Segmentation of breast lesions in Ultrasound images: A survey

Joan Massich[a,b], Joan Martí[a], Fabrice Meriaudeau[b]

[a] *Computer Vision and Robotics Group, Universitat de Girona, Campus Montilivi, Edifici PIV, s/n, 17071 Girona, Spain*

[b] *Le2i-UMR CNRS 6306, UniversitÃľ de Bourgogne, 12 rue de la Fonderie, 712000 Le Creusot, France*

Abstract

Breast cancer still has huge impact due to its place as the leading cause of cancer death among female population. However, medical imaging is a key for breast cancer mortality reduction, since it can increase the success of treatment contributing to its early detection through screening, diagnosis, image-guided biopsy, treatment follow-up and suchlike procedures. Recently, Ultra-Sound (US) imaging has grown into an essential tool to detect and analyze breast abnormalities, specially those present in very dense tissue. It is accepted that when dealing with US images, the most discriminative signs for diagnose are subject to the lesion delimitation. Therefore, the importance to develop segmentation procedures to properly delineate lesions in breast US images in order to improve Computer Aided Diagnosis (CAD) systems. This paper presents a taxonomy of methodologies used for segmenting breast lesions in US images, including a review of the evaluation methodologies used to assess their performance.

Email address: `jmassich@eia.udg.edu` (Joan Massich)

Keywords:

Breast cancer; Ultrasound (US) imaging; Segmentation; Segmentation evaluation.

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 Ferlay et al. [2010]. In terms of mortality, breast cancer is the fifth most common cause of cancer death. However, it places as the leading cause of cancer death among females both in western countries and in economically developing countries Jemal et al. [2011].

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 Smith et al. [2003]. Although Digital Mammography (DM) remains the reference imaging modality, US imaging has proven to be a successful adjunct image modality for breast cancer screening Smith et al. [2003]; Berg et al. [2004], specially as a consequence of the discriminative capabilities that US offers for differentiating between solid lesions that are benign or malignant Stavros et al. [1995] so that the amount of unnecessary biopsies, which is estimated to be between $65 \sim 85\%$ of the prescribed biopsies Yuan et al. [2010], can be reduced Ciatto et al. [1994] in replacing them by short-term US screening follow-up Gordon and Goldenberg [1995].

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 Ensminger and Stulen [2008] 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 Manning et al. [2005].

Therefore the major goals of medical imaging researchers in general, and also in particular for breast lesion assessment using US data, have 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 Stavros et al. [1995]; Mendelson et al. [2001, 2003]; Stavros [2004] in order to reduce diagnosis inconsistencies among readers Baker et al. [1999]. 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 Altman and Bland [1994] 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 CAD as a second opinion has been proven to achieve comparable results Giger et al. [2008], CAD systems are used to alleviate the time demand from the radiologists.

CAD systems applied to aid radiologist when reading US images of the breast take advantage of either low-level features, high-level features or both Cheng et al. [2009]. Jalalian et al. Jalalian et al. [2012] reported that the majority of such hig-level features describing the lesions which bring reliable information to the systems, can be found in these lexicon tools already used by radiologists Mendelson et al. [2003]; Stavros [2004]. However, to take advantage of these descriptors, procedures to accurately segment the lesions are needed. This need comes from the fact that when the images are read by an expert radiologist, the underlying delineation of the lesion is instantly understood.

This article reviews recent advances in breast lesion segmentation in US data procedures which can be used for further extracting reliable high-level features for improving CAD systems applied to breast US images.

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

Segmentation is a fundamental procedure for a CAD system. Figure 1 illustrates the idea that procedures for segmentating breast lesions in US data can be found within a CAD system workflow as part of Computer Aided Detection (CADe), as part of Computer Aided Diagnosis (CADx) or as a stand alone step using detection information and providing further information that can be used for conducting a diagnosis.

Segmentation procedures integrated within CAD systems can either be

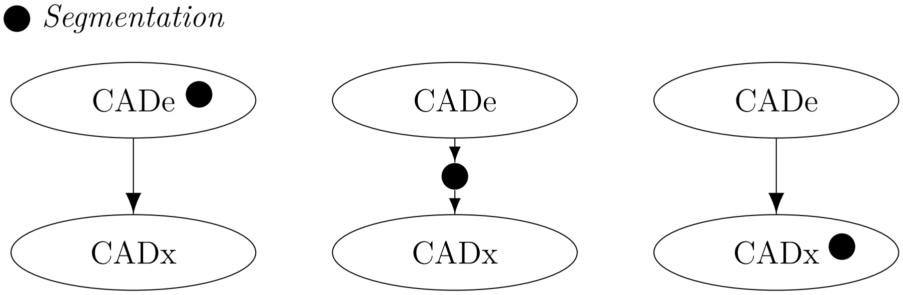


Figure 1: Illustrative idea of the role of segmentation within a CAD framework showing that it can either be a separate process between a CADe and a CADx or it can belong to any of the two CAD typologies: CADe, CADx

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. Hong et al. [2005], which describes a system working on Breast Imaging-Reporting and Data System (BI-RADS) descriptors supplied by an expert based on the reading of images.

Figure 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.

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

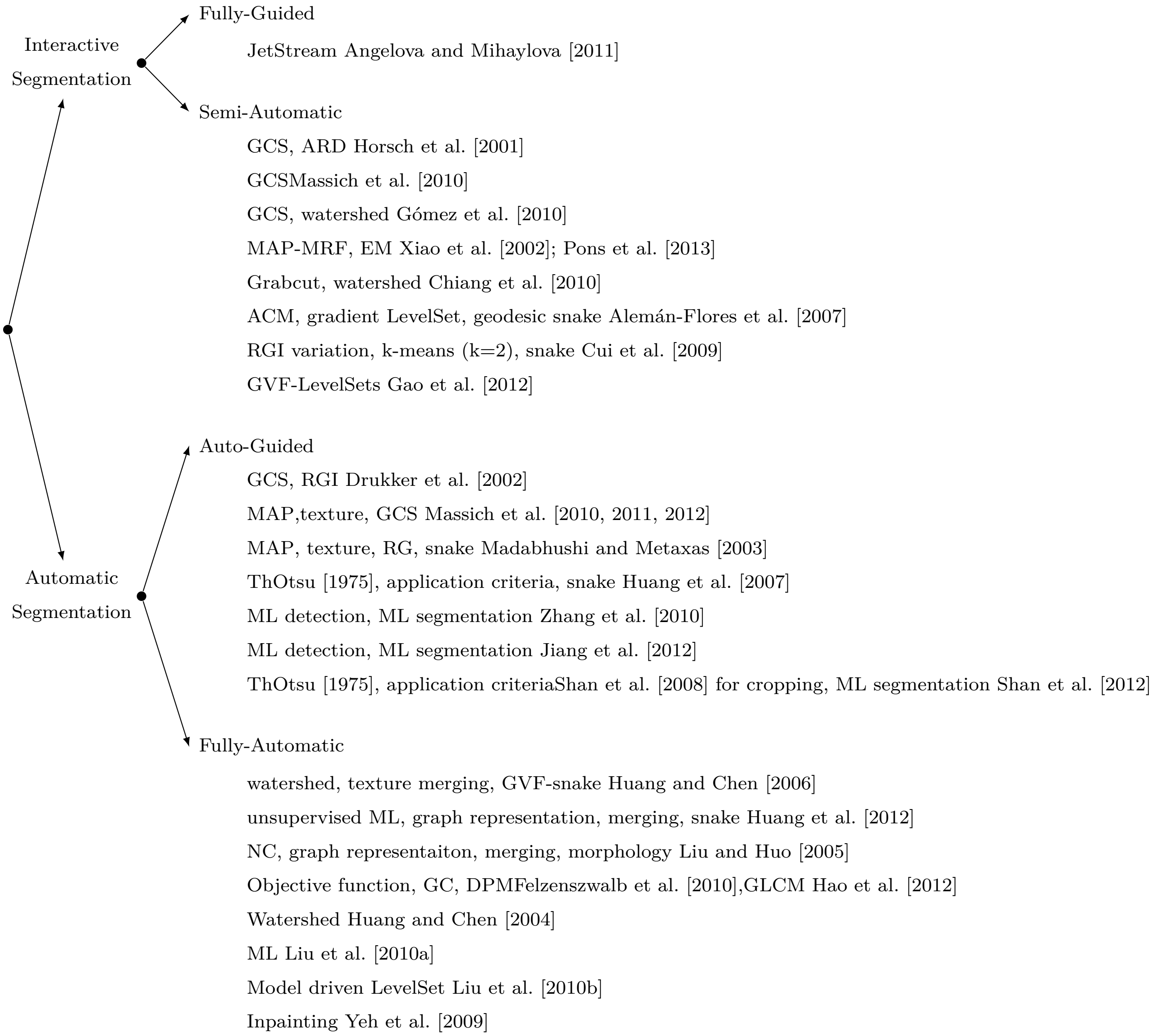

Interactive
Segmentation
Fully-Guided
JetStream Angelova and Mihaylova [2011]
Semi-Automatic
GCS, ARD Horsch et al. [2001]
GCSMassich et al. [2010]
GCS, watershed Gómez et al. [2010]
MAP-MRF, EM Xiao et al. [2002]; Pons et al. [2013]
Grabcut, watershed Chiang et al. [2010]
ACM, gradient LevelSet, geodesic snake Alemán-Flores et al. [2007]
RGI variation, k-means (k=2), snake Cui et al. [2009]
GVF-LevelSets Gao et al. [2012]
Automatic
Segmentation
Auto-Guided
GCS, RGI Drukker et al. [2002]
MAP,texture, GCS Massich et al. [2010, 2011, 2012]
MAP, texture, RG, snake Madabhushi and Metaxas [2003]
ThOtsu [1975], application criteria, snake Huang et al. [2007]
ML detection, ML segmentation Zhang et al. [2010]
ML detection, ML segmentation Jiang et al. [2012]
ThOtsu [1975], application criteriaShan et al. [2008] for cropping, ML segmentation Shan et al. [2012]
Fully-Automatic
watershed, texture merging, GVF-snake Huang and Chen [2006]
unsupervised ML, graph representation, merging, snake Huang et al. [2012]
NC, graph representaiton, merging, morphology Liu and Huo [2005]
Objective function, GC, DPMFelzenszwalb et al. [2010],GLCM Hao et al. [2012]
Watershed Huang and Chen [2004]
ML Liu et al. [2010a]
Model driven LevelSet Liu et al. [2010b]
Inpainting Yeh et al. [2009]

Figure 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* Mortensen and Barrett [1998], or *Jetstream* segmentation Pérez et al. [2001], both deriving from the *live-wire* technique Falcão et al. [1998], 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* Li et al. [2004] and *grabcut* Rother et al. [2004] both based on the work proposed by Boykov and Jolly Boykov and Jolly [2001] which takes advantage of Graph-Cut (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.

1.1.1. 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 Angelova and Mihaylova [2011, 2009] implemented a jetstream Pérez et al. [2001] 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.

1.1.2. 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. [2001] propose a method 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. Massich et al. [2010] proposed a methodology inspired by GCS with different user interactability levels that falls 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 Massich et al. [2010], and the methodology proposed by Massich et al. demands less user interaction. Another work based on GCS Horsch et al. [2001] is the work proposed by

Gomez et al. Gómez et al. [2010] 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. Xiao et al. [2002], 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 Xiao et al. [2002] 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 Pons et al. [2013] 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. Chiang et al. [2010] or the various manually initialized implementations of the popular Active Contour Models (ACMs) technique Alemán-Flores et al. [2007]; Cui et al. [2009]; Gao et al. [2012]. These ACM methodologies reported really good results achieving a mean

AOV of 0.883 for the implementation presented in Alemán-Flores et al. [2007]. Within the group of methodologies using ACM, Alemán-Flores et al. Alemán-Flores et al. [2007] 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. Cui et al. [2009] 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. Gao et al. [2012] 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 Kovesi [2000] along with the gradient information, achieving a mean AOV of 0.863.

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 design-

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

ing 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. Drukker et al. [2002] , Horsch et al. Horsch et al. [2001]) or (Shan et al. Shan et al. [2008], (Shan et al. Shan et al. [2012]).

1.2.1. 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. Drukker et al. [2002] where an automatic detection procedure is added to the original GCS segmentation Horsch et al. [2001] 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 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 Horsch et al. [2001]. 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. Massich et al. [2010] also proposed a methodology based on GCS as Drukker et al. [2002] 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 Horsch et al. [2001] 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

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

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) Warfield et al. [2004] 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 Horsch et al. [2001] 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 Massich et al. [2011, 2012]. 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 Massich et al. [2012]). 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 Madabhushi and Metaxas [2003] proposed using the *Stavros Criteria* Stavros [2004] 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. Huang et al. [2007] 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 Otsu [1975] 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. Shan et al. [2008]. A DSC of 0.876 is reported using a dataset of 118 images.

Zhang et al. Zhang et al. [2010] and Jiang et al. Jiang et al. [2012] 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. Xiao et al. [2002]. In contrast, if the statistical models used to drive the second ML step producing the final segmentation in Zhang et al. [2010]; Jiang et al. [2012] were inferred from dataset annotations, then both methodologies would be considered fully-guided and would resemble the work proposed by Hao et al. Hao et al. [2012] 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. Shan et al. [2012] proposed to use the detection just to simplify the following segmentation procedure. The lesion detection procedure described in Shan et al. [2008] 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 Massich et al. [2010], energy-based phase information Kovesi [1999] and distance to the initially detected contour Shan et al. [2008] 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. Shan et al. [2012] 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.

1.2.2. 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 Huang and Chen [2006] proposed using an ACM to perform the final segmentation Lobregt and Viergever [1995] 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 Beucher et al. [1992]; Najman and Schmitt [1996] 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. Huang et al. [2012] and Liu et al. Liu and Huo [2005] also split the image into regions or segments as a first step for further analysis. To determine the image segments, Huang et al. Huang et al. [2012] use unsupervised learning and Liu et al. Liu and Huo [2005] use normalized cuts Shi and Malik [2000] in order to achieve an image over-segmentation as that obtained when applying the watershed transform in Huang and Chen [2006]. The difference between the three works lies in how the segments are managed once determined since both Huang et al. [2012]; Liu and Huo [2005] 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 Boykov and Jolly [2001]. 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 Fulkerson et al. [2009] and have also been applied to breast lesion segmentation in US data Hao et al. [2012]. 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. Hao et al. [2012] proposed to automatically segment breast lesions using an objective function combining Deformable Part Model (DPM) Felzenszwalb et al. [2010] detection with intensity histograms, a GLCM based texture descriptor and position information using a Graph-Cut minimization tool and normalized cuts Shi and Malik [2000] as image segments. The proposed methodology reported an average AOV of 0.75 of a 480 image database.

In contrast, Huang and Chen Huang and Chen [2004] only performed the spliting of the image using watershed transform, while Liu et al. Liu et al. [2010a] 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. Liu et al. [2010b] 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 Noble and Wells [2009] 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. Yeh et al. [2009], 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. 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. Cremers et al. [2007] 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 Cremers et al. [2007]. Although the chronological difference cannot be appreciated for breast lesion segmentation since early applications

[3] this value has been calculated from the TP, FN and FP values reported in Yeh et al. [2009]

such as Xio et al. Xiao et al. [2002] were already taking advantage of optimization methods. A tendency to move towards optimization methodologies, as can be seen Jiang et al. [2012], in lieu of methodologies driven by obscure heuristics in a full manner such as in Drukker et al. [2002]; Horsch et al. [2001]; Massich et al. [2010] or partially like Madabhushi and Metaxas [2003].

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 Boykov et al. [2001] 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. Delong et al. [2012] as a continuation of the work started by Boykov et al. Boykov et al. [2001] in their seminal paper of Graph-Cut (GC).

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. Kass et al. [1988] 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 Osher and Fedkiw [2003]. 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 Liu et al. [2010b].

Figure 3 maps the methodologies presented in section 1 (see fig. 2) regarding its usage of ML, ACM, and other strategies.

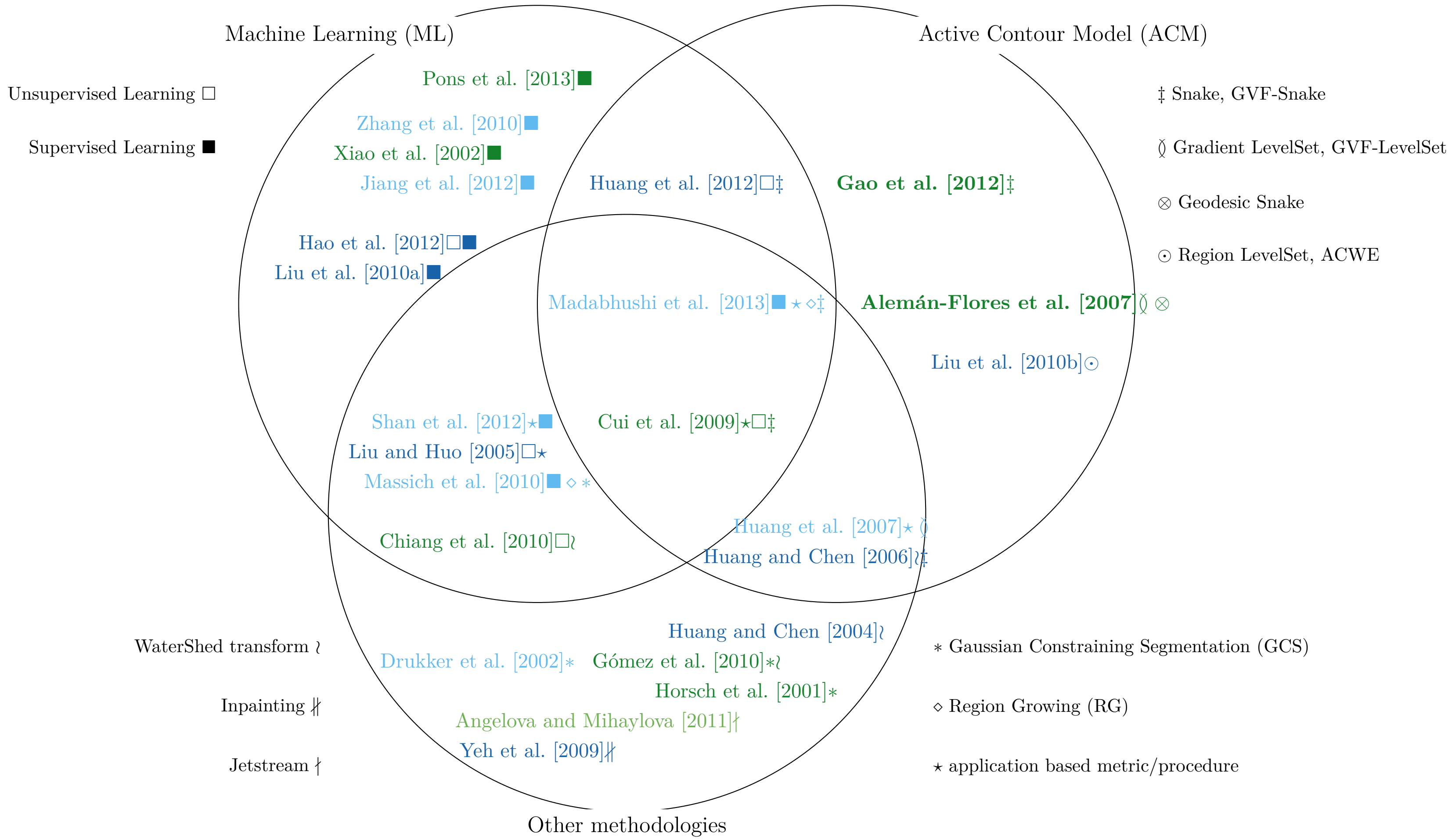


Figure 3: Conceptual map of the segmentation strategy used in the methodologies reported in figure 2. The methods have been grouped according to the segmentation methodology: ML,ACM or others. Each circle has its own iconography representing the sub-strategies that can be found in each class. The color 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).

2.1. Active Contour Models (ACMs)

ACM segmentation techniques are widely applied in US applications such as organ delineation Noble and Boukerroui [2006] or breast lesion segmentation Alemán-Flores et al. [2007]; Cui et al. [2009]; Gao et al. [2012]; Madabhushi and Metaxas [2003]; Huang et al. [2007]; Huang and Chen [2006]; Huang et al. [2012]; Liu et al. [2010b]. Notice in figures 2 and 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. Jumaat et al. [2010] 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 Cohen [1991] reported better performance than GVF-Snakes Xu and Prince [1998].

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. Liu et al. [2010b] 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. 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 Shan et al. [2012]; Hao et al. [2012].
- use information from the target image itself to drive the segmentation Zhang et al. [2010]; Jiang et al. [2012].

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 Xiao et al. [2002]; Pons et al. [2013]; or the information must be provided by another automatic procedure leading to an auto-guided procedure such as Zhang et al. [2010]. However, for detection application, only information from other images with accompanying GTs are used Massich et al. [2011, 2012]; Madabhushi and Metaxas [2003], since user interaction would already solve the detection problem. Taking this into account, figure 4 illustrates the 5 possible scenarios.

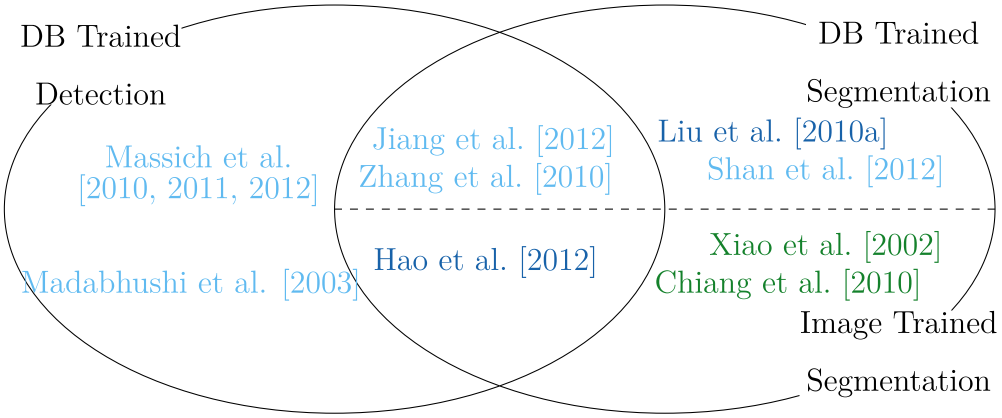


Figure 4: Supervised Machine Learning (ML) training and goals, ending up with a combination of 5 different strategies. The references are colored indicating the user interaction: semi-automatic (light-green), auto-guided(light-Blue), and fully automatic(dark-blue).

Database Trained Detection: generates statistic models from a training dataset to detect lesions in a target image using any sort of ML and features Massich et al. [2010, 2011, 2012]; Madabhushi and Metaxas [2003]; Zhang et al. [2010]; Jiang et al. [2012]; Hao et al. [2012].

Image Trained Segmentation: from information supplied by the user, an ML procedure is trained from the target image in order to produce a segmentation Xiao et al. [2002]; Pons et al. [2013].

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 Liu et al. [2010a]; Shan et al. [2012]. In such a scenario, the features for constructing the models need to be robust to significative differences between the images.

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 Zhang et al. [2010]; Jiang et al. [2012].

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 weather 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 Hao et al. [2012].

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 Horsch et al. [2001]; Massich et al. [2010]; Drukker et al. [2002]; Gómez et al. [2010] are inspired by the work of Kupinski et al. Kupinski and Giger [1998] which was initially adapted to US data by Horsch et al.Horsch et al. [2002]. 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 Sethian [1996]. 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 Huang and Chen [2006]; Huang et al. [2012]

or higher level features extraction from a coherent area so that it can be used in ML procedures Hao et al. [2012]; Chiang et al. [2010]. 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 Cui et al. [2009], and robust graph based clustering Huang et al. [2012]. 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 Huang and Chen [2006]; Chiang et al. [2010]; Huang and Chen [2004] and Normalized Cuts (NC) Shi and Malik [2000]; Hao et al. [2012]; Liu and Huo [2005] are popular techniques used to obtain an over-segmentation, also known as super pixels Achanta et al. [2012].

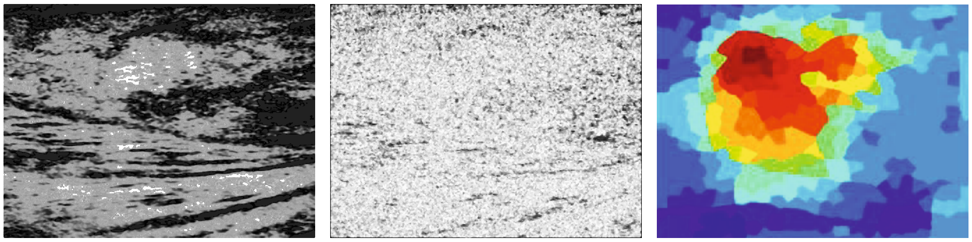
Finally, Yeh et al. Yeh et al. [2009] proposed a totally different approach for breast lesion segmentation based on inpainting of degraded typology. The image is transformed into a binary image using local thresholding and then the largest object within the binary image is reconstructed as the final segmentation.

2.4. 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. Alemán-Flores et al. [2007]. 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 Ensminger and Stulen [2008]. However, images does look like a combination of texture granularities depending on the tissue which has encouraged the exploration of texture descriptors Massich et al. [2010, 2011, 2012]; Madabhushi and Metaxas [2003]; Liu et al. [2012]; Hao et al. [2012]; Huang and Chen [2006]. However, the use of a naive descriptor, like the one used in Massich et al. [2010, 2011]; Madabhushi and Metaxas [2003], 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 5, where the latent information contained in the texture (fig. 5b) is less than that contained in the intensity feature (fig. 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



(a) (b) (c)

Figure 5: Qualitative assessment of feature planes: (a) Maximum A Posteriori (MAP) of intensity feature, (b) MAP of texture feature used in Massich et al. [2010]; Madabhushi and Metaxas [2003] and (c) quantified DPM feature Hao et al. [2012](image taken from the original work in Hao et al. [2012]).

textures can be seen as the outcome of an ML process. Hao et al. Hao et al. [2012] propose to synthesize texture from a lesion detection process (DPM) that takes advantage of Histogram of Gradients (HOG) taken at different scales. Figure 5c illustrates the feature plane inferred from the DPM process.

3. Segmentation assessment

Comparing all the methodologies reviewed in section 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 Noble and Boukerroui [2006]; Noble and Wells [2009]; Cheng et al. [2009]. 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
- 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 unpropper detection

The dificulty of comparing the methodologies using distinct datasets, distinct assessing criteria and distinct metrics is clear. Section 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 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 1 and fig. 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 Pons et al. [2013]. In this regard, later in this article (see section: 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 criterion is not set. Massich et al. Massich et al. [2010] take all the segmentations into account even if the segmentation has been wrongly initialized by the automatic detection procedure. Meanwhile, Zhang et al. Zhang et al. [2010] 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.

3.1. Evaluation criteria

Although multiple criteria arise when assessing segmentations, these criteria can be grouped into two families depending on whether they are area or distance based metrics as illustrated in figure 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.

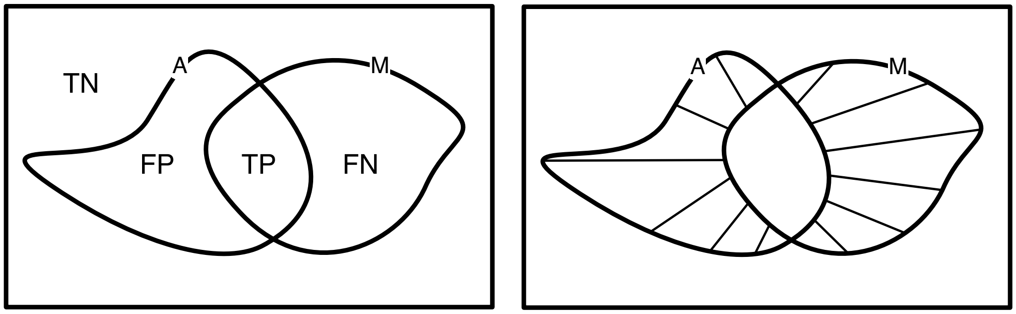
For the sake of simplicity, the name of the reported similarity indexes has been unified.

3.1.1. 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. 6b), 4

		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 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.

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 6a.

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: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) Gao et al. [2012] or Similarity Index (SI) Shan et al. [2012][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 1. The AOV metric has been used to assess the following works: Horsch et al. [2001]; Gómez et al. [2010]; Alemán-Flores et al. [2007]; Cui et al. [2009]; Massich et al. [2010]; Shan et al. [2012]; Hao et al. [2012]; Liu et al. [2010b]

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

Dice Similarity Coefficient (DSC), also found under the name of SI Huang et al. [2007]; Huang and Chen [2006][5], is another widely used overlap

[4] Notice that Similarity Index (SI) is also used formulated as the Dice Similarity Coefficient (DSC) in Huang et al. [2007]; Huang and Chen [2006] which differs from the SI definition in Shan et al. [2012].

[5] Notice that Similarity Index (SI) is also used formulated as the Area Overlap (AOV)

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 and the relation between AOV or JSC and the DSC similarity indexes is expressed by equation 3. Notice that the DSC similiarity index is expected to be greater than the AOV index Pons et al. [2013]. The DSC metric has been used to assess the following works:Pons et al. [2013]; Huang et al. [2007]; Zhang et al. [2010]; Huang and Chen [2006]

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

$$DSC = \frac{2 \cdot AOV}{1 + AOV} \tag{3}$$

True-Positive Ratio (TPR), also known as the recall rate, sensitivity (at pixel level) Pons et al. [2013]; Jiang et al. [2012] or Overlap Fraction (OF) Huang et al. [2007], quantifies the amount of properly labeled pixels as lesion with respect to the amount of lesion pixels from the reference delineation (eq: 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: Madabhushi and Metaxas [2003]; Huang et al. [2007]; Shan et al. [2012]; Jiang et al. [2012]; Huang and Chen [2006]; Huang et al. [2012]; Liu et al. [2010b]; Yeh et al. [2009]

in Shan et al. [2012] which differs from the SI definition in Huang et al. [2007]; Huang and Chen [2006].

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{|M|} = \frac{|A| \wedge |M|}{|M|} \qquad \in [0, 1] \tag{4}$$

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 5. PPV is also greater than AOV. The PPV metric is also used to assess the work in Pons et al. [2013].

$$PPV = \frac{TP}{FP + TP} = \frac{TP}{|A|} = \frac{|A| \wedge |M|}{|A|} \qquad \in [0, 1] \tag{5}$$

Normalized Residual Value (NRV), also found as the Precission Ratio(PR) Huang and Chen [2004], corresponds to the area of disagreement between the two delineations regularized by the size of the reference delineation, as described in equation: 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: Gómez et al. [2010]; Liu and Huo [2005]; Huang and Chen [2004].

$$NRV = \frac{|A \oplus M|}{|M|} \qquad \in \left[0, 1 + \frac{A}{|M|}\right] \tag{6}$$

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 7. The FPR' metric has been used to assess the following works:Madabhushi and Metaxas [2003]; Shan et al. [2012]; Huang et al. [2012]; Liu et al. [2010b]; Yeh et al. [2009] The FPR' has also been found in its complementary form $1-TPR$ under the name of Match Rate (MR) Huang and Chen [2004].

$$FPR' = \frac{FP}{TP+FN} = \frac{FP}{|M|} = \frac{|A \vee M - M|}{|M|} \qquad \in \left[0, \frac{A}{|M|}\right] \tag{7}$$

Notice that the FPR' calculated in equation 7 differs from the classic False-Positive Ratio (FPR) obtained from the table in figure 6a, which corresponds to the ratio between FP and its column marginal ($FP + TN$), as indicated in equation 8. The FPR, when calculated according to equation 8, corresponds to the complement of specificity (described below).

$$FPR = \frac{FP}{FP+TN} = 1 - SPC \qquad \in [0,1] \tag{8}$$

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 9. Notice that it also corresponds to the inverse of the TPR since $TP \cup FN = M$. The FNR metric has been used to assess the following works: Madabhushi and Metaxas [2003]; Huang et al. [2012]; Yeh et al. [2009]

$$FNR = \frac{FN}{|M|} = \frac{|A \vee M - A|}{|M|} = 1 - TPR \qquad \in [0,1] \tag{9}$$

Specificity corresponds to the amount of background correctly labeled. Specificity is described in equation 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 8. The specificity index is also used to assess the work in Pons et al. [2013]; Jiang et al. [2012].

$$SPC = \frac{TN}{TN + FP} = \frac{|\overline{A} \wedge \overline{M}|}{|\overline{M}|} = 1 - FPR \qquad \in [0, 1] \tag{10}$$

3.1.2. 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: Alemán-Flores et al. [2007]; Gómez et al. [2010]; Gao et al. [2012]; Madabhushi and Metaxas [2003]; Shan et al. [2012]; Zhang et al. [2010]; Huang et al. [2012]. 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.

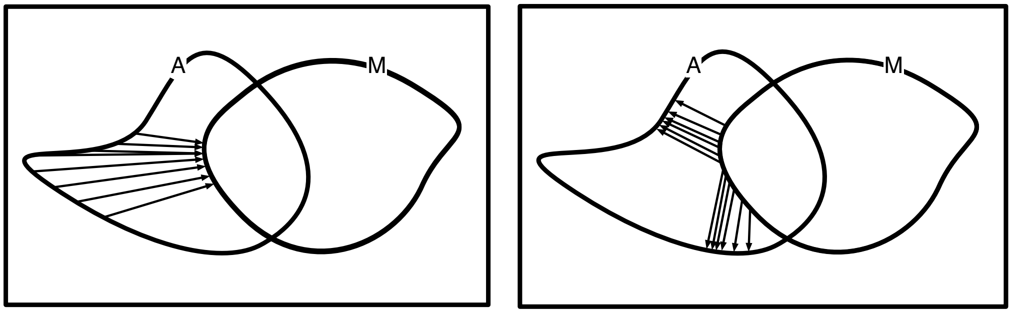
Zhang et al. Zhang et al. [2010] propose using average contour-to-contour distance (E_{cc}) for assessing their work. However, no definition or reference is found on it. Huang et al. Huang et al. [2012] propose using ARE, defined in equation 11, where a set of n radial rays are generated from the center 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 $|C_r(i) - C_0|$.

$$ARE = \frac{1}{n}\sum_{i=1}^{n} \frac{|C_s(i) - C_r(i)|}{|C_r(i) - C_0|} \tag{11}$$

The rest of the works base their similitude indexes on the analysis of the Minimum Distance (MD) coefficients. The MD is defined in equation 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.

$$\mathrm{MD}(a_i, M) = \min_{m_j \in M} \|a_i - m_j\| \tag{12}$$

Hausdorff Distance (HD),. or Hausdorff error, measures the worst possible discrepancy between the two delineations A and M as defined in 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 7. The HD as defined in equation 13 has been used for assessing the segmentation results in Gao et al. Gao et al. [2012]. Meanwhile, Madabhushi and Metaxas Madabhushi and Metaxas [2003] and Shan et al. Shan et al. [2012] only take into account the discrepancy between the assessed delineation A with reference delineation M, here denoted as HD' (see eq. 14). In Madabhushi and Metaxas [2003]; Shan et al. [2012], the HD' is also reported in a normalized form $\frac{\mathrm{HD}'}{\eta}$, where η is the length of the contour of reference M.



(a) (b)

Figure 7: Illustration of the non-symmetry property of the Minimum Distance (MD) metric. (a) MD(a_i, M), (b) MD(m_i, A)

$$\mathrm{HD}(A, M) = \max\left\{\max_{a_i \in A} \mathrm{MD}(a_i, M), \max_{m_i \in M} \mathrm{MD}(m_i, A)\right\} \tag{13}$$

$$\mathrm{HD'}(A, M) = \max_{a_i \in A} \mathrm{MD}(a_i, M) \tag{14}$$

Average Minimum Euclidian Distance (AMED),. defined in equation 15, is the average MD between the two outlines. Gao et al. [2012]. Similar to the case of the HD' distance, Madabhushi and Metaxas Madabhushi and Metaxas [2003] and Shan et al. Shan et al. [2012] only take into account the discrepancy between the assessed delineation A with reference to the delineation M to calculate the AMED' index (see eq. 16). The AMED index can be found under the name of Mean Error (ME) in Madabhushi and Metaxas [2003] and Mean absolute Distance (MD) in. Shan et al. [2012].

$$\mathrm{AMED}(A, M) = \frac{1}{2} \cdot \left[\frac{\sum_{a_i \in A} \mathrm{MD}(a_i, M)}{|A|} + \frac{\sum_{m_i \in M} \mathrm{MD}(m_i, A)}{|M|}\right] \tag{15}$$

$$\mathrm{AMED'}(A, M) = \frac{\sum_{a_i \in A} \mathrm{MD}(a_i, M)}{|A|} \tag{16}$$

Proportional Distance (PD),. used in Alemán-Flores et al. [2007]; Gómez et al. [2010], takes into account the AMED regularized with the area of the reference delineation according to equation 17

$$\mathrm{PD}(A, M) = \frac{1}{2\sqrt{\frac{Area(M)}{\pi}}} \cdot \left[\frac{\sum_{a_i \in A} \mathrm{MD}(a_i, M)}{|A|} + \frac{\sum_{m_i \in M} \mathrm{MD}(m_i, A)}{|M|} \right] * 100 \tag{17}$$

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. Pons et al. [2013] 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. Cui et al. [2009] 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. Yeh et al. [2009] 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. Alemán-Flores et al. [2007], 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. Massich et al. [2010] and Pons et al. Pons et al. [2013]. 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) estimation that has been inferred from the segmentations provided by the experts as a probability map. Massich et al. Massich et al. [2010] propose to assess the segmentation against a thresholded HGT and weight the AOV index with the HGT. The authors in Massich et al. [2010] 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.

4. Discussion

As has been said all along in section 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 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 Gómez et al. [2010], the standard deviation Alemán-Flores et al. [2007]; Cui et al. [2009]; Zhang et al. [2010]; Hao et al. [2012]; Huang et al. [2012]; Liu et al. [2010b], or median Alemán-Flores et al. [2007]; Hao et al. [2012]. Some other works prefer to report the distribution of their results graphically Massich et al. [2010]; Gao et al. [2012]; Yeh et al. [2009]. Finally, in Huang et al. [2007]; Shan et al. [2012], 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 1 and graphically displayed in figure 8 in order to analyze some trends. In table 1, the works presented are grouped depending on the user interaction according to the four categories described in section 1 corresponding to: 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 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 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 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 Pons et al. [2013] 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 8a rearranges the works presented according to the categories shown in figure 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 (Madabhushi and Metaxas [2003]; Cui et al. [2009]) have been chosen to belong to the combination of the two categories best describing the method. So, Madabhushi and Metaxas Madabhushi and Metaxas [2003] is treated as a combination of ML and ACM, and Cui et al. Cui et al. [2009] as an ACM and other methodology combinations. Figure 8b arranges the presented works according to the user interaction. Figure 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 (see fig:4) plus the unsupervised methods. Finally, Figure 8d represents the methodologies belonging to the ACM class, arranged by type (see fig:3 and section 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. Xiao et al. [2002] when quantitatively assessed in Pons et al. [2013], 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 8a and corroborated when comparing the areas of the plots in figures 8b and 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 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 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. Xiao et al. [2002] 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 Zhang et al. [2010]; Shan et al. [2012] 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. Gómez et al. [2010] proposed a methodology based on the popular GCS Horsch et al. [2001], 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. Massich et al. [2010] 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 Zhang et al. [2010] 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 Madabhushi and Metaxas [2003] 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 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 Liu et al. [2010b], 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.

5. references

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

Alemán-Flores, M., Álvarez, L., Caselles, V., 2007. Texture-oriented anisotropic filtering and geodesic active contours in breast tumor ultrasound segmentation. J Math Imaging Vis 28, 81–97.

Altman, D.G., Bland, J.M., 1994. Statistics notes: Diagnostic tests 2: predictive values. Bmj 309, 102.

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

Angelova, D., Mihaylova, L., 2011. Contour segmentation in 2d ultrasound

Table 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
Angelova and Mihaylova [2011]	20	1	∼	∼
Horsch et al. [2001]	400	1	AOV 0.73	73.0%
Gómez et al. [2010]	50	1	AOV 85%, NRV 16%, PD 6.5%	85.0%
Xiao et al. [2002]; Pons et al. [2013]	352	6	Sensitivity(TPR) 0.56, Specificity 0.99, PPV 0.73, AOV 0.51, DSC 0.61	50.8%
Pons et al. [2013]	352	6	Sensitivity(TPR) 0.61, Specificity 0.99, PPV 0.80, AOV 0.55, DSC 0.66	54.9%
Chiang et al. [2010]	16	1	∼	∼
Alemán-Flores et al. [2007]	32	4	AOV 0.88, PD 6.86%	88.3%
Cui et al. [2009]	488	2	AOV 0.73±0.14 AOV 0.74±0.14	74.5%
Gao et al. [2012]	20	1	TPR>0.91, FPR 0.04, JSC(AOV) 0.86, DSC 0.93, AMED 2pix., HD=7pix.	86.3%
Drukker et al. [2002]	757	1	Results reported as detection	∼
Massich et al. [2010]	25	7	AOV 0.64	64.0%
Madabhushi and Metaxas [2003]	42	1	FPR 0.20, FNR 0.25, TPR 0.75 ME(AMED') 6.6pix.	62.0%
Huang et al. [2007]	118		SI(DSC) 0.88 OF(TPR) 0.86	77.6%
Zhang et al. [2010]	347		AOV 0.84±0.1, ECC 3.75±2.85pix.	84.0%
Jiang et al. [2012]	112	1	∼	∼

Shan et al. [2012]	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%

Huang and Chen [2006]	20		SI(DSC) 0.88, OF(TRP) 0.81	78.6%
Huang et al. [2012]	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%
Liu and Huo [2005]	40	1	NRV 0.96 (benign); NRV 0.92 (malignant)	∼
Hao et al. [2012]	480	1	JSC(AOV) 0.75±0.17	75%
Huang and Chen [2004]	60	1	PR(NRV) 0.82, MR(FPR) 0.95	∼
Liu et al. [2010a]	112		Diagnosis results reported only	∼
Liu et al. [2010b]	76	1	TPR 0.94, FPR 0.07, AOV 0.88	88.1%
Yeh et al. [2009]	6	10	TPR>0.85, FNR<0.15, FP<0.16	73.3%

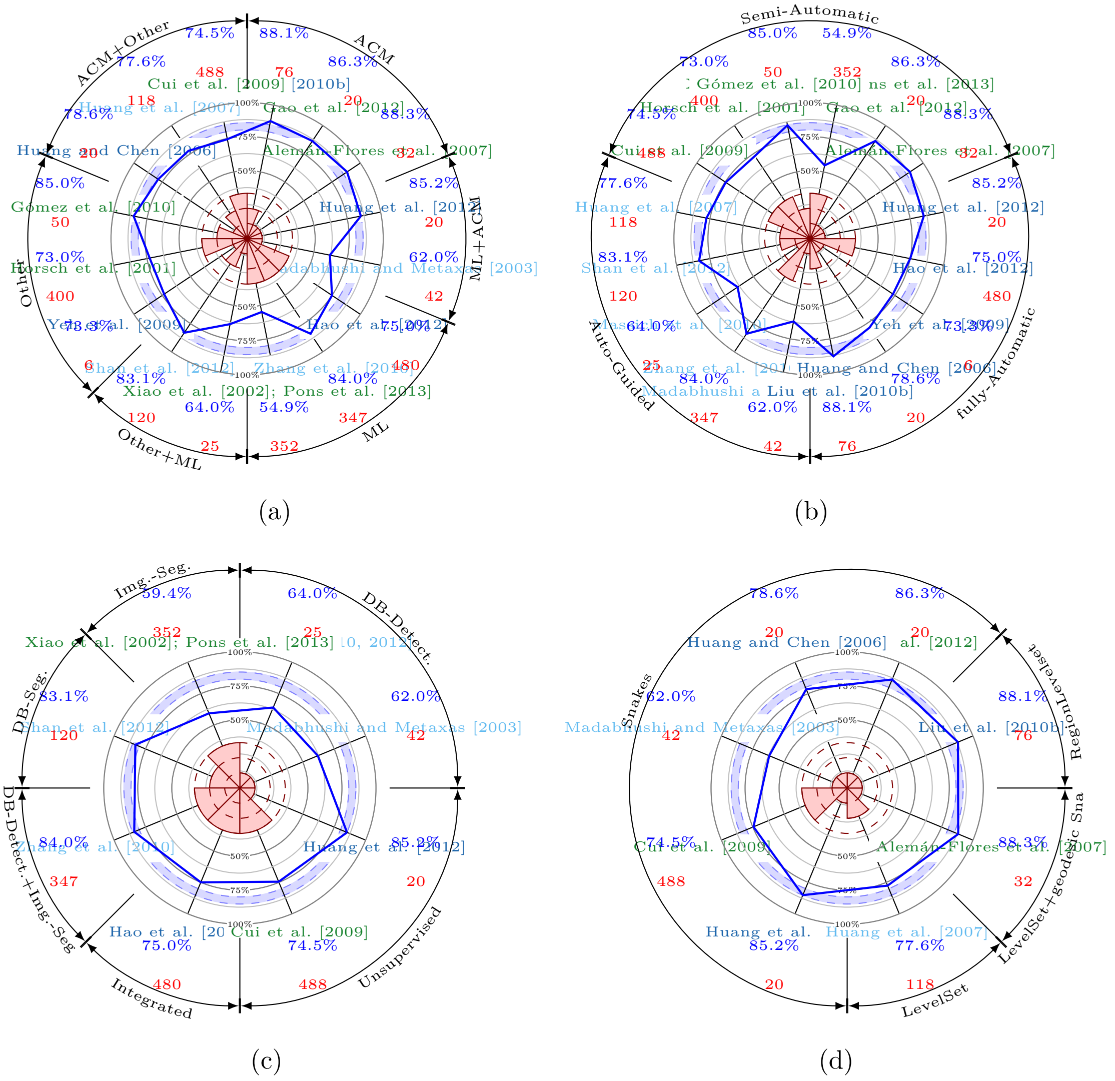


Figure 8: Graphical comparison of the methods presented that reported Area Overlap (AOV) or enough data to be inferred. The inner part of the plot illustrates the size of the dataset used in terms of small, medium, large. The blue swatch illustrates the inter- and intra-observer experiment results carried out in Pons et al. [2013]. The coloring of the reference indicates the user interacthability: semi-automatic (light-green), auto-guided(light-Blue), and fully automatic(dark-blue).

medical images with particle filtering. Machine Vision and Applications 22, 551–561.

Baker, J., Kornguth, P., Soo, M.S., Walsh, R., Mengoni, P., 1999. Sonography of solid breast lesions: observer variability of lesion description and assessment. AJR. American journal of roentgenology 172, 1621–1625.

Berg, W.A., Gutierrez, L., NessAiver, M.S., Carter, W.B., Bhargavan, M., Lewis, R.S., Ioffe, O.B., 2004. Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. Radiology 233, 830–849.

Beucher, S., et al., 1992. The watershed transformation applied to image segmentation. SCANNING MICROSCOPY-SUPPLEMENT- , 299–299.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23, 1222–1239.

Boykov, Y.Y., Jolly, M.P., 2001. 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. pp. 105–112.

Cheng, H.D., Shan, J., Ju, W., Guo, Y., Zhang, L., 2009. Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition 43, 299–317. doi:10.1016/j.patcog.2009.05.012.

Chiang, H.H., Cheng, J.Z., Hung, P.K., Liu, C.Y., Chung, C.H., Chen, C.M., 2010. 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. pp. 177–180.

Ciatto, S., Rosselli del Turco, M., Catarzi, S., Morrone, D., et al., 1994. The contribution of ultrasonography to the differential diagnosis of breast cancer. Neoplasma 41, 341.

Cohen, L.D., 1991. On active contour models and balloons. CVGIP: Image understanding 53, 211–218.

Cremers, D., Rousson, M., Deriche, R., 2007. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. International journal of computer vision 72, 195–215.

Cui, J., Sahiner, B., Chan, H.P., Nees, A., Paramagul, C., Hadjiiski, L.M., Zhou, C., Shi, J., 2009. A new automated method for the segmentation and characterization of breast masses on ultrasound images. Medical Physics 36, 1553.

Delong, A., Osokin, A., Isack, H.N., Boykov, Y., 2012. Fast approximate energy minimization with label costs. International Journal of Computer Vision 96, 1–27.

Drukker, K., Giger, M.L., Horsch, K., Kupinski, M.A., Vyborny, C.J., Mendelson, E.B., 2002. Computerized lesion detection on breast ultrasound. Medical Physics 29, 1438–46.

Ensminger, D., Stulen, F.B., 2008. Ultrasonics: Data, equations, and their practical usesĂŐ , 520.

Falcão, A.X., Udupa, J.K., Samarasekera, S., Sharma, S., Hirsch, B.E., Lotufo, R.d.A., 1998. User-steered image segmentation paradigms: Live wire and live lane. Graphical models and image processing 60, 233–260.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 1627–1645.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., Parkin, D.M., 2010. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International Journal of Cancer 127, 2893–2917.

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

Gao, L., Liu, X., Chen, W., 2012. Phase- and GVF-Based level set segmentation of ultrasonic breast tumors. Journal of Applied Mathematics 2012, 1–22.

Giger, M.L., Chan, H.P., Boone, J., 2008. Anniversary paper: History and status of CAD and quantitative image analysis: the role of medical physics and AAPM. Medical physics 35, 5799.

Gómez, W., Leija, L., Alvarenga, A.V., Infantosi, A.F.C., Pereira, W.C.A., 2010. Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation. Medical Physics 37, 82.

Gordon, P.B., Goldenberg, S.L., 1995. Malignant breast masses detected only by ultrasound. A retrospective review. Cancer 76, 626–630.

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

Hong, A.S., Rosen, E.L., Soo, M.S., Baker, J.A., 2005. BI-RADS for sonography: positive and negative predictive values of sonographic features. AJR Am J Roentgenol 184, 1260–5.

Horsch, K., Giger, M.L., Venta, L., Vyborny, C., 2001. Automatic segmentation of breast lesions on ultrasound. Medical Physics .

Horsch, K., Giger, M.L., Venta, L., Vyborny, C., 2002. Computerized diagnosis of breast lesions on ultrasound. Medical Physics .

Huang, Q.H., Lee, S.Y., Liu, L.Z., Lu, M.H., Jin, L.W., Li, A.H., 2012. A robust graph-based segmentation method for breast tumors in ultrasound images. Ultrasonics 52, 266–275.

Huang, Y.L., Chen, D.R., 2004. Watershed segmentation for breast tumor in 2-D sonography. Ultrasound in Medicine & Biology 30, 625–32.

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

Huang, Y.L., Jiang, Y.R., Chen, D.R., Moon, W.K., 2007. Level set contouring for breast tumor in sonography. Journal of digital imaging 20, 238–247.

Jalalian, A., Mashohor, S.B., Mahmud, H.R., Saripan, M.I.B., Ramli, A.R.B., Karasfi, B., 2012. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clinical imaging .

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., 2011. Global cancer statistics. CA: A Cancer Journal for Clinicians 61, 69–90.

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

Jumaat, A.K., Rahman, W.E.Z.W., Ibrahim, A., Mahmud, R., 2010. 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. pp. 505–509.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active contour models. International journal of computer vision 1, 321–331.

Kovesi, P., 1999. Image features from phase congruency. Videre: Journal of computer vision research 1, 1–26.

Kovesi, P., 2000. Phase congruency: A low-level image invariant. Psychological Research 64, 136–148.

Kupinski, M., Giger, M.L., 1998. Automated seeded lesion segmentation on digital mammograms. IEEE Transactions on medical imaging .

Li, Y., Sun, J., Tang, C.K., Shum, H.Y., 2004. Lazy snapping. ACM Transactions on Graphics (ToG) 23, 303–308.

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

Liu, B., Cheng, H.D., Huang, J., Tian, J., Tang, X., Liu, J., 2010b. Probability density difference-based active contour for ultrasound image segmentation. Pattern Recognition .

Liu, X., Huo, Z., 2005. Automated segmentation of breast lesions in ultrasound images , 1–3.

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

Lobregt, S., Viergever, M.A., 1995. A discrete dynamic contour model. Medical Imaging, IEEE Transactions on 14, 12–24.

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

Manning, D., Gale, A., Krupinski, E., 2005. Perception research in medical imaging. British journal of radiology 78, 683–685.

Massich, J., Meriaudeau, F., Pérez, E., Martí, R., Oliver, A., Martí, J., 2010. Lesion segmentation in breast sonography. Digital Mammography , 39–45.

Massich, J., Meriaudeau, F., Pérez, E., Martí, R., Oliver, A., Martí, J., 2011.

Seed selection criteria for breast lesion segmentation in ultra-sound images. MICCAI Workshop on Breast Image Analysis , 55–64.

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

Mendelson, E., Baum, J., WA, B., et al., 2003. 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.

Mendelson, E.B., Berg, W.A., Merritt, C.R., 2001. Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound, in: Seminars in roentgenology, Elsevier. pp. 217–225.

Mortensen, E.N., Barrett, W.A., 1998. Interactive segmentation with intelligent scissors. Graphical models and image processing 60, 349–384.

Najman, L., Schmitt, M., 1996. Geodesic saliency of watershed contours and hierarchical segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 18, 1163–1173.

Noble, J.A., Boukerroui, D., 2006. Ultrasound image segmentation: A survey. IEEE Transactions on medical imaging .

Noble, J.A., Wells, P.N.T., 2009. Ultrasound image segmentation and tissue characterization. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 224, 307–316.

Osher, S., Fedkiw, R., 2003. Level set methods and dynamic implicit surfaces. volume 153. Springer Verlag.

Otsu, N., 1975. A threshold selection method from gray-level histograms. Automatica 11, 23–27.

Pérez, P., Blake, A., Gangnet, M., 2001. Jetstream: Probabilistic contour extraction with particles, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, IEEE. pp. 524–531.

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

Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut: Interactive foreground extraction using iterated graph cuts, in: ACM Transactions on Graphics (TOG), ACM. pp. 309–314.

Sethian, J.A., 1996. A fast marching level set method for monotonically advancing fronts. Proceedings of the National Academy of Sciences 93, 1591–1595.

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

Shan, J., Cheng, H.D., Wang, Y., 2012. Completely automated segmentation approach for breast ultrasound images using multiple-domain features. Ultrasound in Medicine & Biology 38, 262–275.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22, 888–905.

Smith, R.A., Saslow, D., Sawyer, K.A., Burke, W., Costanza, M.E., Evans, W., Foster, R.S., Hendrick, E., Eyre, H.J., Sener, S., 2003. American cancer society guidelines for breast cancer screening: update 2003. CA: a cancer journal for clinicians 53, 141–169.

Stavros, A.T., 2004. Breast ultrasound. Lippincott Williams & Wilkins.

Stavros, A.T., Thickman, D., Rapp, C.L., Dennis, M.A., Parker, S.H., Sisney, G.A., 1995. Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions. Radiology 196, 123–34.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23, 903–921.

Xiao, G., Brady, M., Noble, J.A., Zhang, Y., 2002. Segmentation of ultrasound B-mode images with intensityinhomogeneity correction. IEEE Transactions on medical imaging 21, 48–57.

Xu, C., Prince, J.L., 1998. Snakes, shapes, and gradient vector flow. Image Processing, IEEE Transactions on 7, 359–369.

Yeh, C., Chen, Y., Fan, W., Liao, Y., 2009. A disk expansion segmentation method for ultrasonic breast lesions. Pattern Recognition .

Yuan, Y., Giger, M.L., Li, H., Bhooshan, N., Sennett, C.A., 2010. Multi-modality computer-aided breast cancer diagnosis with ffdm and dce-mri. Academic radiology 17, 1158.

Zhang, J., Zhou, S.K., Brunke, S., Lowery, C., Comaniciu, D., 2010. Database-guided breast tumor detection and segmentation in 2D ultrasound images, in: SPIE Medical Imaging, International Society for Optics and Photonics. pp. 762405–762405.