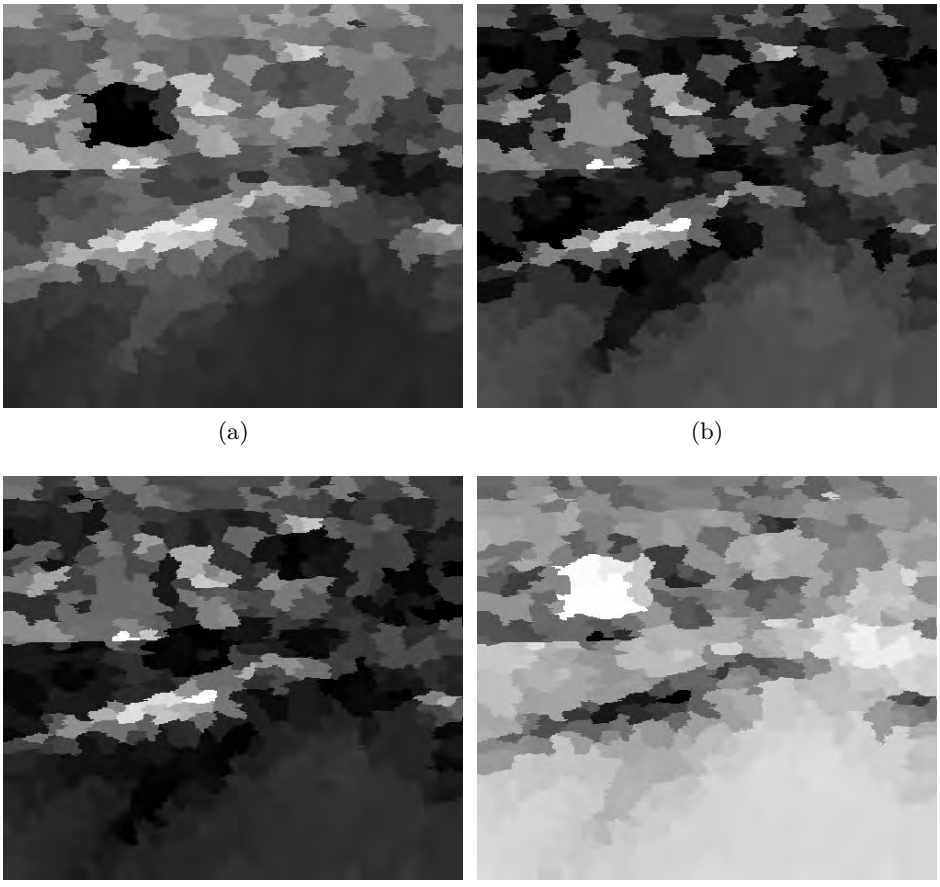


(a)



(b)

Figure 3.14: Brightness appearance feature based on comparing superpixel and image statistics. (a) example image illustrating the following region types: anechoic (3), hypoechoic (8), isoechoic (2,4) and hyperechoic (1,5,6,7). (b) illustrates the feature's signature of the regions highlighted in (a).



(a) (b)

(c) (d)

Figure 3.15: Qualitative examination of the brightness features of the example image used in fig. 3.14, where (a) corresponds to $|\mu(s) - \min(I(\mathcal{S}))|$, (b) corresponds to $|\mu(s) - \mu(I(\mathcal{S}))|$, (c) corresponds to $|\mu(s) - Md(I(\mathcal{S}))|$, and, (d) corresponds to $|\mu(s) - \max(I(\mathcal{S}))|$. The intensity of the images has been stretched for visualization purposes.

reference, the histogram of every and each superpixel is re-sampled using 100 beams equally spaced between its lowest and highest intensity value, and further normalized to ensure that the area of each histogram is one. To build up the models for each and every one of the tissue classes in the GT (*i.e:* air, fat, fibro-glandular, lesion, *etc.*), the histograms of the superpixels belonging to each tissue class are grouped together to calculate the MAD model of each class accordingly to equation 3.10.

$$\mathrm{MAD}_i^c = \mathrm{median}\Big(\mathrm{hist}_i(s) - \mathrm{median}(\mathrm{hist}_i(s), s \in \mathcal{S}_c); s \in \mathcal{S}_c\Big) \tag{3.10}$$

Where c represents every tissue class in the GT and i represents the beam index of the model. For this case $i \in [1, 100]$ since the histograms have been re-sampled using 100 beams. $\mathrm{hist}_i(s)$ corresponds to the value of the i_{th} beam of the re-sampled and normalized histogram. $\mathcal{S}_c$ represents the set of superpixels that share the same GT label c. Finally, the model needs to be normalized so that the sum of all the beams forming the model add 1.

Once determined the tissues' appearance model, in order to determine the feature describing a superpixel s, the superpixel's brightness spectrum is re-sampled an normalized. The QC histogram distance is now used to compare this histogram with all the normalized MAD models in order to build up the current superpixel feature. The process finalizes by normalizing the feature such that the sum of all the distances to the MAD models adds 1.

Describing the texture appearance of the regions

Texture is a widespread phenomenon, easy to recognize and hard to define [135]. Texture of a point is undefined, it is simply a property of the area that can be related to spatial repetition of structures, similar statistical properties of the area, or both. The texture in US images is produced by speckle (see section 1.3.2), an undesired artifact produced by aleatory backscatter from particles within the tissues depicted that give US images their distinctive appearance [9]. Despite being an unwanted artifact, the texture produced by speckle brings relevant information that helps to discern tissues as reported in *Stavros criteria* [13], BI-RADS assessment [12], and other related work segmenting breast lesions in US data [61], [81], [118].

Due to the aleatory nature of the texture, a stochastic description of the data is preferred over a repetition analysis. Taking into account that US images depict a quantity of acoustic reflection as a result of tissue interfaces and tissue inhomogeneities, producing images with an abundance of high

frequencies, so a stochastic descriptor describing gradient information has been adopted. The solution adopted here consists of using a multi level gradient descriptor, Self-Invariant Feature Transform (SIFT), extracted in a dense grid (one descriptor for every pixel) which wraps up these SIFT descriptors using Bag-of-Words (BoW) as a global texture descriptor at the superpixel level. An equivalent feature to encode information similar to the information encoded in the SIFT descriptors is HOG. HOG has already been used in combination with DPM [83] in [136] to detect and segment breast lesions in an optimization framework similar to the one proposed here. However, Lazebnik et al. [137] established that well-designed BoW methodologies can out-perform more sophisticated methodologies based on parts and relations.

Self-Invariant Feature Transform (SIFT) [138] transforms key-points into scale-invariant coordinates relative to local features. Dense SIFT uses every pixel in the image as a key-point in order to map the whole image in this space. The SIFT descriptor was inspired by a biological vision model proposed by Edelman et al. [139]. Initially, the key-point scale and orientation are determined. The image gradient's magnitude and orientation are then sampled in a search window of sampling regions according to the key-point scale and rotated according to its orientation to achieve rotation invariance. The typical search window is 16x16 sampling regions. These samples are weighted by a Gaussian window to ensure smoothness transition and give more weight to the gradients found close to the key-point. Figure 3.16 illustrates this process and how the final descriptor is generated using a search window of 8x8 sampling regions for illustrative purposes. In order to generate the final descriptor, groups of 4x4 sampling regions are used to group the samples into orientation histograms of 8 bins, leading to a 4x4x8=128 element feature vector when using a 16x16 search window, which is reported to achieve the best performance [138].

The Bag-of-Words (BoW) technique is a well known technique solving the document classification problem consisting of correctly classifying text documents into different categories and analyzing the occurrence of a set of keywords [140]. In recent years, the BoW technique has been introduced to CV applications [141] where it can also have the name Bag-of-Features (BoF). During these years, BoF has been widely tested showing remarkable success in texture and object recognition. The idea remains the same: perform an occurrence study of a keywords set in order to correctly classify an image, subimage, patch, etc. The only difference is that in CV applications, there is no direct occurrence of keywords in the images. Therefore, the images need to be represented in terms of these keywords often

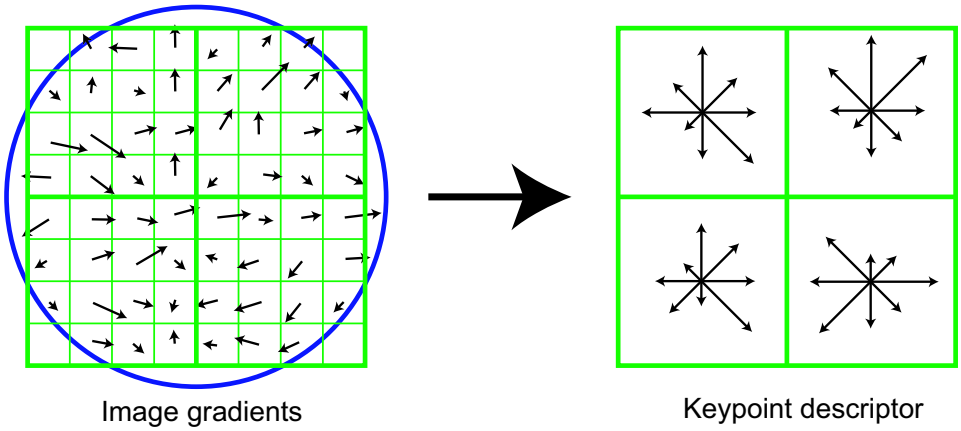


Figure 3.16: Self-Invariant Feature Transform (SIFT) descriptor illustration extracted from [138].

referred to as a visual dictionary. Figure 3.17 illustrates the process. From an image training set, N dimensional local visual features are determined. In our case, the images are transformed by SIFT descriptors, constituting a 128 dimension feature space. The large set of visual features from all the training data is then used to generate a visual dictionary, using an unsupervised learning technique to cluster the visual features data. In our case, k-means is used to generate this visual dictionary. Up to this point, the BoF is building offline. Then, for a test image, this is represented in the feature space and every feature is assigned to the nearest cluster. In this manner the target image is represented in terms of visual words belonging to the defined dictionary so that a histogram word occurrence can be generated and used as the new feature. In our case, a 36 beams histogram is generated for every superpixel to determine the words occurrence from a 36 SIFT word dictionary previously generated. Those histograms are further used to determine the data model.

The selection of the dictionary's size is carried out empirically, such value has to be large since it would condition the separability of the data but it has to remain bounded since it determines the length of the feature and low dimension features are preferred. During the designing process the size of the dictionary has been set to 36 since it complies with the mentioned criteria.

As an illustration of the procedure, figure 3.18 visually represents, in two different ways, a SIFT descriptor further used to interpret a 36 word SIFT

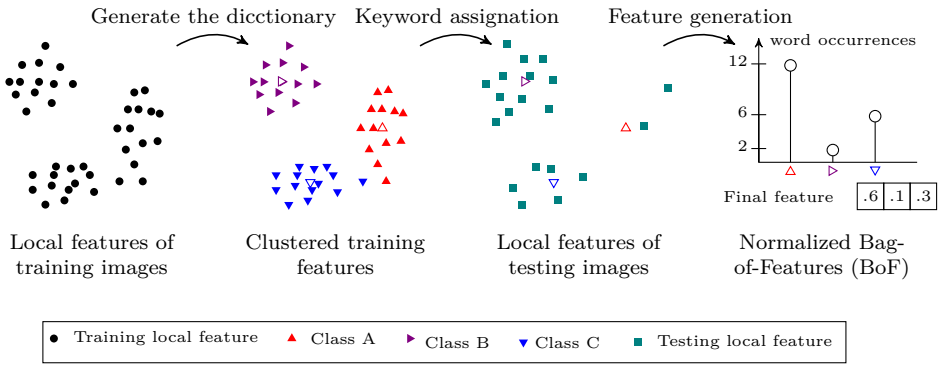


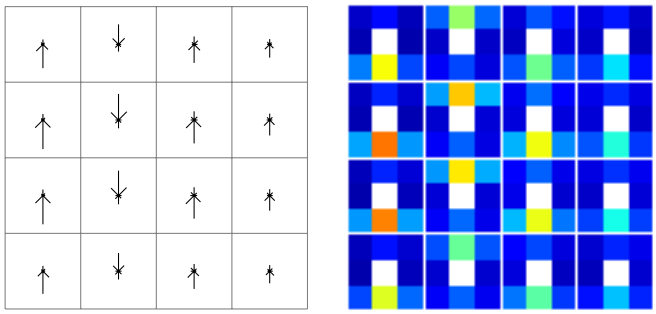
Figure 3.17: Representation of the Bag-of-Features (BoF) procedure. This process is also found in the bibliography under the name Bag-of-Words (BoW).

dictionary (example illustrated in figures 3.19 and 3.20). Figure 3.20 also illustrates the similitude of the words comprising the dictionary and projects the centroids onto a plain using Principal Component Analysis (PCA) in order to visualize the relationship between the clusters. Each word is associated with a color in order to interpret the SIFT descriptors in figure 3.22 extracted from the original US images in figure 3.21. Notice that a dictionary of 36 words already allows texture patterns to be revealed keeping a fairly reduced feature dimension.

Describing the location of the regions

Taking into account the acquisition process of breast US images and the architecture of the breast, (see section 1.3.1) some areas in the images are more likely to present certain tissues or structures. Taking advantage of this information is not unusual [60], [61], [73], [78], [116] and can be found either as a feature or a domain application criteria used to refine the segmentation. The choice preferred for this work was to use spatial information as a feature to drive the segmentation rather than use it to refine the results.

Directly providing superpixel position to the classifier is a valid approach to incorporate the spatial information in the data model. However, this approach implies that the classifier has to be trained with a fair number of samples covering the whole space. Therefore in this application, the generation and use of an atlas has been chosen. The atlas option allows using all the training data to build up a model of the tissue distribution similar to the posterior probability distribution of the lesion's position shown in



(a) (b)

Figure 3.18: Self-Invariant Feature Transform (SIFT) descriptor visualization corresponding to a word within the dictionary example from figures 3.19 and 3.20.

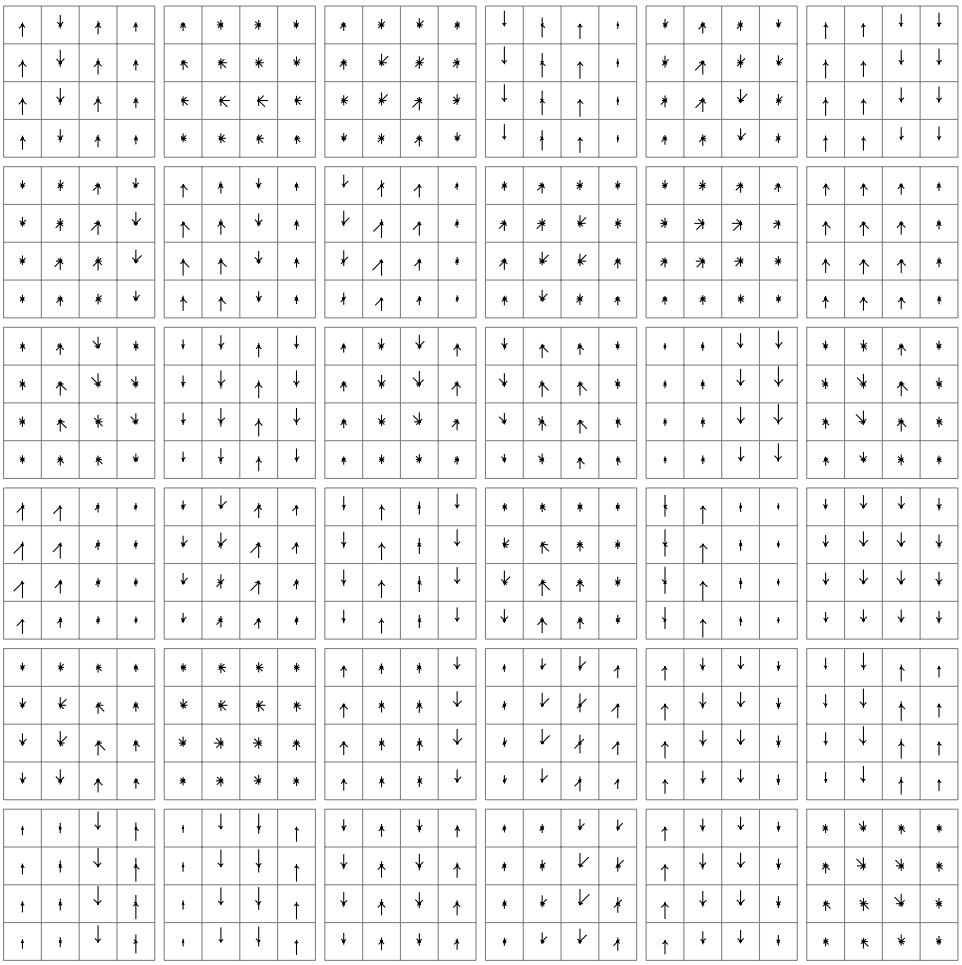
figure 3.3 in section 3.2.2.

Multi-resolution

Multi-resolution offers a natural, hierarchical representation of the information. Such ideas are not new, and they have been around for quite a while inspiring ideas in many fields like mathematics, physics, data analysis, signal processing, etc. [142]. A classical way to exploit multi-resolution in CV and image processing is to incorporate features based on wavelet transform [143] which has already been applied to breast imaging in US data, for lesion detection [144] and lesion diagnosis [145] purposes.

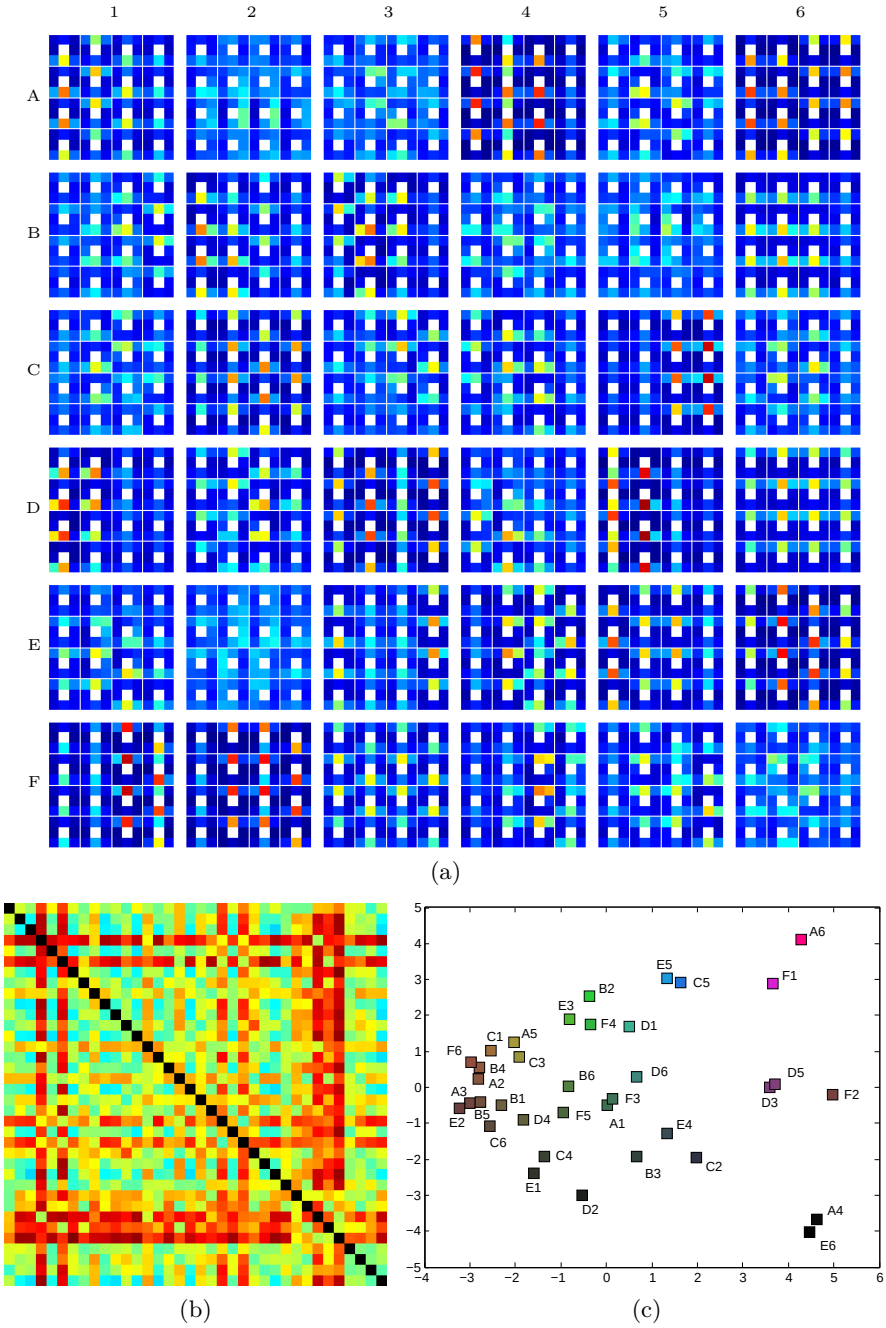
What is here proposed as multi-resolution differs from the multi-resolution aforesaid, since in our case the description of the image information at different scales is somehow already been taken into account when describing the image texture using SIFT, due to the fact that in SIFT, the gradients' analysis is carried out already at multiple scales. The multi-resolution here proposed consists on recompute the statistics (or features) of the superpixel but instead of using only the elements forming the current superpixel, the elements within the neighboring superpixels n-steps further are also used to compute the current superpixel's statistics (or features).

Figures 3.23 to 3.26 qualitatively illustrate for a particular image and superpixel type, how the brightness appearance feature evolves while increasing the step of neighboring superpixels involved in the calculation of the current superpixel's descriptor. As aforesaid, the feature proposed to capture the brightness feature of the superpixels is a quadruple (see eq. 3.8).



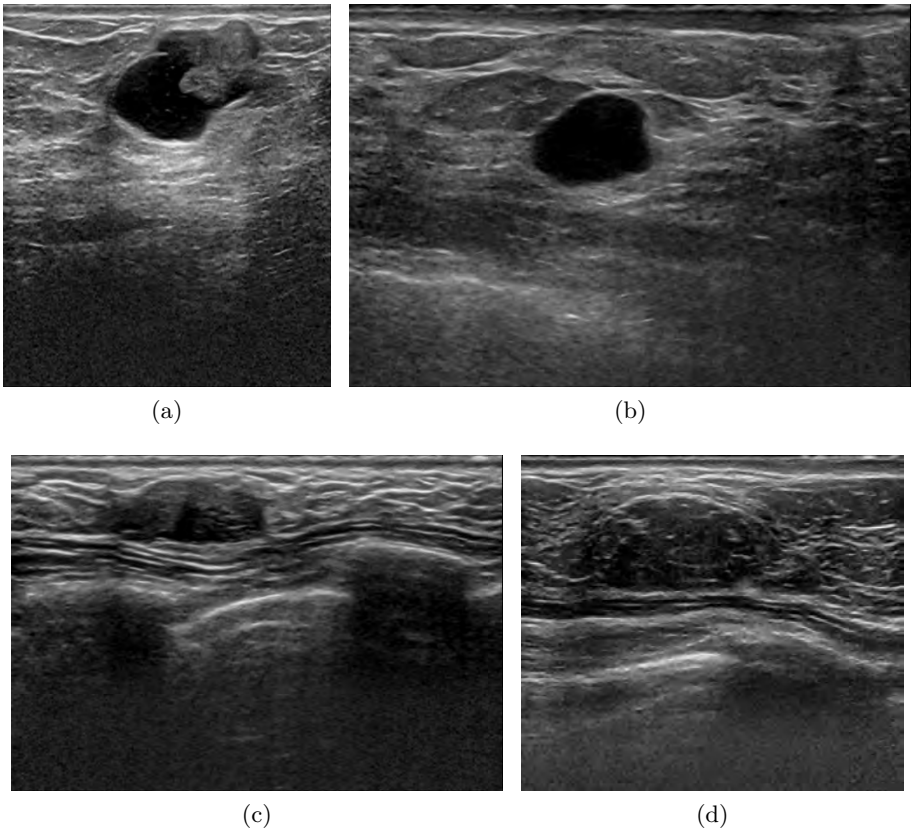
(a)

Figure 3.19: Self-Invariant Feature Transform (SIFT) oriented histogram bins visualization of a 36 words SIFT dictionary generated from a training dataset of US images of the breast



(a) (b) (c)

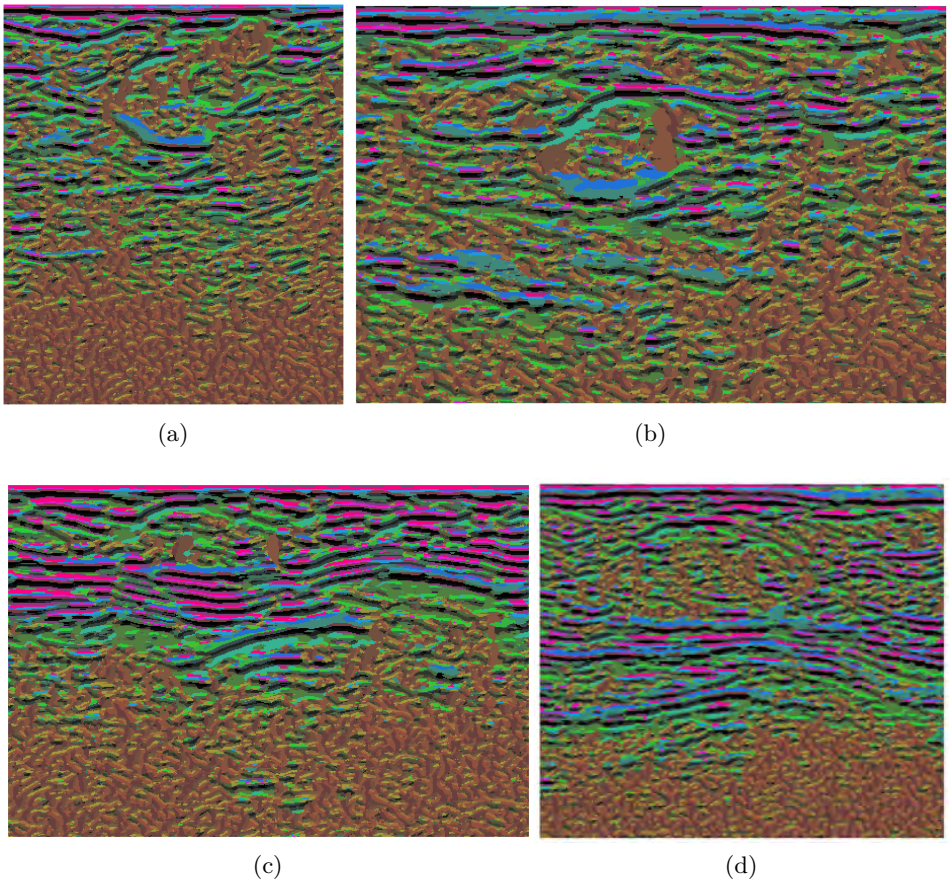
Figure 3.20: SIFT dictionary interpretation. (a) Color coding of the dictionary used in figure 3.18. (b) Illustrates the distances between the words forming the dictionary. (c) Words distance reinterpretation by mapping (b) 2D grid. The colors associated to each word are used to interpret the SIFT features in figure 3.22



(a) (b)

(c) (d)

Figure 3.21: Breast ultrasound image examples used to illustrate the SIFT texture in figure 3.22



(a) (b)

(c) (d)

Figure 3.22: Self-Invariant Feature Transform (SIFT) texture image interpretation. Each position has been colored according to the color associated to each cluster of the SIFT descriptors extracted from the images in figure 3.21. The color association can be found in figure 3.20c.

In this regard, figure 3.23 illustrates the element of the feature computed as the distance between the superpixel's median intensity value and the image's minimum intensity value; figure 3.24 between the superpixel's median and the image's mean; figure 3.25 between the superpixel's and the image's median; and figure 3.26 shows the distance between the superpixel's median and the maximum intensity value of the image.

Any other multi-resolution feature used here would be calculated in the same exact manner.

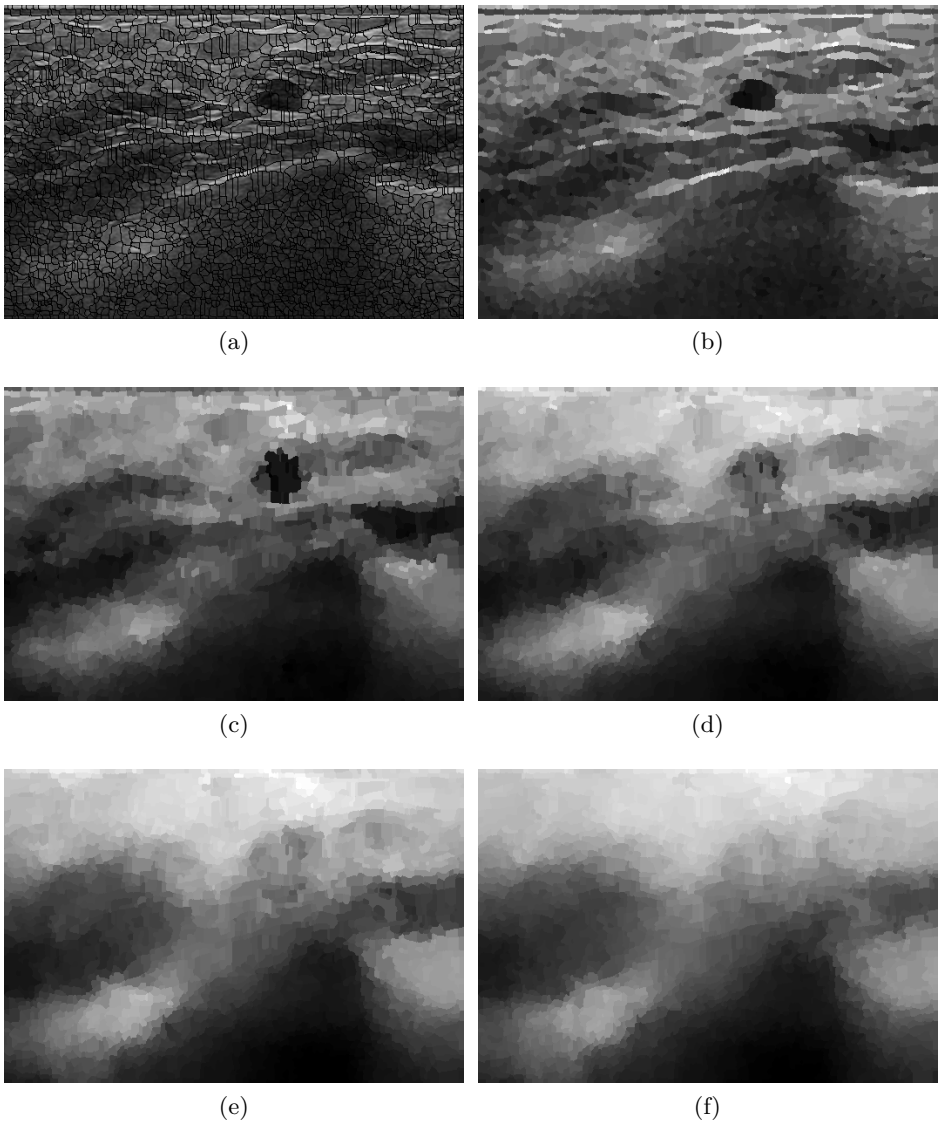
3.3.5 Classification or data model generation

The data term (or data cost) is the associated cost based on the data when assigning a concrete label to a particular superpixel. There are many ways to link this cost to the data, however, using ML or Pattern Recognition (PR) techniques offers standard stochastic frameworks to determine the cost based on feature observations ($\bar{x} \in \mathcal{X}$) and their occurrence within a training set ($\mathcal{D}$) as expressed in equation 3.11.

$$(\bar{x}, \mathcal{D}) \rightarrow \mathcal{R} \tag{3.11}$$

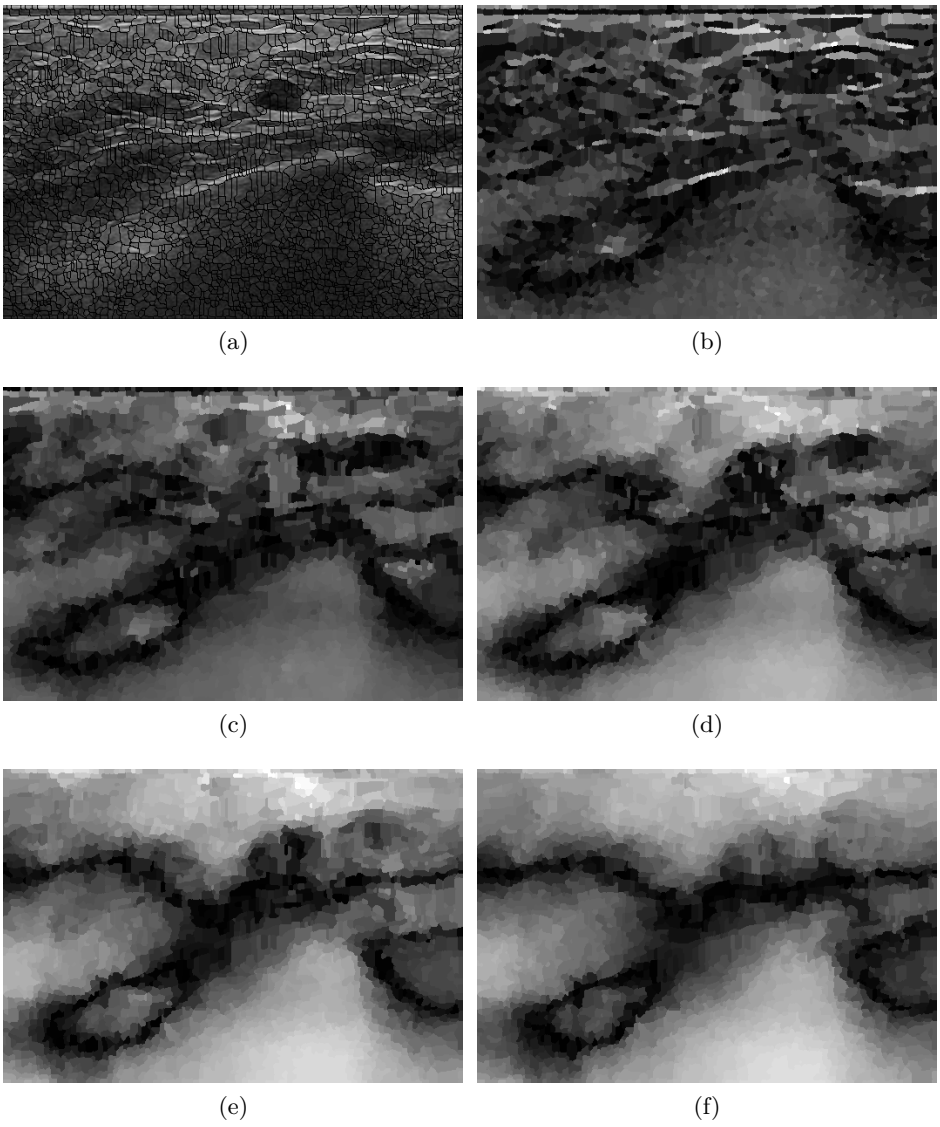
In ML and PR there is a rich body of work offering a wide range of techniques that are able to infer or map a cost term from a training data and a set of features (eq: 3.11). The election criteria to determine which is the most adequate technique for a particular application is diverse, and includes among others:

- The need of incremental training.
- Data typology regarding if it is categorical, numeric or mixed.
- The need of a discrete classification output or a continuous probability or reward.
- The presence of correlation within the features describing the data.
- Linear separability of the data.
- Accuracy.
- The presence of outliers within the training data or poor separability of the data.
- Computational and time requirements or restrictions during the training stage.



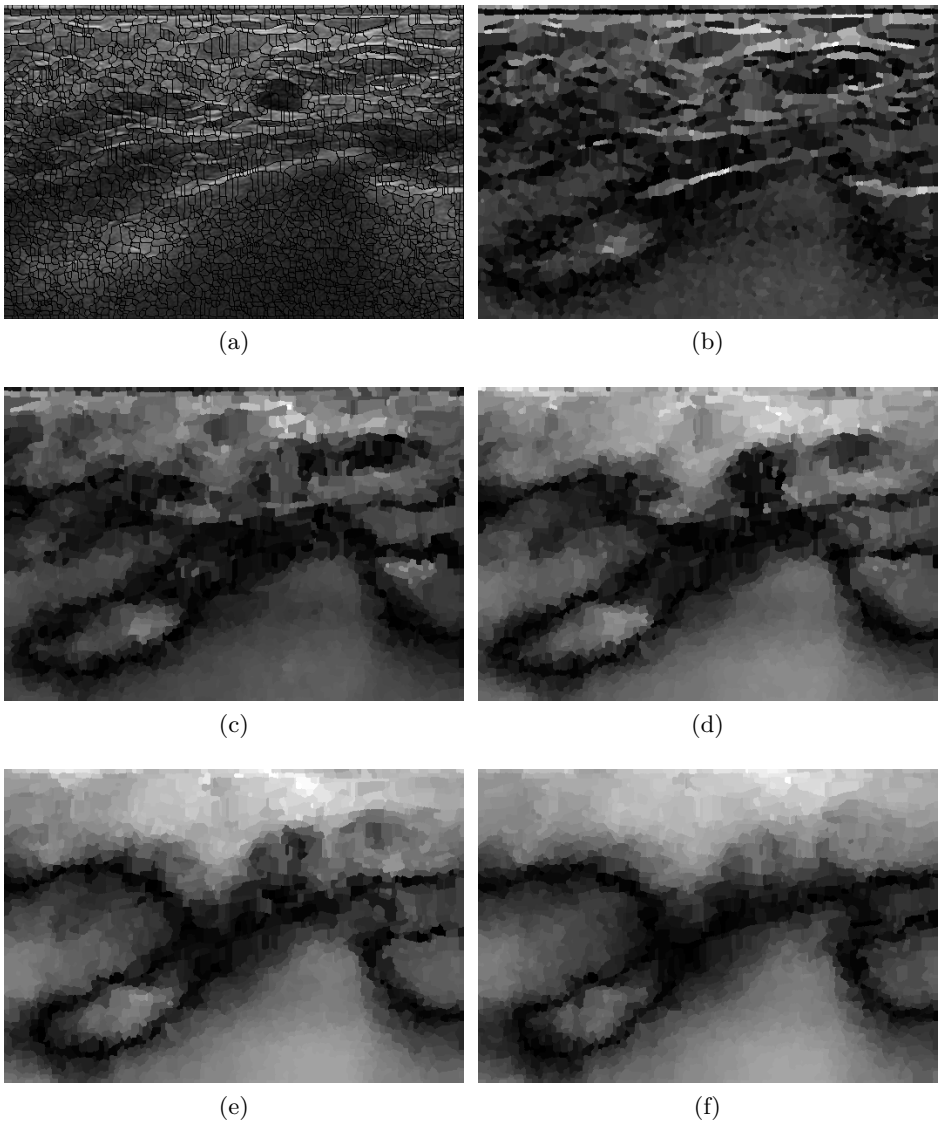
(a) (b) (c) (d) (e) (f)

Figure 3.23: Multi-resolution example for a given image and Global Probability Boundary (gPb) superpixel. (a) original image with the superpixel's delineation as overlay. (b-f) represent the distance between the minimum intensity value of the image and the median intensity value of the different superpixel groups based on their neighboring distance: (b) 0 distance, only the current superpixel is used to compute the group statistic; (c) using neighbors at distance 1; (d) at 2; (e) at 3; and (f) at 4.



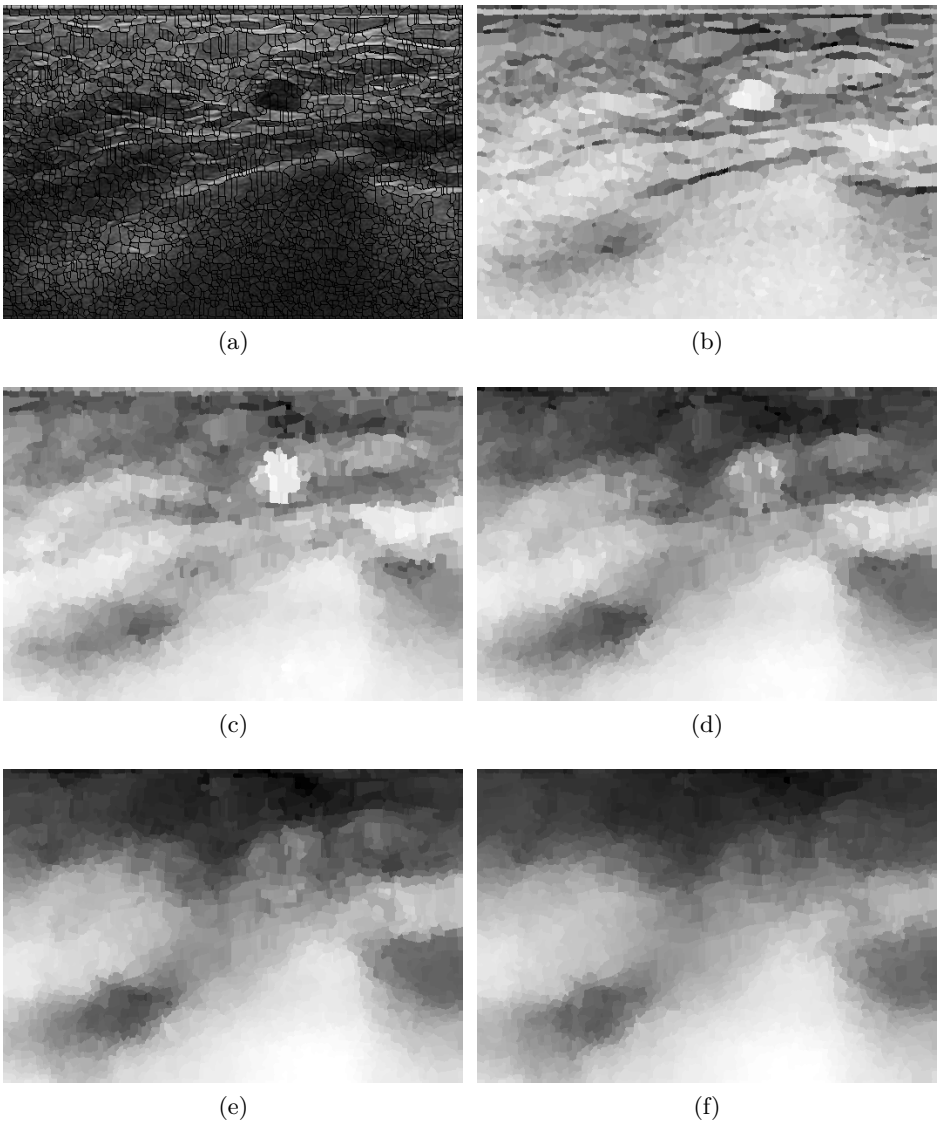
(a) (b) (c) (d) (e) (f)

Figure 3.24: Multi-resolution example for a given image and Global Probability Boundary (gPb) superpixel. (a) original image with the superpixel's delineation as overlay. (b-f) represent the distance between the mean intensity value of the image and the median intensity value of the different superpixel groups based on their neighboring distance: (b) 0 distance, only the current superpixel is used to compute the group statistic; (c) using neighbors at distance 1; (d) at 2; (e) at 3; and (f) at 4.



(a) (b) (c) (d) (e) (f)

Figure 3.25: Multi-resolution example for a given image and Global Probability Boundary (gPb) superpixel. (a) original image with the superpixel's delineation as overlay. (b-f) represent the distance between the median intensity value of the image and the median intensity value of the different superpixel groups based on their neighboring distance: (b) 0 distance, only the current superpixel is used to compute the group statistic; (c) using neighbors at distance 1; (d) at 2; (e) at 3; and (f) at 4.



(a) (b) (c) (d) (e) (f)

Figure 3.26: Multi-resolution example for a given image and Global Probability Boundary (gPb) superpixel. (a) original image with the superpixel's delineation as overlay. (b-f) represent the distance between the maximum intensity value of the image and the median intensity value of the different superpixel groups based on their neighboring distance: (b) 0 distance, only the current superpixel is used to compute the group statistic; (c) using neighbors at distance 1; (d) at 2; (e) at 3; and (f) at 4.

- Computational and time requirements or restrictions during the testing stage.

Once a subset of suitable techniques has been determined, a common practice is to test all the different techniques and configurations against the training dataset to determine the best classifier and configuration. Up to a certain point, it is normal to perform a brute force search to determine which methodology leads to the best results based on the training samples available.

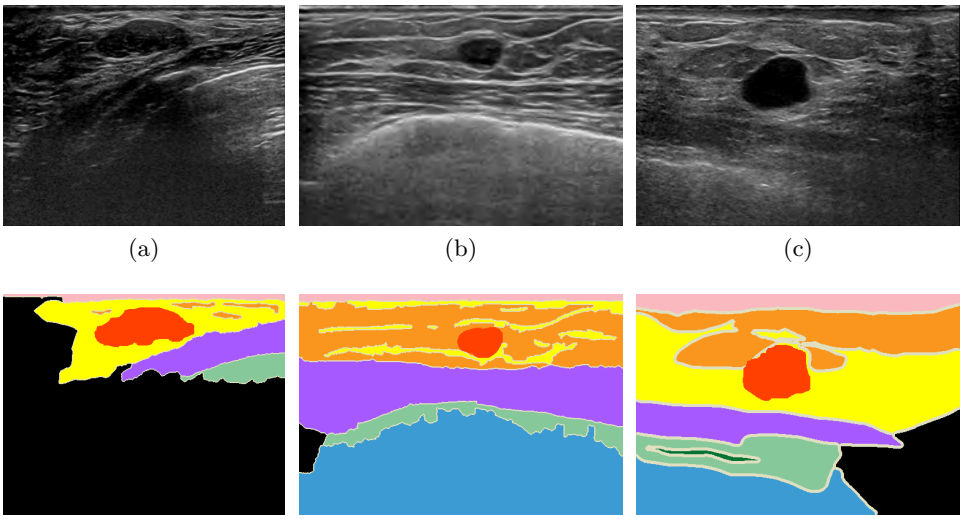
Among the different classification architectures that could be used in this particular problem, Support Vector Machine (SVM) in conjunction with a Radial Basis Function (RBF) kernel has been chosen. The need of a continuous class belonging probability to be used as the data term, the non-linear separability nature of the data, the presence of outliers within the training data as well as the need of accuracy have been specific criteria to lead to that choice. Although other classification architectures have been considered, it has been demonstrated that SVM is a very well-known non-probabilistic framework that can be trained as a MAP solver offering a computational advantage over probabilistic methods [140].

In order to implement the SVM architecture for the optimization framework, the very well-known LIBSVM library [146] has been used. The library, apart from outputting the classification of each testing sample, offers a class belonging probability which is directly used to build up the data term cost. In order to cope with non-linear separable data, such in the present case, the SVM needs classifier needs to be provided with a RBF kernel which has a parameter corresponding to its bandwidth. Following the library author's recommendations, when constructing a RBF-SVM classifier based on F features, this parameter corresponding to the kernel's bandwidth is set as $\frac{1}{F}$.

3.3.6 Pairwise or smoothing modeling

The *pairwise* or smoothing term is used to incorporate biases and assumptions in order to overcome the ambiguity and unreliability of the data-models caused by the same ambiguity, unreliability and incompleteness of the data observed. The *pairwise* term introduces low-level regularization based on MRF in order to impose a coherent labeling similar to that in GT. Figure 3.27 illustrates several GT delineations in order to observe the presence of large homogeneous label regions.

In order to implement the aforesaid homogeneity, the pairwise term in



(a) (b) (c)

(d) (e) (f)

Figure 3.27: Multilabel Ground Truth (GT) examples illustrating label coherence. (a-c) original images, (d-f) GT tissue labeling.

equation 3.6 takes the form of equation 3.12. Where β is a small value that ensures that the smoothing term doesn't take over the data term.

$$V_{s,r}(\omega_s, \omega_r) = \begin{cases} \beta, & \text{if } \omega_s \neq \omega_r \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

3.3.7 Cost minimization

Determining how to search for the optimal solution within the solution space is as important as defining the elements used to build the optimization function, and defining either of the cost terms determining the goodness or cost of the solution. Two major concerns while designing the minimization are (1) proper management of the local minima present in nonconvex functions, and (2) the space and time efficiency of the minimization algorithm. These two essential aspects of the minimization procedure design are fairly contradictory since avoiding local minima within a noneconvex arbitrary function requires an exhaustive search throughout the solution space.

Although there is no algorithm to guarantee a global minimum with good efficiency [147], there is an extensive bibliography of methodologies and strategies to overcome the trap of local minima and reach, suboptimal

solutions at a fairly computational cost [105], [148]–[150]. This field of research is not particular to image segmentation, therefore, cost minimizing techniques come from extremely varied historical backgrounds and follow different paradigms and philosophies [148]. This leads to a multitude of minimization algorithms: deterministics, stochastic, heuristics, inspired by physic events, biological behaviors, etc. (see [150]).

Despite any cost optimization being valid to Computer Vision (CV) applications, historically Iterated Conditional Modes (ICM) [151], Simulated Annealing (SA) [152] and Graph-Cut (GC) [104] have been the most commonly used in this field. Szeliski et al. [149] conducted an exhaustive review in terms of solution quality and runtime of the most common energy minimization algorithms used in CV and proposed a set of energy minimization benchmarks drawn from published CV applications. From that review, it was concluded that the state-of-the-art in energy minimization has advanced significantly since the early ICM or SA methods.

Although an extensive review of energy minimization methodologies is beyond the scope of this work, the methods named in this section are briefly described here to complete the idea of searching through the solution space to find the underlying labeling. No further discussion regarding the choice of using GC for our application is carried out, we simply state the fact that GC is the most popular energy minimization technique in the state-of-the-art applied to CV.

Iterated Conditional Modes (ICM)

ICM [151] is a deterministic minimization performing a local search using a greedy policy to iteratively reach a local minimum. Using an estimate of the labeling (ω^0) for every site $s \in \mathcal{S}$, the label ω_s^{k+1} is chosen to obtain the largest decrease of the energy function $U(\omega^k_{\{\mathcal{S}-s\}} \cup \omega_s^{k+1})$. This process is repeated until convergence is reached, which is reported to be guaranteed and very rapid [149].

Unfortunately, the results are extremely sensitive to the initial estimate (ω^0), especially for high-dimensional spaces with nonconvex energy functions as happens in CV applications. A common practice leading to acceptable results is to use the labeling producing the lowest data cost as initial estimate [149].

A way to avoid local minima is to randomize multiple initializations and apply ICM to perform a local search from the different initial solutions. A procedure with this characteristics would fall into the Greedy Randomized Adaptive Search Procedures (GRASPs) category [153] where problems of

how to correctly randomize the space arise.

Simulated Annealing (SA)

SA [152] is a stochastic methodology that performs a randomized sampling of the search space. This algorithm is motivated from an analogy with the annealing process used to find low-energy states of solids.

For this algorithm, a cooling criteria in order to simulate the annealing is needed. The algorithm starts with an arbitrary labeling ω^{0,T_0}. At every iteration, ω^{k,T_k} is randomly perturbed. A site $s \in \mathcal{S}$ is randomly selected as well as a new label for ω_s^{k',T_k}. This new possible configuration ω^{k',T_k} is either accepted ($\omega^{k+1,T_k} = \omega^{k',T_k}$) or declined ($\omega^{k+1,T_k} = \omega^{k,T_k}$), according to a Metropolis criterion. If the perturbation implies an overall reduction of the cost $U(\omega^{k',T_k}) < U(\omega^{k,T_k})$, then the perturbation is accepted. However, the perturbation can still be accepted based on a random event where the acceptance probability is related to the magnitude of the cost increase ($\triangle U$) and the current state of the parameter *temperature*. Basically, a move is more likely to be accepted if the temperature is high and the cost increase is low. Once stability has been reached for the current temperature, it is lowered according to the cooling criteria. The cooling criteria or schedule along with the temperature stage criteria stability conditions the goodness of the solution reached and its computational cost. As opposed to other methodologies, SA asymptotically converges to the global minimum when assuming an infinite number of iterations, otherwise the global minimum is not guaranteed.

Figure 3.28 intuitively illustrates the behavior of an SA procedure. Figure 3.28a shows a toy energy function defined in a single continuous dimension space W and a current labeling ω^k. Illustrated in two different colors are the elements in the space that would be accepted when randomly sampled. Each color represents a different acceptance policy and is subject to the current temperature of the system. The higher the temperature, the wider the range of accepted transitions. All the random samplings in the labeling space causes the function cost to be expressed across the time (or iterations k) behavior as in figure 3.28b where large increases in energy cost are allowed at the initial stages and restricted in further stages until convergence.

Improvements for this minimization technique, among others, include: variations and more general forms of the acceptance rule rather than the Metropolis criteria [154], or algorithm parallelization [155]. For more details, the reader is referred to [154].

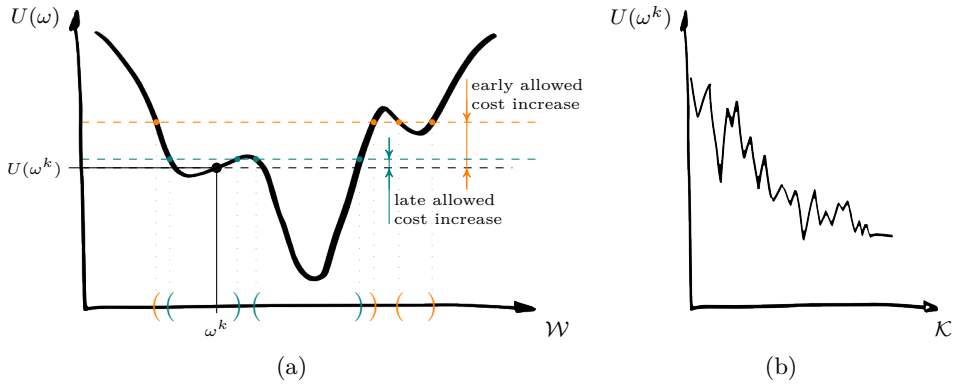


Figure 3.28: Simulated Annealing (SA) behavior. (a) Eligible state evolution for a particular configuration at different temperature stages. (b) Cost evolution.

Graph-Cut (GC)

The use of graphs in order to mathematicaly formulate or solve certain problems dates from Euler's 1736 paper on the bridges of Königsberg, despite the fact that there is no mention of graphs in this early paper [156]. But since then, *Graph Theory* has provided a wide range of techniques and strategies to solve problems and is particularly useful for problems with pairwise relationships within the elements. The main problem consists of how to represent a particular problem as a graph and determine which graph particularity or property corresponds to the goal pursued.

Boykov et al. [104] introduced the use of graph-theory in order to solve the metric labeling problem (see eq. 3.6). The algorithms introduced were the *swap-move* and the *expansion-move*. Both algorithms repeatedly compute the global minimum of a binary labeling problem in their inner loops, rapidly converging in a strong local minima guaranteeing that no labeling with lower energy can be found [149].

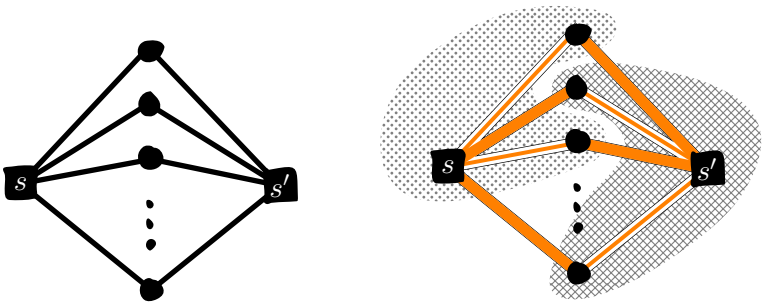
Details of the exact functioning of the *swap-move* and *expansion-move* GC algorithms are not covered here. The reader is refereed to the energy minimization comparative study carried out by Szeliski et al. [149] for a consise description, to Boykov et al. [104] to review the original proposal or to Delong et al. [105] for further work carried out by the same team that proposed the methods.

However, it is in our interest to understand how the metric labeling prob-

lem posted in equation 3.6 is represented as a graph and which graph solution or property leads to the estimated labeling $\hat{\omega}$, so that $\hat{\omega} = \arg\min_{\omega} U(\omega)$.

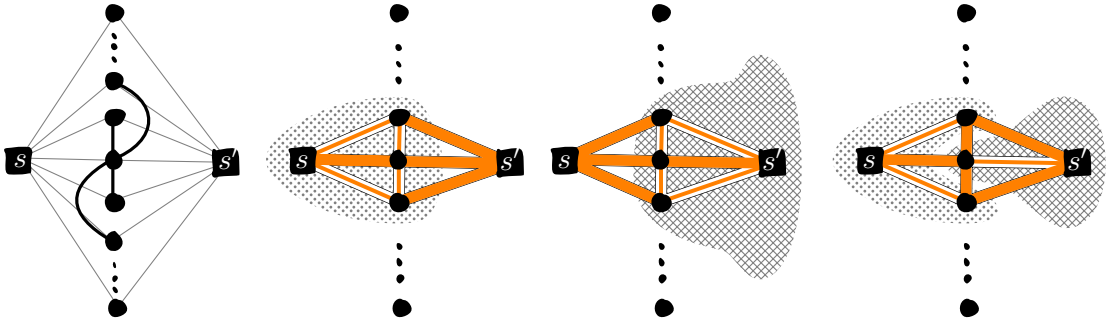
Let's consider a binary case of the type segmenting foreground vs. background ($\mathcal{L} = \{f, b\}$) and only the data term is considered ($V_{s,r} = 0 \,|\forall\{s, r\} \in \mathcal{S}$). Also let the data term be defined as the posterior probability of a Bayesian procedure $D_s(\omega_s = f) = P(f|\bar{x_s}), D_s(\omega_s = b) = P(b|\bar{x_s})$, so that $P(f|\bar{x_s}) + P(b|\bar{x_s}) = 1$, where $\bar{x_s}$ represents the data describing the site s. Figure 3.29a illustrates how the graph is constructed. The sites are represented as nodes in the graph and two extra nodes, illustrated as a squares, denoting source (s) and sink (s') are added. (In some works the sink is denoted as t.) s and s' are each assigned to one of the possible labels; foreground or background. Usually, s is assigned to the foreground and s' to the background, but this is irrelevant, since the solution is the same. Source is connected to each of the nodes in $\mathcal{S}$ with the associated cost $P(f|\bar{x_s})$. Similarly, the sink is also connected to all the nodes with the cost $P(f|\bar{x_s})$. In order to find the labeling $\hat{\omega}$, the *min-cut/max-flow* technique consisting of passing the maximum amount of flow from s to s' is used. For the case considered here, where only the data term is taken into account, the maximum amount of flow corresponds to $\sum_s \min(P(f|\bar{x_s}), P(b|\bar{x_s}))$ since the amount of flow passing through every node is limited by the weakest posterior probability. The saturated edges conditioning the maximum amount possible of flow correspond to those edges needed to partition the graph into two sets with a minimal cost. Once the graph has been partitioned, the nodes still connected to the source are labeled as foreground, while those connected to the sink are labeled as background (see fig. 3.29b). Notice for the case with no pairwise term, the output of the graph-cut corresponds to assigning the label producing the greatest posterior probability.

When adding the pairwise term, connections between sites are made, as can be seen in figure 3.30. This allows flow transfer between the sites so that connections between the sites and the sink (s') that previously were not saturated can now be, if the flow needed to saturate the link to the sink is provided by the connected sites. Notice that the amount of flow is limited by the strength between the source and the sites. If there is not enough flow to saturate the link to the sink, then the link to the source is saturated, unless the links between the sites have already been saturated. If the link to the source is saturated, then the sites are labeled with the label associated to the sink. If the connections between sites are not big enough, they are easily saturated and the solution is the same as the solution minimizing the data term.



(a) (b)

Figure 3.29: Data term graph construction to solve the data part of the labeling problem using *min-cut/max-flow*. (a) Graph construction. (b) Data term solution.



(a) (b) (c) (d)

Figure 3.30: Data and pairwise terms graph construction to solve the complete labeling problem using *min-cut/max-flow*. (a) Graph construction. (b-d) Multiple configurations leading to different solutions.

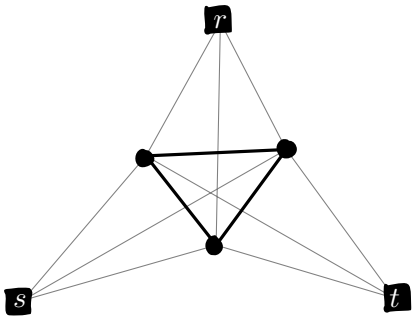


Figure 3.31: Multi-class graph construction example using three sites example.

For the multi-label case, sites are all connected to a node representing each label, as can be seen in figure 3.31. All the possible pairing combinations of labels are assigned as source and sink. The system is flooded repeatedly for each combination pair labels until convergence. Figure 3.31 represents an intuitive idea of a system with three sites and three labels. The links from the labels are colored differently from the links between the sites for better comprehension.

3.3.8 Post-processing

Post-processing is a common practice in breast lesion segmentation in the ultrasound image application. Reducing the amount of false positive segmentations or refining the delineation for a more accurate segmentation can be found in common practices. ML procedures are a common choice to reduce false positive segmentations used as the outlier rejection step [157]. Another common practice is to impose strong criteria from the application domain so that the lesions do not touch the border of the image, to eliminate undesired segmentations [78], [116]. In order to improve the lesion delineation, a common strategy is to use ACM to obtain smooth segmentations [60].

For this work, no post-processing is applied. The outlier rejection is imposed as homogeneity in the pairwise term, application domain constraints are enforced in the data term and the accuracy achieved in the delineation is given by the superpixel ability to attach to the true tissue interface.

3.4 Case of Study

3.4.1 Gathered dataset

The dataset used for this work comes from the collaboration between the *University of Girona* and the *UDIAT Diagnostic Centre of Parc Taulí* in Sabadell (Catalunya) where an image database from UDIAT is being collected and cataloged in order to make it available to the researchers. The collected database consists of a collection of screenings including DM, US, or both, which exceeds the 2300 images.

The resulting US image dataset once discarded US images with burned in overlays coming from the acquisition system consists of 700 B-mode US images screened using the following devices:

- Siemens ACUSON$^{\text{TM}}$ Sequoia equipped with the linear transducer 17L5 HD)$(17-5MHz)$.
- Siemens ACUSON$^{\text{TM}}$ S2000$^{\text{TM}}$ equipped with the linear transducer 18L6 HD $(18-6MHz)$.
- Siemens ACUSON$^{\text{TM}}$ Antares® equipped with the linear transducer VF13-5 $(13-5MHz)$.
- SonoSite® MicroMaxx®.
- SonoSite® Titan®.
- Supersonic-Imagine Aixplorer®.
- Toshiba PowerVision SSA-380A.
- Toshiba Aplio$^{\text{TM}}$ 500.

The overall dataset is composed of Digital Imaging and Communications in Medicine (DICOM) formated images with anonymized metadata-information and an heterogeneous accompanying GT. All the images have, at least, one lesion and a delineation of the lesion structures of each image, provided by an experienced radiologist from UDIAT. Figure 3.32a illustrates the different sub-datasets that appear depending on the GT provided:

276 image dataset constitutes the original dataset from clinical cases acquired by the doctor radiologists at UDIAT. All the images contain a single lesion which their delineation and pathology description and lesion delineation has been provided by radiologists.

150 image dataset consisting of a subset of the previous dataset. Each image is linked with the associated BI-RADS description. This subset has been used to illustrate the BI-RADS image description in section 1.3.3.

115 image dataset consisting of a subset of the 150 images dataset. Each image is linked with seven different manual delineations of lesion carried out by trained experts and technicians from *University of Girona* and the *UDIAT Diagnositc Centre*. All the delineations have been validated by two doctor radiologists with a dilated experience.

700 image dataset with a subset of 424 images pulled out from patients history, complements the original 276 image dataset constituting an entire dataset of 700 images. The advantage of complementing the original database in this way allows to perform further temporal studies and increases the variability in the quality of the images. All the images have been reviewed by a doctor radiologist from UDIAT in order to ensure that there is at least one lesion per image and also to provide an accurate delineation of any lesion present in each image.

16 image dataset it is composed by a randomly sampled subset of the entire dataset for software developing purposes. Some images from this subset are provided with the delineation of all the tissues present in the image, for training purposes. These tissue multi-label delineations have been carried out by a technician an validated by doctor radiologist with dilated experience, members from UDIAT.

In terms of pathology, our data is distributed accordingly to figure 3.32b, where the are of each rectangle represents the amount of lesions presenting a particular pathology. The collected pathologies are distributed as follows:

Benign: 90 Cysts, 69 Fibroadenomas, 8 Ganglions, 6 Hermatomas, and 10 benign lesions categorized as other, with pathologies like: Papillomas, Lipomas, fat necrosis, etc.

Malignant: 67 Ductal Infiltrating Carcinoma (DIC), 12 Infiltrating Lobular Carcinoma (ILC), 7 Intra-Ductal Carcinoma (IDC), and 8 malignant lesions categorized as other, with pathologies like: Mucinous Carcinoma, Lifoma, etc.

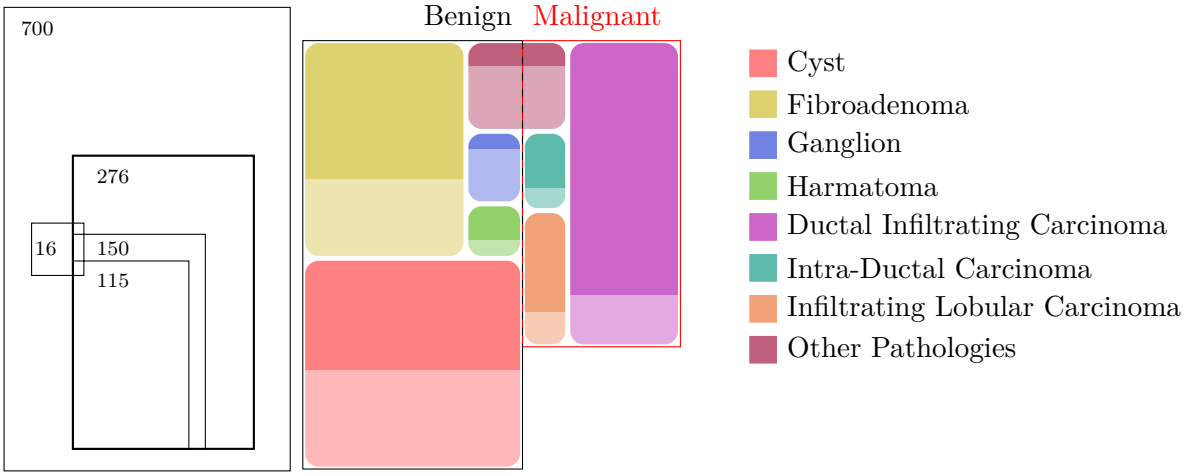


Figure 3.32: B-mode breast US image dataset collection. (a) Represents the datasets collected from an undergoing collaboration between the *University of Girona* and the *UDIAT Diagnositc Centre of Parc Taulí* in Sabadell (Catalunya). In (a) the data is grouped in terms of the GT available. (b) Represents the pathology distribution from the 276 image subset. The highlighted areas in (b) represent the amount of elements of each class forming the 115 images subset.

3.4.2 Experimentation and results

The experimentation has been carried out using the smallest dataset in order to keep the computational cost bounded while exploring to combine all the proposed features. The tested system features are:

- Superpixel type:
  - Quick-Shift (QS) superpixel.
  - Global Probability Boundary (gPb) superpixel.
- Feature description:
  - Superpixel brightness:
    - using mean as superpixel descriptor, B_{μ}.
    - using median as superpixel descriptor, B_{Md}.
    - BoW-SIFT
  - Superpixel overall appearance distance to the appearence of the tissue models.
  - Superpixel BoW representation from a 36 words dictionary of SIFT.
  - Atlas information.
  - Superpixel multi-resolution feature description:
    - Brightness, B_{μ}.
    - Brightness, B_{Md}.
    - BoW-SIFT
- Data model generation using SVM with a RBF kernel in order to compute the MAP.
- Pairwise modelling:
  - No model.
  - Smoothing, homogeneity model set as 10% of the overall data cost[1].

[1]The cost of every and each pairwise link is set as 10% of the total data term for a site, consisting of the cost from the source to the site plus the cost from the site to the sink, which is constant for all the sites.

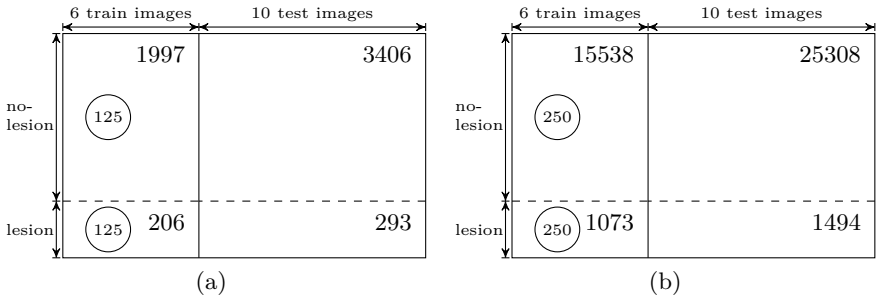


Figure 3.33: Randomized sampling for classifier training from the pool of superpixels. The values represent the amount of superpixels and the circle illustrate the random selection for a training round. The supperpixels' pool correspond to case of: (a) Quick-Shift (QS), (b) Global Probability Boundary (gPb).

Quantitative results

From figure 3.34 to 3.39 it is shown the quantitative results obtained by applying the *cross-validation* procedure. The classic cross-validation, when analyzing image segmentation, consists of splitting the data into K-image subsets named *folds* and use $K-1$ folds to train the system for testing with the fold unused in the training. The extreme case of cross-validation is known as Leave-One-Out Cross-Validation (LOOCV), where every and each image is considered a fold by itself, so that all the images but one are used to train the system for further testing in this one. The training/testing is repeated k-times, so that every image is used once as testing. However, in our case and taking advantage of the fact that the system is based in superpixels, the cross-validation is carried out at the level of superpixels. In our system, all the superpixels are seen as samples of a larger dataset and the cross-validation is carried out as a as multiple randomized sampling of such a pool of superpixels. Since the size of the pool is orders of magnitude larger than the size of the training set, it is unlikely to over-fit the classifier. Figure 3.33 illustrates this idea. Although all the superpixels belonging to the training images have been used to generate the models needed for computing the features, in the figure it can be observed how only two lesion/non-lesion balanced subsets are used for a training round of the classifier. At each round all the images (and hence all the superpixels) are used for testing. All the testing results are collected together as independent instances to conduct the results analysis.

Figure 3.34 shows a quantitative general analysis of the system. Each

and every one of the boxplots present in figure 3.34 represents an experiment where the superpixels and the pairwise model vary depending upon the experiment (the experiment details are summarized in table 3.2). The elements building up the boxplots are the mean Area Overlap (AOV) (in fig. 3.34a), the mean False Positive (FP) rate (in fig. 3.34b) , and the mean False Negative (FN) rate (in fig. 3.34c) achieved across the entire dataset for a particular configuration of the features used to describe the images. Details about lesion detection are beyond the scope of this work, further than to illustrate the improvement that supposes incorporating the pairwise term, so then justifying the need of such a pairwise or smoothing term (see fig. 3.34b). In figure 3.34a some AOV references are represented. Since the segmentation is achieved by labeling superpixels, the final delineation is subject to the underlying superpixels' delineation therefore an AOV of 1 cannot be achieved and the ceiling for each superpixel is represented in the figure. Notice that the ceiling for gPb superpixel is higher than the ceiling for QS superpixel, which is explained by the fact that gPb superpixels are smaller than QS superpixels. In addition to the superpixels' AOV ceiling, the AOV reward achieved by manual segmentations done by trained technicians and expert radiologists is also represented by a swatch compressing the interval between best and lowest performance, as reported in [66]. Finally, for comparison purposes, the AOV results reported by Massich et al. [61] and Pons et al. [66] on their respective proposals are displayed, since the subsets used to test both methodologies come from the same dataset. Due to the difference in the size of the superpixels, two FP rate references are needed in figure 3.34b.

In order to facilitate the comparison between the proposed methodology and the methodologies reviewed in chapter 2, despite the bias of being tested in different datasets, the figure 2.8a is replicated here in figure 3.35 this time showing an extra ring in black at 0.623 representing the best performance in fig. 3.34a, so that it can be easily compared to the previously reviewed methodologies.

All the boxplot pairs in figure 3.34a and figure 3.34b correspond to the same experiment with and without applying the smoothing term, so that odd elements within these figures (1,3,5,7) represent the data model results with no pairwise model applied whereas even elements (2,4,6,8) represent the results once the pairwise model is applied. Notice that in most cases, although there is no statistical difference in terms of segmentation performance when applying the smoothing term, a tiny decrease of the results can be observed even if this is not enough to be considered statistically different. However the improvement in terms of FP reduction justifies to assume the

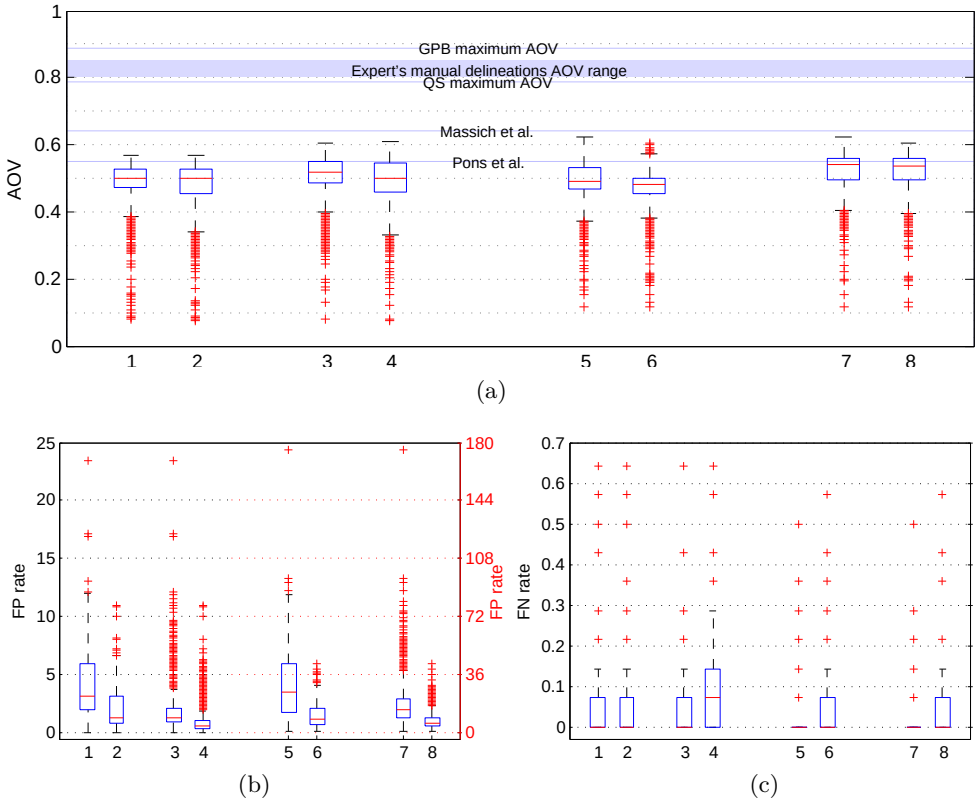


Figure 3.34: Quantitative results. (a) Area Overlap (AOV) distribution depending on the feature descriptions configuration of the system for the set of experiments described in table 3.2. (b) dataset average False Positive (FP) rate distribution. (c) dataset average False Negative (FN) rate distribution.

Table 3.2: Configuration details of the experiments

	Superpixel		Regular features	Multi-resolution feat. [1,2,3]			Pairwise term
	QS	gPb		B_{Md}	B_{μ}	BoW+SIFT	
1	✓		✓	✓	✓		
2	✓		✓	✓	✓		✓
3	✓		✓		✓	✓	
4	✓		✓		✓	✓	✓
5		✓	✓	✓	✓		
6		✓	✓	✓	✓		✓
7		✓	✓		✓	✓	
8		✓	✓		✓	✓	✓

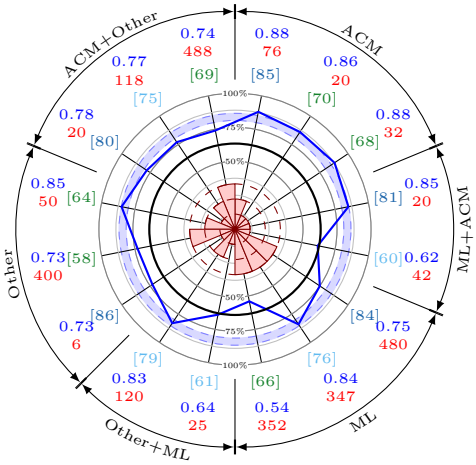


Figure 3.35: AOV comparison between our proposal and the methodologies reported in section 2.4. The figure replicates fig. 2.8a adding a circle representing the top AOV performance achieved 0.6231 to facilitate the comparison against all methods despite the bias of not being tested on the same dataset.

decrease in AOV terms, in favor of a much lower FP rates when combining both the data and the smoothing terms. As downside, encouraging homogeneity, by combining these two terms, also increases the FN rate. However both images still show, represented as gray elements in the figures, a large amount of feature configurations with no FN. The incorporation the multi-resolution feature of BoW-SIFT also produce a FP rate reduction, as can be observed in figure 3.34b. Finally, despite not being a substantial improvement, the configuration best scoring in 3 improves its AOV reward from 0.607 to 0.61 achieving the best performance for the QS superpixel in that particular configuration.

In terms of superpixels, gPb produce better results than QS but it also needs to be taken into account that the AOV ceiling of both superpixel types differ whereas the performance difference is not as large as the ceiling differences.

Figures 3.36 and 3.37 show some screen-shots of the software we use to qualitatively explore the quantitative results achieved for different feature combinations. Both figures represent the experiments where the tested features are: B_μ , B_{Md}, overall appearance, BoW+SIFT, atlas, multi-resolution B_μ and multi-resolution BoW-SIFT. Figure 3.36 represents the case of QS superpixels , which corresponds to experiments 3 and 4, and figure 3.37 represents the case of gPb superpixels, which corresponds to experiments 7 and 8. In the figures, pairs (a,b) correspond to AOV, (c,d) correspond to FP rate, and (e,f) correspond to FN rate. In addition the mentioned pairs, the triplets (a,c,e) in the figures correspond to the experiments 3 and 7 respectively where no pairwise cost is applied whereas the (b,d,f) tripled correspond tho the experiments 4 and 8 respectively where pairwise cost is applied. Every pixel within the figure represents the average reward obtained across the dataset for a particular feature combination. The average value is color coded. For the AOV values the color code is such that an AOV of 0 is represented in blue while an AOV of 1 is represented in red. For the rest of the figure every pair hare the same scale where strict 0 is represented as gray and the rest of the values are coded in a linear manner where the lower bound is represented as blue and the upper bound is represented in red (the limits can be found on the figure captions). On a general view, both figures show some repetitive patterns corresponding to the usage or not of certain features. On the figures it can be observed that when applying the pairwise cost, despite the performance of some configurations do not decrease, the general tendency is that most feature configurations experiment a reduction of the AOV obtained when introducing the pairwise cost. In figure 3.37 the effect of the AOV reduction when applying the pairwise

cost produce a regular patter showing that certain configurations are more affected than others since the color patterns present in (b) are not that clear in (a). Those configurations that get more affected by pairwise term correspond to configurations were a limited amount of features are used. Similar conclusions can be drawn for the (b,c) and (e,f) pairs, which also corroborate the conclusions drawn from figure 3.34. The reduction of the FP is general when applying the pairwise cost where still exist configurations with no FN.

In order to perform a more guided discussion, similar information is represented in figures 3.38 and 3.39 where instead of representing the experiments as a colored table, the experiments are placed in a disc where a binary code illustrates the presence or not of a particular feature, and the obtained results are displayed as a polar plot. From the inner to the outer part, the displayed information is organized as follow: the most inner part in red represents the amount of FP segmentations in a logarithmic way in order to obtain better resolution for cases with low FP rates. Notice that the amount of false positives is given as the average across the entire dataset and it needs to be taken into account that when an image suffers from FP segmentations usually there is more than one FP. The second polar plot, in blue, represents the AOV reward for every particular system configuration. In gray, follows a binary coding illustrating which feature descriptors are active at every time. It needs to be mention that the two less significant bits have been merged for displaying purposes. Finally, the plot offers a degree wheel for easy reference to a particular configuration.

In terms of features, in general, incorporating BoW-SIFT multi-resolution increases the performance in terms of AOV and, as aforesaid, reduces the FP rate. In order to be able to further look to the influence of each feature to the overall performance of the system, the reader is referred to figures 3.36 to 3.39. Figure 3.36 and 3.37 show a set of colored tables where every cell represents a system configuration. Figure 3.36 compares experiments 3 and 4 whereas figure 3.37 compares experiments 7 and 8. Both figures show, at a glance, that some patterns arise in the AOV performance which are related to the features used to describe the superpixels showing the preference for some description configuration. While comparing the experiments with and without smoothing term a small change in color can also be observed as expected. However, what is interesting, is to observe patterns produced by the system configuration regarding the reduction of the FP rate and the increasing of the FN rate.

Once familiarized with the overall behavior of the system depending on its feature description configuration, a deeper analysis can be carried out in figures 3.38 and 3.39 where the AOV, FP and FN rates are displayed at

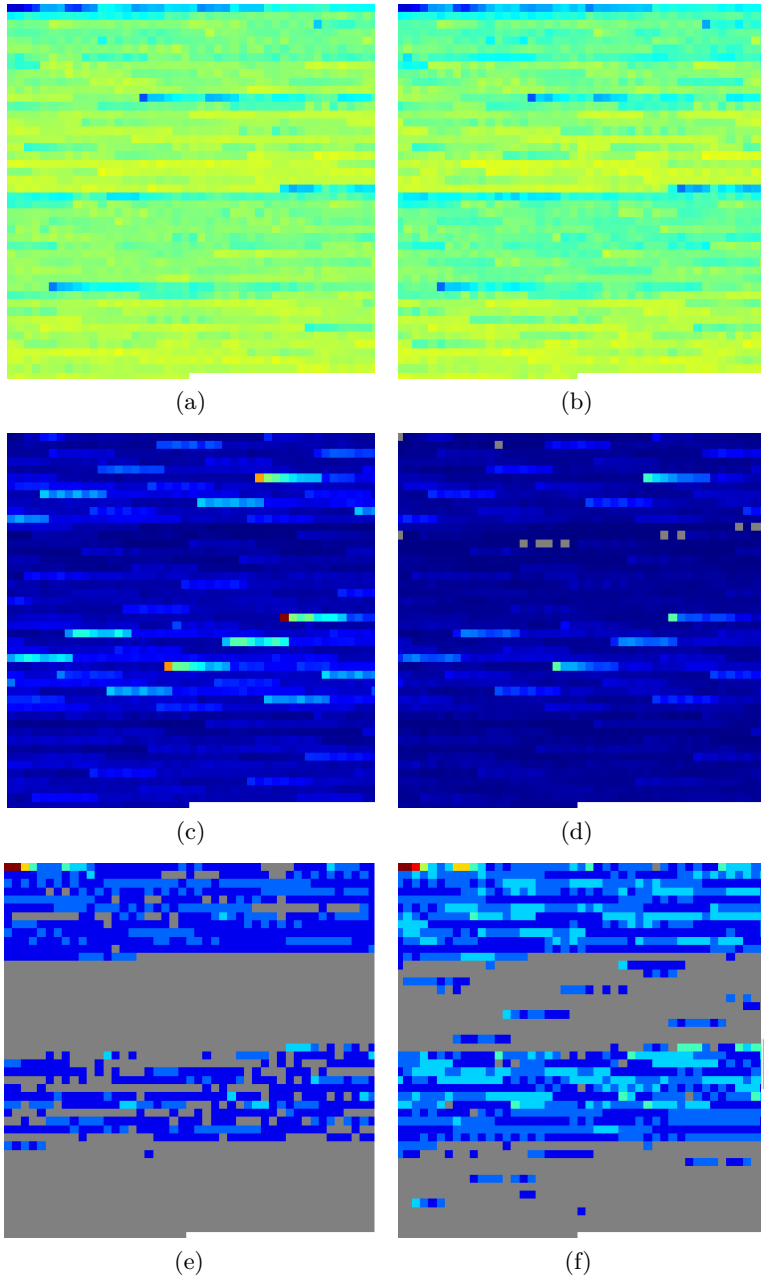


Figure 3.36: Qualitative inspection of the quantitative results obtained by different feature descriptors combination. (a,b) Represent the AOV where the scale 0 to 1 is represented from blue to red. (c,d) Represent the FP rate where dark blue corresponds to nearly 0 values and red corresponds to an average FP rate of 23.4 FP per image. (e,f) Represent the FN rate where dark blue corresponds to nearly 0 values and red corresponds to an average FP rate of 0.6 FP per image. (a,c,e) correspond to experiment 3 and (b,d,f) to experiment 4.

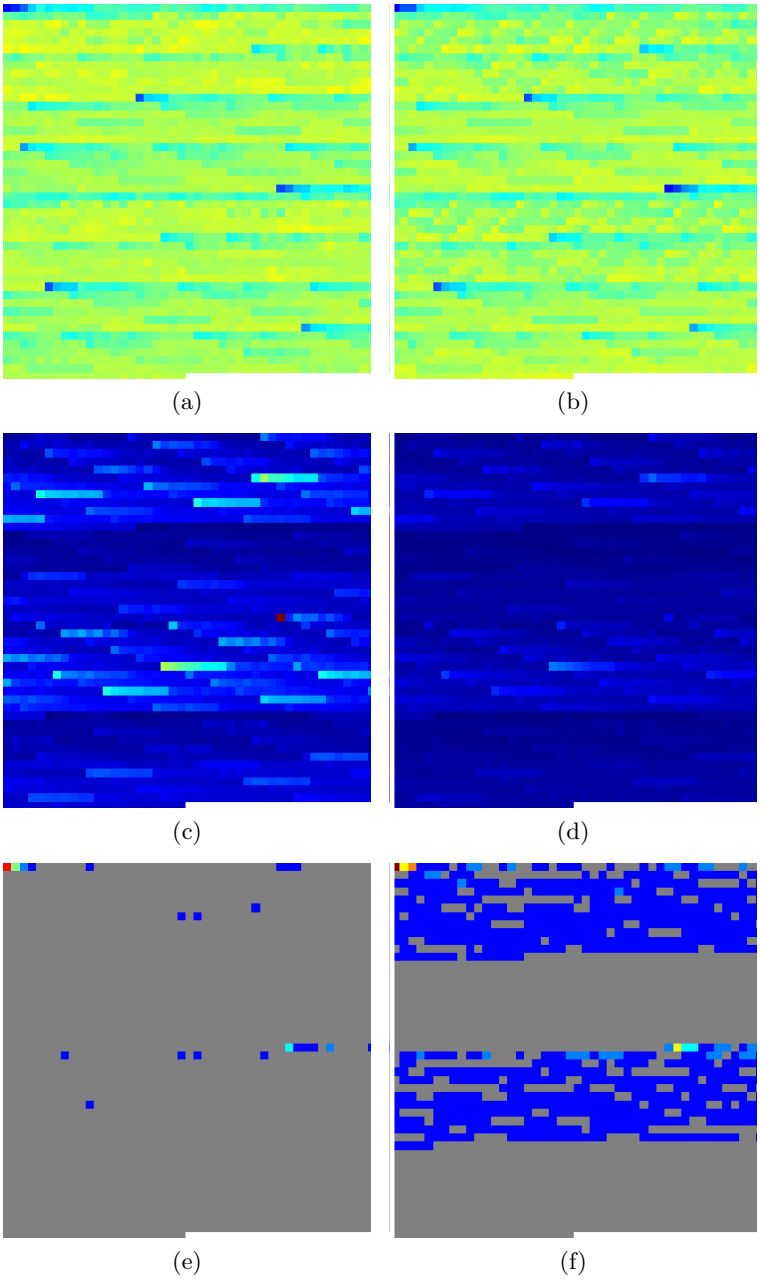


Figure 3.37: Qualitative inspection of the quantitative results obtained by different feature descriptors combination. (a,b) Represent the AOV where the scale 0 to 1 is represented from blue to red. (c,d) Represent the FP rate where dark blue corresponds to nearly 0 values and red corresponds to an average FP rate of 175.6 FP per image. (e,f) Represent the FN rate where dark blue corresponds to nearly 0 values and red corresponds to an average FP rate of 0.5 FP per image. (a,c,e) correspond to experiment 7 and (b,d,f) to experiment 8.

once, along with the system configuration.

Figure 3.38 shows the details of the boxplot in fig. 3.34a and 3.34b for the experiment 4. In it can be observed that the recommendable feature description configuration lies between angles 90° to 167° and 270° to 315° which produce higher results with less FP rates. In the figure is also noticeable the AOV drop at every 90° which is explained by the low amount of descriptors. The fact that the glitches at 0° and 180° are smaller compared to those in 90° and 270° is due to the usage of the atlas feature. The usage of the atlas feature, also explains the increase in AOV and the reduction of FP for the quarters from 90° to 180°, and from 270° to 365°. The atlas feature also gives an overall stability of the results since the AOV plot shows less jitter in the quarters where atlas has been used. This reaffirms the usefulness of the position information. Similar configurations with a reduced set of features such in 45° and 225° despite not producing large drop in terms of AOV the rise of the FP is quite substantial. Again, the presence of position information from the atlas contains the spike in 135° and 315° but a small increase in the FP rate can still be observed. In such FP peaks at 135° and 315°, a decrease of the AOV rate can also be observed which repeats in some of the other peaks of the FP plot discouraging, even more, to use configurations that produce high amount of FP. The AOV plot offers a crescendo tendency at every quarter, which can also be observed at the overall AOV plot if read counterclockwise from 90°, indicating that the feature descriptors designed properly capture the lesions.

Similar conclusions can be drawn from figure 3.39 which replicates the same experiment but this time using gPb superpixels. The influence of the atlas is also clear specially the fact that atlas produce more stable results since the high jitter is present in the AOV present in the configurations compressed from 0° to 90°, and from 180° to 270° where the atlas feature is not used. A difference between 3.38 and 3.39 is that now the presence of less features which happens at every 45° is more noticeable. The AOV drops close to angles 326°, 331°, 338°, and 342° are subject to not using neither of the BoW+SIFT multi-resolution features at neighboring level of 2 and 3. This pattern, despite not being as clear as in such examples, can be found all over the disc which indicates that at least for capturing the texture larger superpixels are recommendable. Specially since the gain of using BoW+SIFT (regions $0-45°$, $90-135°$, $180-225°$, and $270-315°$) is minimal compared to their counterparts shifted 45° but the increase of FP is notable. For this particular experiments configurations within the ranges between 135° to 165°, and 315° to 345° would be preferred.

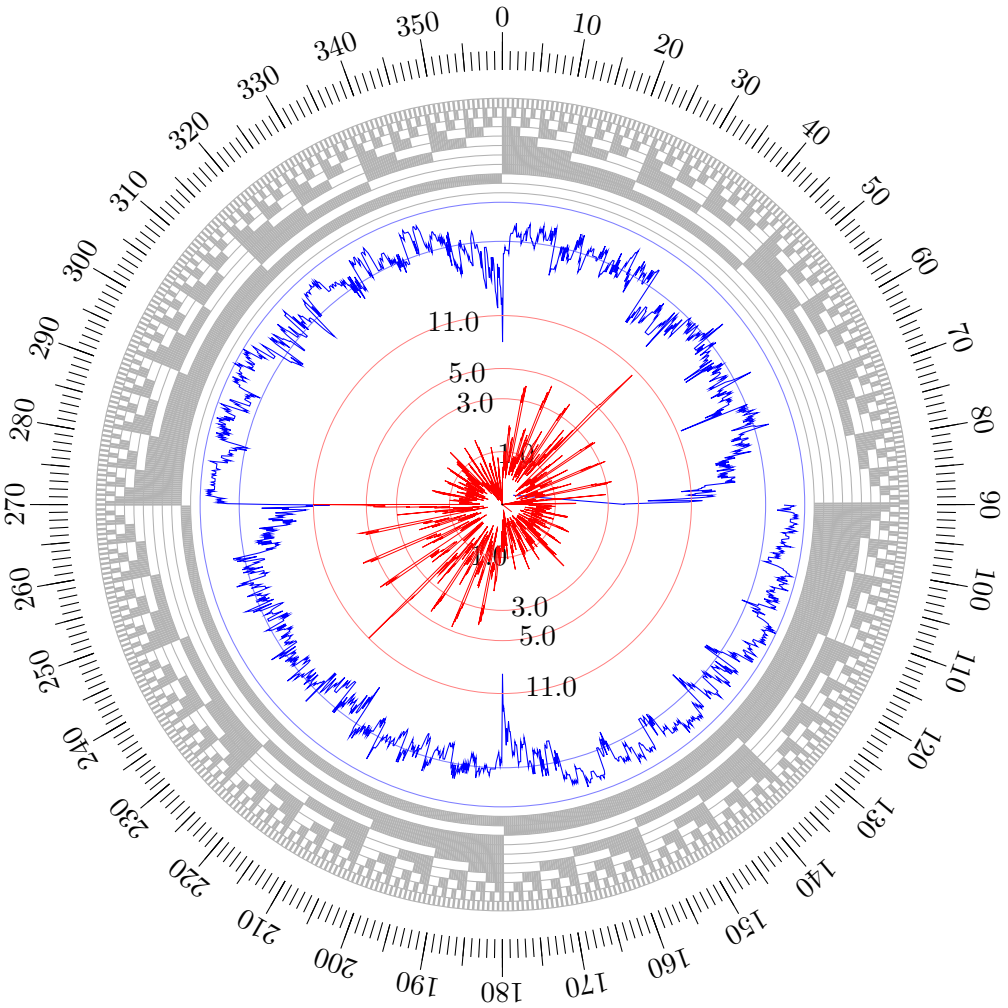


Figure 3.38: Experiment 4 detailed results where each angle represents a particular configuration. From inner to outer part: False Positive (FP) rate (in red), AOV rate (in blue), active feature swatch (gray when active), and degree wheel for rapid referencing.

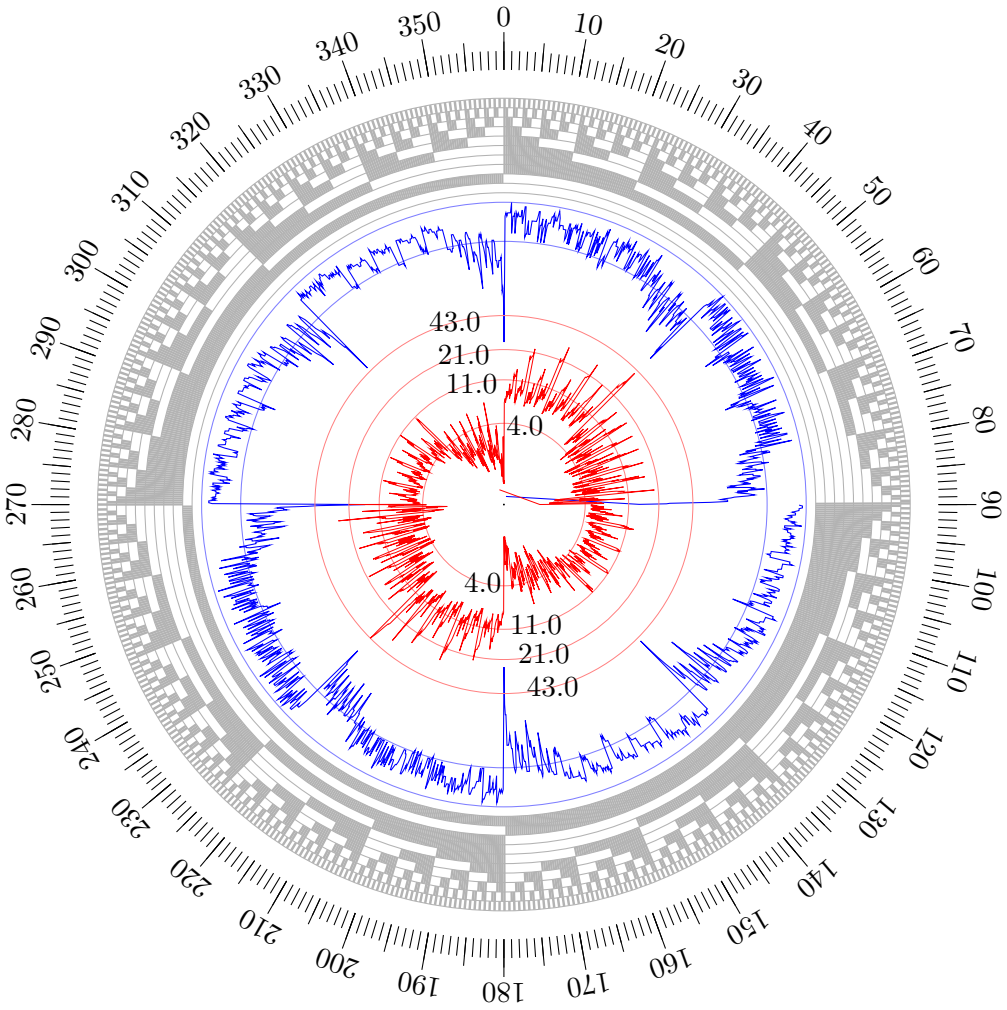
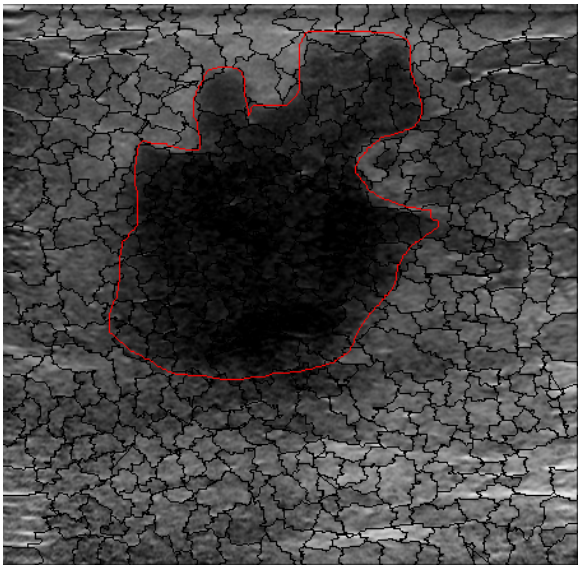


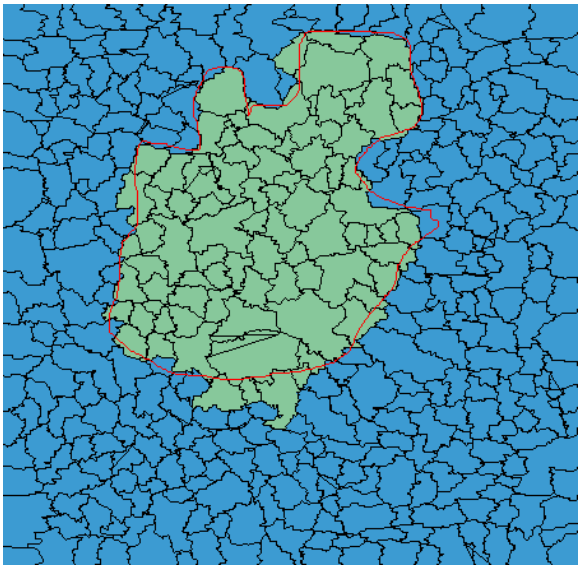
Figure 3.39: Experiment 8 detailed results where each angle represents a particular configuration. From inner to outer part: False Positive (FP) rate (in red), AOV rate (in blue), active feature swatch (gray when active), and degree wheel for rapid referencing.

Qualitative results

Figures 3.40 to 3.42 show few qualitative results complementing the quantitative results shown already. Apart from the discussion regarding if the three lowest superpixels labeled as lesion should belong or not to the lesion in figure 3.40, the figure shows the dissimilarities between a perfect segmentation produced by the system and the GT, which leads to the performance ceilings represented in figure 3.34a. Figure 3.41 illustrates the effect of applying homogeneity in the pairwise term for reducing the FP rate. Figure 3.42 illustrates the particular cases of FN cases where the lesion cannot be properly represented due to the fact that the superpixels are larger than the lesion. In such cases, there is no way to properly characterize the superpixel containing the lesion and therefore missclassification is inevitable. In such a cases when the superpixel containing the lesion is labeled as such, a large amount of FP is present (see fig. 3.42c) otherwise the lesion is missed like in figure 3.42d.

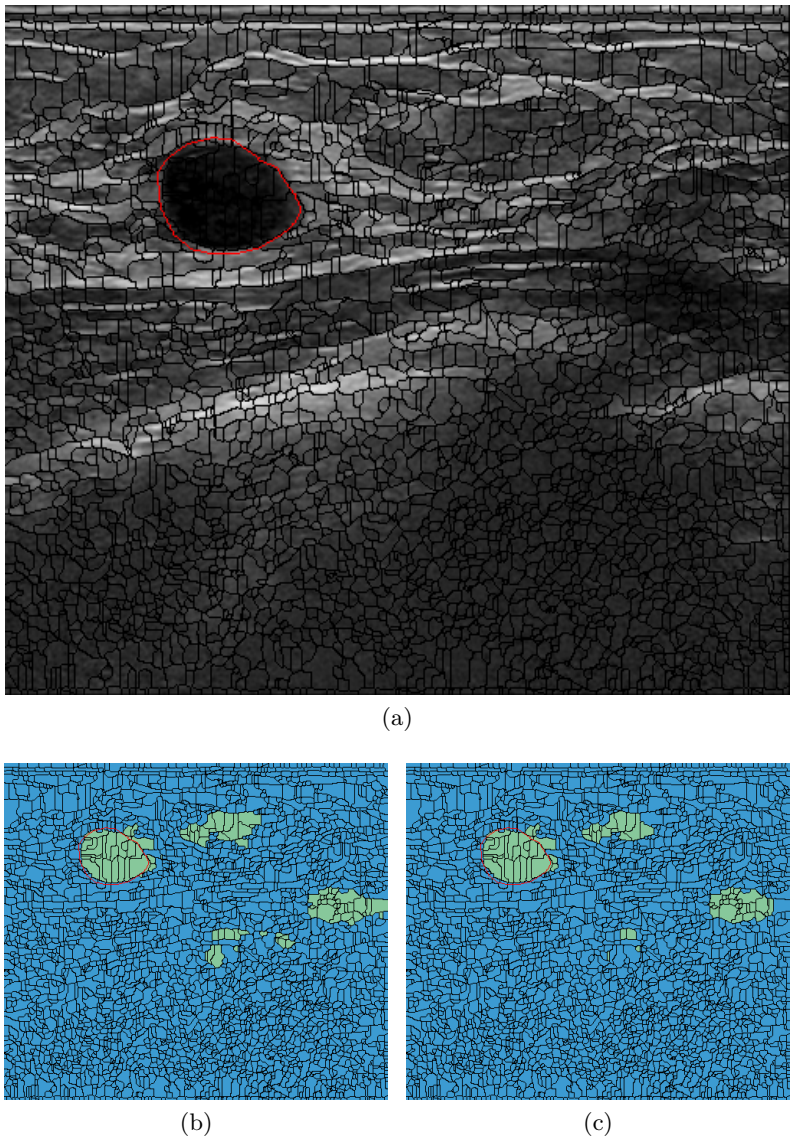


(a)



(b)

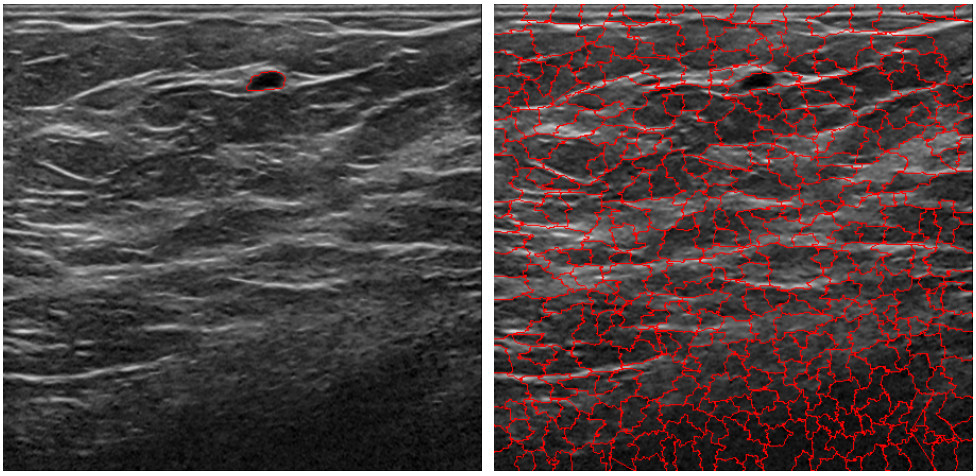
Figure 3.40: Qualitative result example from experiment 4. (a) Original image with GT overlay in red and superpixels' boundaries overlay in black. (b) Segmentation obtained using: Appearance model, Atlas, BoW+SIFT multi-resolution (1,2,3) and mean Brightness multi-resolution (1,2,3) features. This feature set is close to 163° in fig. 3.38.



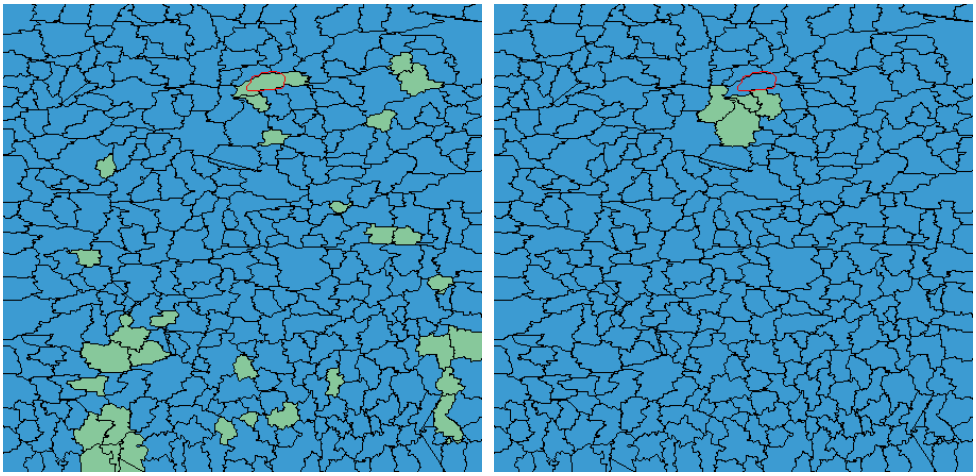
(a)

(b) (c)

Figure 3.41: Qualitative result example from experiment 7 and 8 to illustrate the effect of the homogeneous pairwise term. (a) Original image with GT overlay in red and superpixels' boundaries overlay in black. (b) Segmentation obtained without using the pairwise term. (c) Segmentation obtained with pairwise term. The Segmentations in (b) and (c) are obtained using: Atlas, mean Brightness, median Brightness, BoW+SIFT multi-resolution (1,2,3) and mean Brightness multi-resolution (1) features. This feature set is close to 315.5° in fig. 3.39.



(a) (b)



(c) (d)

Figure 3.42: Qualitative result from experiment 3 and 4. (a) Original image with the GT overlay in red. (b) Original image with superpixels' boundaries overlay in red. (c) segmentation obtained from experiment 3. (d) segmentation obtained from experiment 4.

Chapter 4

Conclusions and further work

On ne termine pas un poème, on l'abandonne

Paul Valéry

Automatic analysis of US images is challenging specially in the case of breast US imaging. However, it is of our interest to address such a problematic in order to automatize massive screening since it has been proved that early detection decreases the breast cancer mortality which remains the leading cause of cancer death among females.

First chapter introduces the imaging modalities used in breast screening placing special emphasis in US screening of the breast. This chapter is used to familiarize the reader with the US images of the breast spotting its strengths for diagnosis purposes, as well as, describing its limitations and reading difficulties such as its strong noise and artifacts. This introductory chapter is also used to analyze standardized procedures that radiologist doctors use to carry out the image readings. Such standard procedures for analyzing US breast images rely in the fact than when read by an expert radiologist the delineation of the lesion is instantly understood. Therefore the need to improve automatic procedures for accurately delineate lesions to be able to extract high level features similar to those already proven to be useful for the doctors in order to improve CAD systems.

Second chapter is devoted to survey the state-of-the-art in segmentation of breast lesions in US data. This chapter reviews the methodologies along with the manner their results are reported in order to set them all in a common framework for comparison. However, the lack of a common dataset to

test all the methodologies with makes impossible a fair comparison between methods. This is easily observed when comparing the segmentation results reported from automatic methodologies those outperform manual segmentations done by trained expert radiologists.

Finally, in chapter three the bulk of work carried out is reported. In it, a novel segmentation scheme based on an optimization framework is presented, and the methodology based in GCS presented in [61] is reviewed. Although the new segmentation technique achieves results are comparable only to some of the results published in the bibliography (see fig. 3.35), the proposed methodology has large room for improvement compared to our previous proposal which was pretty tuned up already (see section 3.2). In this last affirmation it needs to be taken into account such methodologies against the rest of the methodologies in the literature is unfeasible due to the lacking common dataset.

The main advantage of the proposed framework is that it splits the problem of segmenting the tissues present in US the images into subtasks that can be taken care of individually. The correctness of the final delineation is relative to the correctness achieved during the partitioning of the image while generating the superpixels. The characterization and proper labeling of the superpixels with the desired tissue label becomes a ML problem that can take advantage of the large range of solutions in that field. Finally the obtained results from the classification stage can be improved by elaborated pairwise cost functions allowing an inhomogeneous severe smoothing.

As a summary, the main contribution is not in term of results yet but in facilitating a framework that splits the segmentation task in concrete subtasks allowing easily testing them. Another contribution is the collection of dataset of US images with large amount of annotation information regarding both medical and image information to put in use for the scientific community.

4.1 Short term perspective

In short term perspective, the system is set and ready to be tested in larger datasets of the data gathered and cataloged from the collaboration between the *University of Girona* and the *UDIAT Diagnositc Centre of Parc Taulí*. Data that at the same time are ready to be published to make it available to the scientific community in order to take advantage of data and the GT that has been already collected in addition to the GT which is being completed under the undergoing collaboration between the institutions mentioned.

Regarding the proposed framework, short term efforts should be placed in improving the the data term where feature extraction procedures can be applied at any time. However, before that, there exists still room to improve the current feature descriptors of the images. Here are some ideas to improve each and every one of the features reported:

Brightness feature A naive segmentation, as a presegmentations, of the tissue can be made in order to condition the brightness reference only to the part anterior to the chest-wall, or have multiple references rather than the statistics of the whole image.

Overall appearance Instead of creating superpixels' appearance models to compare with, based on the GT classes, a spatial clustering in order to obtain more sparse models arises as a plausible way to improve the results. However, it needs to be taken into account that if a large set of models is generated there exist a growth in the feature space which suggest that a feature extraction procedure such as PCA might be recommended.

BoW+SIFT During the assignation of the local feature SIFT to is closes visual word, and even when generating such visual words, these processes are done without taking into account the continuous nature of the visual data by using a hard quantization methodology such as k-means and nearest neighbor assignation. This is a known drawback an there are several solutions in order to reduce such quantization errors (see van Gemet et al. [158] for further details). New approaches are arising in order to improve BoW by including spatial information within the features and generate dictionaries based on both the features and their spatial relation [159].

Atlas Similarly to the improvement suggested for the brightness feature, a naive pre-segmentation of the lungs in order to modify the atlas feature based on lungs location. Another solution to explore is to perform registration of the atlas in order to build up a more reliable prior for driving the data term since it causes a huge influence on it when used.

Multi-resolution it is desirable to compare gathering the superpixels in a inheritance fashion with respect to the current multi-resolution based on neighboring.

Also, as a short term perspective, a larger set of configurations regarding the superpixels is recommendable in order to find the best parameters to describe the images in terms of superpixels.

4.1.1 Long term perspective

The encouraging of smoothness by the pairwise term has been demonstrated important, however it seems reasonable to apply it in a heterogeneous manner encouraging severe smoothing in some areas and no smoothing in other areas of the image in order to take a greater advantage of the pairwise term. Therefore it is recommended to explore more sophisticated ways to determine the cost of the link between superpixels. A plausible solution in order to do that is to apply a supervised ML procedure in order to determine the MAP cost of the link based on the GT information.

And last but not least, a system in order to efficiently explore the system configuration space, rather than the brute force applied here, is also desirable.

Bibliography

[1] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *International Journal of Cancer*, vol. 127, no. 12, pp. 2893–2917, 2010, ISSN: 1097-0215.

[2] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[3] R. A. Smith, D. Saslow, K. A. Sawyer, W. Burke, M. E. Costanza, W. Evans, R. S. Foster, E. Hendrick, H. J. Eyre, and S. Sener, "American cancer society guidelines for breast cancer screening: update 2003," *CA: a cancer journal for clinicians*, vol. 53, no. 3, pp. 141–169, 2003.

[4] W. A. Berg, L. Gutierrez, M. S. NessAiver, W. B. Carter, M. Bhargavan, R. S. Lewis, and O. B. Ioffe, "Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer," *Radiology*, vol. 233, no. 3, pp. 830–849, 2004.

[5] A. T. Stavros, D Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions," *Radiology*, vol. 196, no. 1, pp. 123–34, 1995.

[6] Y. Yuan, M. L. Giger, H. Li, N. Bhooshan, and C. A. Sennett, "Multimodality computer-aided breast cancer diagnosis with ffdm and dce-mri.," *Academic radiology*, vol. 17, no. 9, p. 1158, 2010.

[7] S Ciatto, M Rosselli del Turco, S Catarzi, D Morrone, *et al.*, "The contribution of ultrasonography to the differential diagnosis of breast cancer.," *Neoplasma*, vol. 41, no. 6, p. 341, 1994.

[8] P. B. Gordon and S. L. Goldenberg, "Malignant breast masses detected only by ultrasound. A retrospective review," *Cancer*, vol. 76, no. 4, pp. 626–630, 1995.

[9] D. Ensminger and F. B. Stulen, *Ultrasonics: Data, Equations, and Their Practical Uses*. CRC Press, 2008, p. 520.

[10] D. Manning, A Gale, and E. Krupinski, "Perception research in medical imaging," *British journal of radiology*, vol. 78, no. 932, pp. 683–685, 2005.

[11] E. B. Mendelson, W. A. Berg, and C. R. Merritt, "Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound," in *Seminars in roentgenology*, Elsevier, vol. 36, 2001, pp. 217–225.

[12] E. Mendelson, J. Baum, B. WA, *et al.*, *BI-RADS: Ultrasound, 1st edition in: D'Orsi CJ, Mendelson EB, Ikeda DM, et al: Breast Imaging Reporting and Data System: ACR BIRADS – Breast Imaging Atlas*. American College of Radiology, 2003.

[13] A. T. Stavros, *Breast ultrasound*. Lippincott Williams & Wilkins, 2004.

[14] J. Baker, P. Kornguth, M. S. Soo, R Walsh, and P Mengoni, "Sonography of solid breast lesions: observer variability of lesion description and assessment.," *AJR. American journal of roentgenology*, vol. 172, no. 6, pp. 1621–1625, 1999.

[15] D. G. Altman and J. M. Bland, "Statistics notes: diagnostic tests 2: predictive values," *Bmj*, vol. 309, no. 6947, p. 102, 1994.

[16] M. L. Giger, H.-P. Chan, and J. Boone, "Anniversary paper: History and status of CAD and quantitative image analysis: the role of medical physics and AAPM," *Medical physics*, vol. 35, no. 12, p. 5799, 2008.

[17] J Ferlay, P Autier, M Boniol, M Heanue, M Colombet, and P. Boyle, "Estimates of the cancer incidence and mortality in Europe in 2006," *Annals of oncology*, vol. 18, no. 3, pp. 581–592, 2007.

[18] B. S. Hulka and P. G. Moorman, "Breast cancer: hormones and other risk factors," *Maturitas*, vol. 38, no. 1, pp. 103–113, 2001.

[19] P. Autier, M. Boniol, C. LaVecchia, L. Vatten, A. Gavin, C. Héry, and M. Heanue, "Disparities in breast cancer mortality trends between 30 european countries: retrospective trend analysis of who mortality database," *BMJ: British Medical Journal*, vol. 341, 2010.

[20] M Angell, J. Kassirer, and A. Relman, "Looking back on the millennium in medicine [editorial]," *New England Journal Medicine*, vol. 342, no. 1, pp. 42–49, 2000.

[21] S Moore, "Better breast cancer detection," *IEEE Spectrum*, 2001.

[22] R. Bird, T. Wallace, and B. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology*, vol. 184, no. 3, pp. 613–617, 1992.

[23] N. F. Boyd, H. Guo, L. J. Martin, L. Sun, J. Stone, E. Fishell, R. A. Jong, G. Hislop, A. Chiarelli, S. Minkin, *et al.*, "Mammographic density and the risk and detection of breast cancer," *New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[24] K. Evers, "Diagnostic breast imaging mammography, sonography, magnetic resonance imaging, and interventional procedures," *American Journal of Roentgenology*, vol. 177, no. 5, pp. 1094–1094, 2001.

[25] E. A. Sickles, R. A. Filly, and P. W. Callen, "Breast cancer detection with sonography and mammography: comparison using state-of-the-art equipment," *American Journal of Roentgenology*, vol. 140, no. 5, pp. 843–845, 1983.

[26] A. P. Smith, P. A. Hall, D. M. Marcello, *et al.*, "Emerging technologies in breast cancer detection," *Radiology management*, vol. 26, no. 4, pp. 16–27, 2004.

[27] J.-H. Chung, V. Rajagopal, T. A. Laursen, P. M. Nielsen, and M. P. Nash, "Frictional contact mechanics methods for soft materials: application to tracking breast cancers," *Journal of biomechanics*, vol. 41, no. 1, pp. 69–77, 2008.

[28] M. Moskowitz, S. A. Feig, C. Cole-Beuglet, S. Fox, J. Haberman, H. Libshitz, and A Zermeno, "Evaluation of new imaging procedures for breast cancer," in *Early Detection of Breast Cancer*, Springer, 1984, pp. 55–61.

[29] J. M. Lewin, R. E. Hendrick, C. J. D'Orsi, P. K. Isaacs, L. J. Moss, A. Karellas, G. A. Sisney, C. C. Kuni, and G. R. Cutter, "Comparison of full-field digital mammography with screen-film mammography for cancer detection: Results of 4,945 paired examinations," *Radiology*, vol. 218, no. 3, pp. 873–880, 2001.

[30] A. Smith, "Fundamentals of breast tomosynthesis," *White Paper, Hologic Inc., WP-00007*, 2008.

[31] R. R. Entrekin, B. A. Porter, H. H. Sillesen, A. D. Wong, P. L. Cooperberg, and C. H. Fix, "Real-time spatial compound imaging: application to breast, vascular, and musculoskeletal ultrasound," in *Seminars in ultrasound, CT and MRI*, Elsevier, vol. 22, 2001, pp. 50–64.

[32] S. Huber, M. Wagner, M. Medl, and H. Czembirek, "Real-time spatial compound imaging in breast ultrasound," *Ultrasound in medicine & biology*, vol. 28, no. 2, pp. 155–163, 2002.

[33] R Entrekin, P Jackson, J. Jago, and B. Porter, "Real time spatial compound imaging in breast ultrasound: technology and early clinical experience," *medicamundi*, vol. 43, no. 3, pp. 35–43, 1999.

[34] P. B. Gordon *et al.*, "Ultrasound for breast cancer screening and staging," *Radiologic Clinics of North America*, vol. 40, no. 3, p. 431, 2002.

[35] H. Lewis-Jones, G. Whitehouse, and S. Leinster, "The role of magnetic resonance imaging in the assessment of local recurrent breast carcinoma," *Clinical radiology*, vol. 43, no. 3, pp. 197–204, 1991.

[36] R. L. Egan, "Experience with mammography in a tumor institution evaluation of 1,000 studies," *Radiology*, vol. 75, no. 6, pp. 894–900, 1960.

[37] J. J. Wild and J. M. Reid, "Further pilot echographic studies on the histologic structure of tumors of the living intact human breast," *The American journal of pathology*, vol. 28, no. 5, p. 839, 1952.

[38] P. J. Dempsey, "The history of breast ultrasound," *Journal of ultrasound in medicine*, vol. 23, no. 7, pp. 887–894, 2004.

[39] W Teh and A. Wilson, "The role of ultrasound in breast cancer screening. A consensus statement by the european group for breast cancer screening," *European Journal of cancer*, vol. 34, no. 4, pp. 449–450, 1998.

[40] M. B. Kossoff, "Ultrasound of the breast," *World journal of surgery*, vol. 24, no. 2, pp. 143–157, 2000.

[41] L. L. Humphrey, M. Helfand, B. K. Chan, and S. H. Woolf, "Breast cancer screening: a summary of the evidence for the US preventive services task force," *Annals of internal medicine*, vol. 137, no. 5, pp. 347–360, 2002.

[42] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Occult cancer in women with dense breasts: detection with screening US–diagnostic yield and tumor characteristics," *Radiology*, vol. 207, no. 1, pp. 191–199, 1998.

[43] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations1," *Radiology*, vol. 225, no. 1, pp. 165–175, 2002.

[44] C. Kuhl, "MRI of breast tumors," *European radiology*, vol. 10, no. 1, pp. 46–58, 2000.

[45] I. Andersson, D. M. Ikeda, S. Zackrisson, M. Ruschin, T. Svahn, P. Timberg, and A. Tingberg, "Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings," *European radiology*, vol. 18, no. 12, pp. 2817–2825, 2008.

[46] V Jackson, "The role of US in breast imaging," *Radiology*, 1990.

[47] O. Graf, T. H. Helbich, M. H. Fuchsjaeger, G. Hopf, M. Morgun, C. Graf, R. Mallek, and E. A. Sickles, "Follow-up of palpable circumscribed noncalcified solid breast masses at mammography and US: can biopsy be averted?" *Radiology*, vol. 233, no. 3, pp. 850–856, 2004.

[48] N. Hines, P. J. Slanetz, and R. L. Eisenberg, "Cystic masses of the breast," *American Journal of Roentgenology*, vol. 194, no. 2, W122–W133, 2010.

[49] D. Leucht, H. Madjar, and W. Leucht, *Teaching atlas of breast ultrasound*. Thieme, 1996.

[50] J. A. Jensen, "Field: A program for simulating ultrasound systems," in *10th Nordicbaltic Conference on Biomedical Imaging*, Citeseer, vol. 4, 1996, pp. 351–353.

[51] A. S. Hong, E. L. Rosen, M. S. Soo, and J. A. Baker, "BI-RADS for sonography: positive and negative predictive values of sonographic features," *AJR Am J Roentgenol*, vol. 184, no. 4, pp. 1260–5, 2005.

[52] E. Lazarus, M. B. Mainiero, B. Schepps, S. L. Koelliker, and L. S. Livingston, "BI-RADS lexicon for US and mammography: Interobserver variability and positive predictive value1," *Radiology*, vol. 239, no. 2, pp. 385–391, 2006.

[53] N. Abdullah, B. Mesurolle, M. El-Khoury, and E. Kao, "Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses," *Radiology*, vol. 252, no. 3, pp. 665–672, 2009.

[54] M. Calas, R. Almeida, B Gutfilen, and W. Pereira, "Intraobserver interpretation of breast ultrasonography following the BI-RADS classification," *European journal of radiology*, vol. 74, no. 3, pp. 525–528, 2010.

[55] L. B. Lusted, "Medical electronics," *New England Journal of Medicine*, vol. 252, no. 14, pp. 580–585, 1955.

[56] H.-P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, H. MacMahon, *et al.*, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis.," *Investigative radiology*, vol. 25, no. 10, p. 1102, 1990.

[57] B Liu, H Cheng, J Huang, J Tian, X Tang, and J Liu, "Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images," *Pattern Recognition*, 2010.

[58] K. Horsch, M. L. Giger, L Venta, and C Vyborny, "Automatic segmentation of breast lesions on ultrasound," *Medical Physics*, 2001.

[59] Y.-L. Huang and D.-R. Chen, "Watershed segmentation for breast tumor in 2-D sonography," *Ultrasound in Medicine & Biology*, vol. 30, no. 5, pp. 625–32, 2004.

[60] A Madabhushi and D Metaxas, "Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions," *IEEE Transactions on medical imaging*, 2003.

[61] J. Massich, F. Meriaudeau, E. Pérez, R. Martí, A. Oliver, and J. Martí, "Lesion segmentation in breast sonography," *Digital Mammography*, pp. 39–45, 2010.

[62] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: a survey," *Pattern Recognition*, vol. 43, no. 1, pp. 299–317, 2009. DOI: 10.1016/j.patcog.2009.05.012.

[63] D. Angelova and L. Mihaylova, "Contour segmentation in 2d ultrasound medical images with particle filtering," *Machine Vision and Applications*, vol. 22, no. 3, pp. 551–561, 2011.

[64] W Gómez, L Leija, A. V Alvarenga, A. F. C Infantosi, and W. C. A Pereira, "Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation," *Medical Physics*, vol. 37, no. 1, p. 82, 2010.

[65] G Xiao, M Brady, J. A. Noble, and Y Zhang, "Segmentation of ultrasound B-mode images with intensityinhomogeneity correction," *IEEE Transactions on medical imaging*, vol. 21, no. 1, pp. 48–57, 2002.

[66] G. Pons, J. Martí, R. Martí, S. Ganau, J. Vilanova, and J. Noble, "Evaluating lesion segmentation in breast ultrasound images related to lesion typology," *Journal of Ultrasound in Medicine*, 2013.

[67] H.-H. Chiang, J.-Z. Cheng, P.-K. Hung, C.-Y. Liu, C.-H. Chung, and C.-M. Chen, "Cell-based graph cut for segmentation of 2D/3D sonographic breast images," in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, IEEE, 2010, pp. 177–180.

[68] M. Alemán-Flores, L. Álvarez, and V. Caselles, "Texture-oriented anisotropic filtering and geodesic active contours in breast tumor ultrasound segmentation," *J Math Imaging Vis*, vol. 28, no. 1, pp. 81–97, 2007.

[69] J. Cui, B. Sahiner, H.-P. Chan, A. Nees, C. Paramagul, L. M. Hadjiiski, C. Zhou, and J. Shi, "A new automated method for the segmentation and characterization of breast masses on ultrasound images," *Medical Physics*, vol. 36, no. 5, p. 1553, 2009.

[70] L. Gao, X. Liu, and W. Chen, "Phase- and GVF-Based level set segmentation of ultrasonic breast tumors," *Journal of Applied Mathematics*, vol. 2012, pp. 1–22, 2012.

[71] K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Medical Physics*, vol. 29, no. 7, pp. 1438–46, 2002.

[72] J. Massich, F. Meriaudeau, E. Pérez, R. Martí, A. Oliver, and J. Martí, "Seed selection criteria for breast lesion segmentation in ultrasound images," *MICCAI Workshop on Breast Image Analysis*, pp. 55–64, 2011.

[73] J. Massich, F. Meriaudeau, M. Santís, S. Ganau, E. Pérez, R. Martí, A. Oliver, and J. Martí, "Automatic seed placement for breast lesion segmentation on US images," *Digital Mammography*, pp. 308–315, 2012.

[74] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[75] Y.-L. Huang, Y.-R. Jiang, D.-R. Chen, and W. K. Moon, "Level set contouring for breast tumor in sonography," *Journal of digital imaging*, vol. 20, no. 3, pp. 238–247, 2007.

[76] J. Zhang, S. K. Zhou, S. Brunke, C. Lowery, and D. Comaniciu, "Database-guided breast tumor detection and segmentation in 2D ultrasound images," in *SPIE Medical Imaging*, International Society for Optics and Photonics, vol. 7624, 2010, pp. 762 405–762 405.

[77] P. Jiang, J. Peng, G. Zang, E. Cheng, V. Megalooikonomou, and H. Ling, "Learning-based automatic breast tumor detection and segmentation in ultrasound images," pp. 1–4, 2012.

[78] J. Shan, H Cheng, and Y. Wang; "A novel automatic seed point selection algorithm for breast ultrasound images," *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1 –4, 2008.

[79] J. Shan, H. D. Cheng, and Y. Wang, "Completely automated segmentation approach for breast ultrasound images using multiple-domain features," *Ultrasound in Medicine & Biology*, vol. 38, no. 2, pp. 262–275, 2012.

[80] Y.-L. Huang and D.-R. Chen, "Automatic contouring for breast tumors in 2-D sonography," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005*, IEEE, 2006, pp. 3225–3228.

[81] Q.-H. Huang, S.-Y. Lee, L.-Z. Liu, M.-H. Lu, L.-W. Jin, and A.-H. Li, "A robust graph-based segmentation method for breast tumors in ultrasound images," *Ultrasonics*, vol. 52, no. 2, pp. 266–275, 2012.

[82] X. Liu, Z. Huo, and J. Zhang, "Automated segmentation of breast lesions in ultrasound images," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, IEEE, 2006, pp. 7433–7435.

[83] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[84] Z. Hao, Q. Wang, Y. K. Seong, J.-H. Lee, H. Ren, and J.-y. Kim, "Combining CRF and multi-hypothesis detection for accurate lesion segmentation in breast sonograms," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, Springer, 2012, pp. 504–511.

[85] B. Liu, H. D. Cheng, J. Huang, J. Tian, X. Tang, and J. Liu, "Probability density difference-based active contour for ultrasound image segmentation," *Pattern Recognition*, 2010.

[86] C Yeh, Y Chen, W Fan, and Y Liao, "A disk expansion segmentation method for ultrasonic breast lesions," *Pattern Recognition*, 2009.

[87] E. N. Mortensen and W. A. Barrett, "Interactive segmentation with intelligent scissors," *Graphical models and image processing*, vol. 60, no. 5, pp. 349–384, 1998.

[88] P. Pérez, A. Blake, and M. Gangnet, "Jetstream: Probabilistic contour extraction with particles," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, IEEE, vol. 2, 2001, pp. 524–531.

[89] A. X. Falcão, J. K. Udupa, S. Samarasekera, S. Sharma, B. E. Hirsch, and R. d. A. Lotufo, "User-steered image segmentation paradigms: Live wire and live lane," *Graphical models and image processing*, vol. 60, no. 4, pp. 233–260, 1998.

[90] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 303–308, 2004.

[91] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, ACM, vol. 23, 2004, pp. 309–314.

[92] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, IEEE, vol. 1, 2001, pp. 105–112.

[93] D. Angelova and L. Mihaylova, "Contour extraction from ultrasound images viewed as a tracking problem," in *Information Fusion, 2009. FUSION'09. 12th International Conference on*, IEEE, 2009, pp. 284–291.

[94] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychological Research*, vol. 64, no. 2, pp. 136–148, 2000.

[95] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[96] P. Kovesi, "Image features from phase congruency," *Videre: Journal of computer vision research*, vol. 1, no. 3, pp. 1–26, 1999.

[97] S. Lobregt and M. A. Viergever, "A discrete dynamic contour model," *Medical Imaging, IEEE Transactions on*, vol. 14, no. 1, pp. 12–24, 1995.

[98] S. Beucher *et al.*, "The watershed transformation applied to image segmentation," *Scanning microscopy-supplement*, pp. 299–299, 1992.

[99] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 12, pp. 1163–1173, 1996.

[100] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[101] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Computer Vision, 2009 IEEE 12th International Conference on*, IEEE, 2009, pp. 670–677.

[102] J. A. Noble and P. N. T. Wells, "Ultrasound image segmentation and tissue characterization," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 224, no. 2, pp. 307–316, 2009.

[103] D. Cremers, M. Rousson, and R. Deriche, "A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape," *International journal of computer vision*, vol. 72, no. 2, pp. 195–215, 2007.

[104] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[105] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 1–27, 2012.

[106] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[107] S. Osher and R. Fedkiw, *Level set methods and dynamic implicit surfaces*. Springer Verlag, 2003, vol. 153.

[108] J. A. Noble and D Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Transactions on medical imaging*, 2006.

[109] A. K. Jumaat, W. E. Z. W. Rahman, A. Ibrahim, and R. Mahmud, "Comparison of balloon snake and GVF snake in segmenting masses from breast ultrasound images," in *Computer Research and Development, 2010 Second International Conference on*, IEEE, 2010, pp. 505–509.

[110] L. D. Cohen, "On active contour models and balloons," *CVGIP: Image understanding*, vol. 53, no. 2, pp. 211–218, 1991.

[111] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *Image Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 359–369, 1998.

[112] M Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Transactions on medical imaging*, 1998.

[113] K. Horsch, M. L. Giger, L Venta, and C Vyborny, "Computerized diagnosis of breast lesions on ultrasound," *Medical Physics*, 2002.

[114] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proceedings of the National Academy of Sciences*, vol. 93, no. 4, pp. 1591–1595, 1996.

[115] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," 2012.

[116] Y. Liu, H. D Cheng, J. Huang, Y. Zhang, and X. Tang, "An effective approach of lesion segmentation within the breast ultrasound image based on the cellular automata principle," *Journal of Digital Imaging*, pp. 1–11, 2012.

[117] T Chan and L Vese, "Active contours without edges," *IEEE Transactions on image Processing*, 2001.

[118] A Madabhushi and D Metaxas, "Automatic boundary extraction of ultrasonic breast lesions," *Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on*, pp. 601 –604, 2002.

[119] C. Kotropoulos and I. Pitas, "Segmentation of ultrasonic images using support vector machines," *Pattern Recognition Letters*, vol. 24, no. 4, pp. 715–727, 2003.

[120] R. Martí, J. Martí, J. Freixenet, R. Zwiggelaar, J. Vilanova, and J. Barceló, "Optimally discriminant moments for speckle detection in real b-scan images," *Ultrasonics*, vol. 48, no. 3, pp. 169 –181, 2008, ISSN: 0041-624X.

[121] T. Deselaers, D. Keysers, and H. Ney, "Discriminative training for object recognition using image patches," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, IEEE, vol. 2, 2005, pp. 157–162.

[122] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, ACM, 2007, pp. 401–408.

[123] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[124] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool, "The Pascal Visual Object Classes challenge 2006 (VOC 2006) results," 2006.

[125] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition(CVPR), 2008 IEEE Conference on*, IEEE, 2008, pp. 1–8.

[126] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[127] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[128] O. Veksler, Y. Boykov, and P. Mehrani, "Superpixels and supervoxels in an energy optimization framework," in *Computer Vision–ECCV 2010*, Springer, 2010, pp. 211–224.

[129] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2290–2297, 2009.

[130] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Computer Vision–ECCV 2008*, Springer, 2008, pp. 705–718.

[131] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[132] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 898–916, 2011.

[133] O. Pele and M. Werman, "The quadratic-chi histogram distance family," in *Computer Vision–ECCV 2010*, Springer, 2010, pp. 749–762.

[134] L. Sachs and Z. Reynarowych, *Applied statistics: a handbook of techniques*. Springer-Verlag New York, 1984.

[135] J. Ponce, D. Forsyth, E.-p. Willow, S. Antipolis-Méditerranée, R. d'activité RAweb, L. Inria, and I. Alumni, "Computer vision: a modern approach," *Computer*, vol. 16, p. 11, 2011.

[136] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International journal of computer vision*, vol. 73, no. 2, pp. 213–238, 2007.

[137] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition(CVPR), 2006 IEEE Conference on*, IEEE, vol. 2, 2006, pp. 2169–2178.

[138] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[139] S. Edelman, N. Intrator, and T. Poggio, "Complex cells and object recognition," *unplublished*, 1997. [Online]. Available: `http://kybele.psych.cornell.edu/~edelman/archive.html`.

[140] K. P. Murphy, *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[141] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, 2004, p. 22.

[142] A. Rosenfeld *et al.*, *Multiresolution image processing and analysis*. Springer-Verlag New York: 1984, vol. 12.

[143] I. Daubechies *et al.*, *Ten lectures on wavelets*. SIAM, 1992, vol. 61.

[144] K. V. Mogatadakala, K. D. Donohue, C. W. Piccoli, and F. Forsberg, "Detection of breast lesion regions in ultrasound images using wavelets and order statistics," *Medical physics*, vol. 33, p. 840, 2006.

[145] D.-R. Chen, R.-F. Chang, W.-J. Kuo, M.-C. Chen, and Y.-L. Huang, "Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks," *Ultrasound in medicine & biology*, vol. 28, no. 10, pp. 1301–1310, 2002.

[146] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, 27:1–27:27, 3 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[147] S. Z. Li, *Markov random field modeling in image analysis*. Springer, 2009.

[148] G. A. Kochenberger *et al.*, *Handbook in Metaheuristics*. Springer, 2003.

[149] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 6, pp. 1068–1080, 2008.

[150] M. Gendreau and J.-Y. Potvin, "Metaheuristics in combinatorial optimization," *Annals of Operations Research*, vol. 140, no. 1, pp. 189–213, 2005.

[151] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.

[152] S. Kirkpatrick, D. G. Jr., and M. P. Vecchi, "Optimization by simmulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[153] T. A. Feo and M. G. Resende, "Greedy randomized adaptive search procedures," *Journal of global optimization*, vol. 6, no. 2, pp. 109–133, 1995.

[154] E. Aarts, J. Korst, and W. Michiels, "Simulated annealing," in *Search methodologies*, Springer, 2005, pp. 187–210.

[155] Z. J. Czech and P. Czarnas, "Parallel simulated annealing for the vehicle routing problem with time windows," in *Parallel, Distributed and Network-based Processing, 2002. Proceedings. 10th Euromicro Workshop on*, IEEE, 2002, pp. 376–383.

[156] N. L. Bigg, E. K. Lloyd, and R. J. Wilson, *Graph Theory: 1736-1936*. Oxford University Press, 1976.

[157] K Drukker, M. Giger, C. Vyborny, and E. Mendelson, "Computerized detection and classification of cancer on breast ultrasound.," *Academic radiology*, vol. 11, no. 5, p. 526, 2004.

[158] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision–ECCV 2008*, Springer, 2008, pp. 696–709.

[159] F. B. Silva, S. Goldenstein, S. Tabbone, and R. d. S. Torres, "Image classification based on bag of visual graphs," in *Image Processing, 2013. Proceedings. 2013 International Conference on*, IEEE, 2013.