

Article

Mammographic Segmentation Using WaveCluster

Michael Barnathan

Living Discoveries, Morganville, NJ 07751, USA; E-Mail: michael@livingdiscoveries.com;
Tel.: +1-732-328-8268

Received: 2 June 2012; in revised form: 10 July 2012 / Accepted: 20 July 2012 /

Published: 10 August 2012

Abstract: Segmentation of clinically relevant regions from potentially noisy images represents a significant challenge in the field of mammography. We propose novel approaches based on the WaveCluster clustering algorithm for segmenting both the breast profile in the presence of significant acquisition noise and segmenting regions of interest (ROIs) within the breast. Using prior manual segmentations performed by domain experts as ground truth data, we apply our method to 150 film mammograms with significant acquisition noise from the University of South Florida's Digital Database for Screening Mammography. We then apply a similar segmentation procedure to detect the position and extent of suspicious regions of interest. Our approach was able to segment the breast profile from all 150 images, leaving minor residual noise adjacent to the breast in three. Performance on ROI extraction was also excellent, with 81% sensitivity and 0.96 false positives per image when measured against manually segmented ground truth ROIs. When not utilizing image morphology, our approach ran in linear time with the input size. These results highlight the potential of WaveCluster as a useful addition to the mammographic segmentation repertoire.

Keywords: segmentation; wavelets; clustering; WaveCluster; mammography

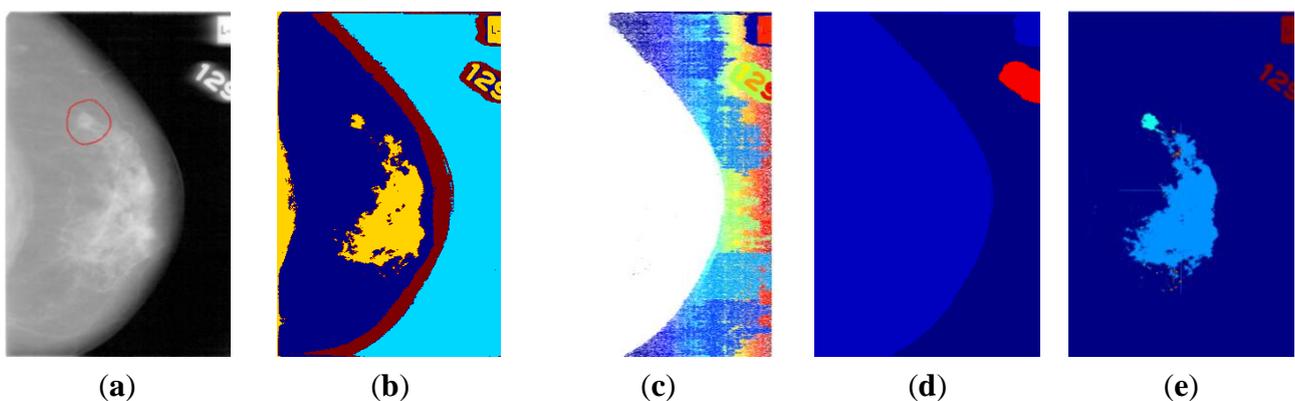
1. Introduction

Segmentation of clinically relevant regions of interest (ROIs) represents a significant challenge in the medical imaging domain, central to the performance of computer-assisted diagnostic systems. Within the context of screening mammography, segmentation is typically performed to extract and delineate possible lesions from the surrounding normal tissue. Additionally, although the transition to full-field digital mammography is ameliorating the need for mammographic background removal,

segmentation of the entire breast from a film mammogram with potentially significant acquisition noise represents a challenge in the many facilities which are as yet unable to transition to fully digital solutions, and in mining legacy data. Interestingly, these segmentation tasks are related: as we will show, the same approaches used to segment the entire breast profile are also useful for ROI extraction.

However, existing breast and ROI segmentation approaches, such as approaches based on k-means [1], fuzzy c-means [2,3], gray level co-occurrence matrices (GLCM) [4], and the watershed transform [5], suffer several shortcomings. For example, k-means and fuzzy c-means tend to find spherical clusters, k-means cannot differentiate between significant and nonsignificant clusters (*i.e.*, all pixels receive a label), GLCM makes assumptions on the texture of the underlying region, and watershed transforms have high sensitivity to noise and a well-known tendency to oversegment (See Figure 1). Many clustering algorithms are natively insensitive to locality, and may consider spatially dispersed regions part of the same cluster without the inclusion of additional spatial features—a property which is sometimes useful, but is often undesirable for the purpose of segmentation. Additionally, many clustering algorithms scale superlinearly with image area, and are therefore impractical for mining large medical imaging datasets. These properties typically render clustering algorithms unsuitable on their own for segmentation, necessitating construction of complex multi-step segmentation methodologies on top of them.

Figure 1. (a) Left craniocaudal view of a patient with a malignant mass, manually segmented by a domain expert (circled). Low-power salt and pepper acquisition noise, not visible at scale, corrupts the image; (b) k-means clustering ($k = 4$) identifies regions of interest (ROIs), but loses part of the image border and does not distinguish between important and unimportant ROIs (e.g. the pectoral muscle, the lesion, and normal tissue all fall into the same cluster). Increasing k simply resulted in empty clusters; (c) Watershed segmentation of the gradient identifies the overall contour of the breast, but fails to identify any regions within and oversegments the noisy background; (d) WaveCluster with a low density threshold (15%) segments the entire breast, while (e) WaveCluster with a higher threshold (85%) isolates and highlights the suspicious regions.



In this paper, we make use of the WaveCluster [6] clustering algorithm in the context of mammographic segmentation to address these challenges. WaveCluster has linear runtime with respect to image area and can provide a usable segmentation without additional preprocessing steps beyond

basic image normalization and noise removal. The algorithm is also versatile enough to segment both the breast profile and regions of interest within the breast simply by varying one parameter.

However, one major limitation of WaveCluster as initially proposed is its inability to cluster real-valued data, such as an intensity image—the algorithm as originally proposed can only operate on binary images [6]. Therefore, we make use of a novel “weighted WaveCluster” extension, which we have originally derived for use in fMRI imaging of the brain [7], which aggregates grid cell values via a weighted sum of neighborhood intensity values rather than a count of points. In the remainder of this paper, we demonstrate the efficacy of the approach in the domain of mammographic segmentation.

2. Background

2.1. Mammographic Segmentation

Though a great deal of work has been conducted in breast profile and ROI segmentation, both remain open challenges. While some model-driven supervised approaches exist [2,8], the majority of segmentation algorithms utilize semisupervised or unsupervised approaches, particularly region growing [8,9], contour detection [8,10], polygon fitting [10], and image clustering [1,2,3,8]. Our work falls into the clustering category and represents the first application of WaveCluster to the mammographic domain. Image morphology is commonly used as a preprocessing step to remove small regions of noise and homogenize local intensity variations, and is also used to supply connected component detection to many segmentation algorithms. A Hough transform is often used to remove the pectoral muscle from the image prior to analysis [11].

In all cases, the objective of segmentation is extraction of “interesting” regions of an image from the surrounding background, either for display or as a precursor to further region-restricted analysis.

2.2. WaveCluster

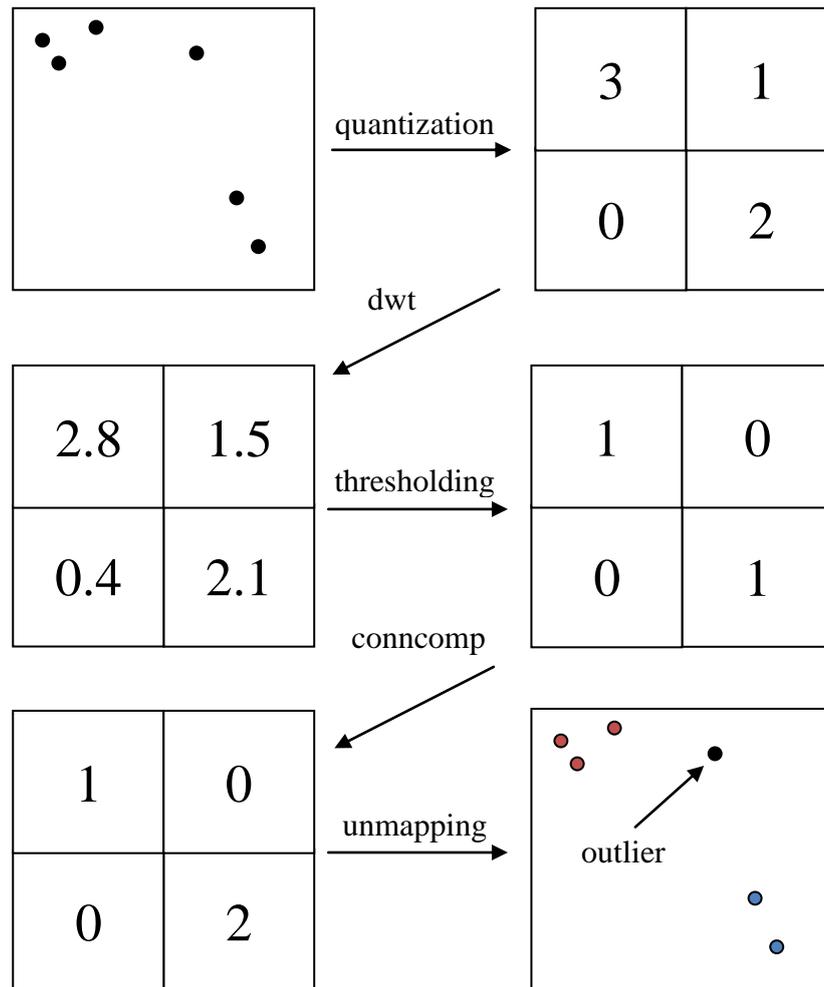
2.2.1. WaveCluster Algorithm

WaveCluster [6] is a grid and density-based clustering technique unique for performing clustering directly in wavelet space. This confers a number of advantages associated with wavelet transformations, including native multiresolution and multilevel analysis, linear runtime, and localization in space and time, rather than solely in frequency. The idea behind WaveCluster is as follows:

- Quantize data to a grid, using the cell counts in place of the original data.
- Apply a wavelet transformation using a hat-shaped wavelet (such as the 2.2 or 4.2 biorthogonal wavelets), retaining the approximation coefficients. These wavelets have the effect of emphasizing dense regions and suppressing sparse ones.
- Threshold cells in the transformed space. Cells with values above a user-specified density threshold are “significant”, while other cells are dropped.
- Apply a connected component algorithm to the significant cells to discover and label clusters.
- Map the cells back to the data using a lookup table built during quantization.

This approach is illustrated in Figure 2:

Figure 2. Illustrated step-by-step operation of the WaveCluster algorithm.



WaveCluster is notable for its robustness to noise, linear runtime, ability to discover complex shapes, cluster significance thresholding, and native multiresolution analysis. Unlike k-means, the number of clusters need not be specified. However, WaveCluster has several additional parameters: grid cell size, significance threshold, wavelet name, and wavelet level. An additional drawback of WaveCluster is its tendency to underestimate border regions (which makes it useful for outlier detection). As outlined in [7], we adopt a variant of WaveCluster replacing grid cell counts with a weighted sum of intensity values to extend to grayscale segmentation.

2.2.2. WaveCluster Parameters

As previously mentioned, the WaveCluster algorithm is parameterized by the number of cells in the grid (inversely proportional to cell size), the wavelet function used to perform the transformation, and the density threshold to apply in order to determine whether a cell is “significant” after performing the wavelet transform. The density threshold can be an absolute cell value, but is often more convenient to think about in terms of a percentile among cell values, since malignant regions tend to manifest as regions of localized density relative to surrounding tissue. Unlike k-means, WaveCluster is not parameterized by the number of clusters, which is determined automatically by the algorithm.

Varying the parameters has several effects:

- Increasing the number of cells increases the resolution of the analysis, allowing the algorithm to detect finer patterns but reducing the amount of neighborhood information which is incorporated in the clustering.
- Increasing the density threshold reduces the number of cells which are considered “significant”, resulting in fewer “bridges” between clusters and a smaller number of “stronger” but more isolated clusters. The optimal value of this parameter is application-dependent: too low and the entire image may be merged into one cluster, too high and regions of the same cluster will break apart. At extremely high values, no clusters are detected whatsoever.
- The wavelet function can be thought of as a convolution kernel applied to each pixel of the image. “Hat-shaped” wavelets are useful for clustering applications because they emphasize dense regions and suppress sparse ones, but more complex wavelet functions may be useful in niche applications, such as clustering highly anisotropic data.

3. Results and Discussion

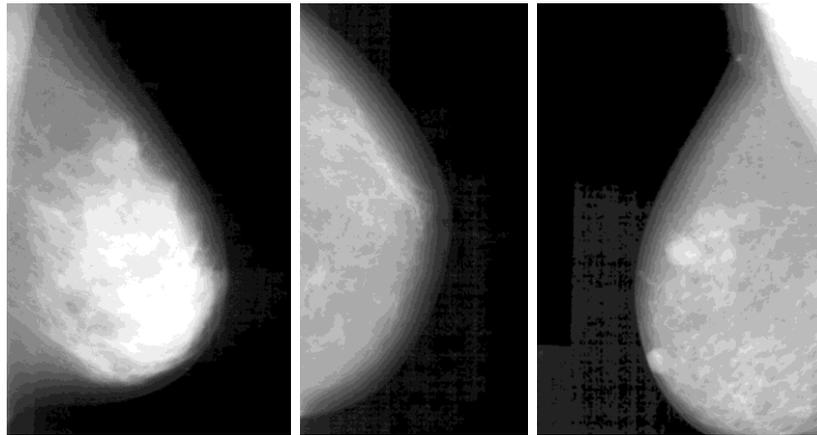
3.1. Dataset

We made use of a random sample of 100 malignant and 50 normal images from the public USF Digital Database for Screening Mammography [12]. Images are scanned film mammograms in CC and MLO views, acquired from 1991–1998 at a median resolution of 43.5 $\mu\text{m}/\text{pixel}$ (range 42–50 $\mu\text{m}/\text{pixel}$). Labels represent pathology-confirmed diagnoses and domain experts have manually traced all ROIs. Many of the images are noisy, and some have white borders and other artifacts of imperfect digitization.

3.2. Breast Profile Segmentation

The breast was successfully segmented in all 150 images. Three images (all malignant) contained a small amount of residual noise adjacent to the breast (see Figure 3); this number increased to 14 when the significance threshold was lowered to 5%, illustrating the importance of this parameter in the quality of the final segmentation. Conversely, increasing the significance threshold to 25% eliminated the noise; however, small holes began appearing in the segmented borders, as isolated cells near the edges of the image began to fall below the threshold. Minor adjustments to the density threshold therefore represent a tradeoff between the false dismissal and false alarm rates. Unfortunately, ground truth breast margins were not provided with the dataset, therefore we were unable to conduct an ROC analysis against this parameter.

Figure 3. The three noisy images. Due to the periodic texture of the noise, it is likely that a gray-level co-occurrence or entropy-based approach could effectively filter it.



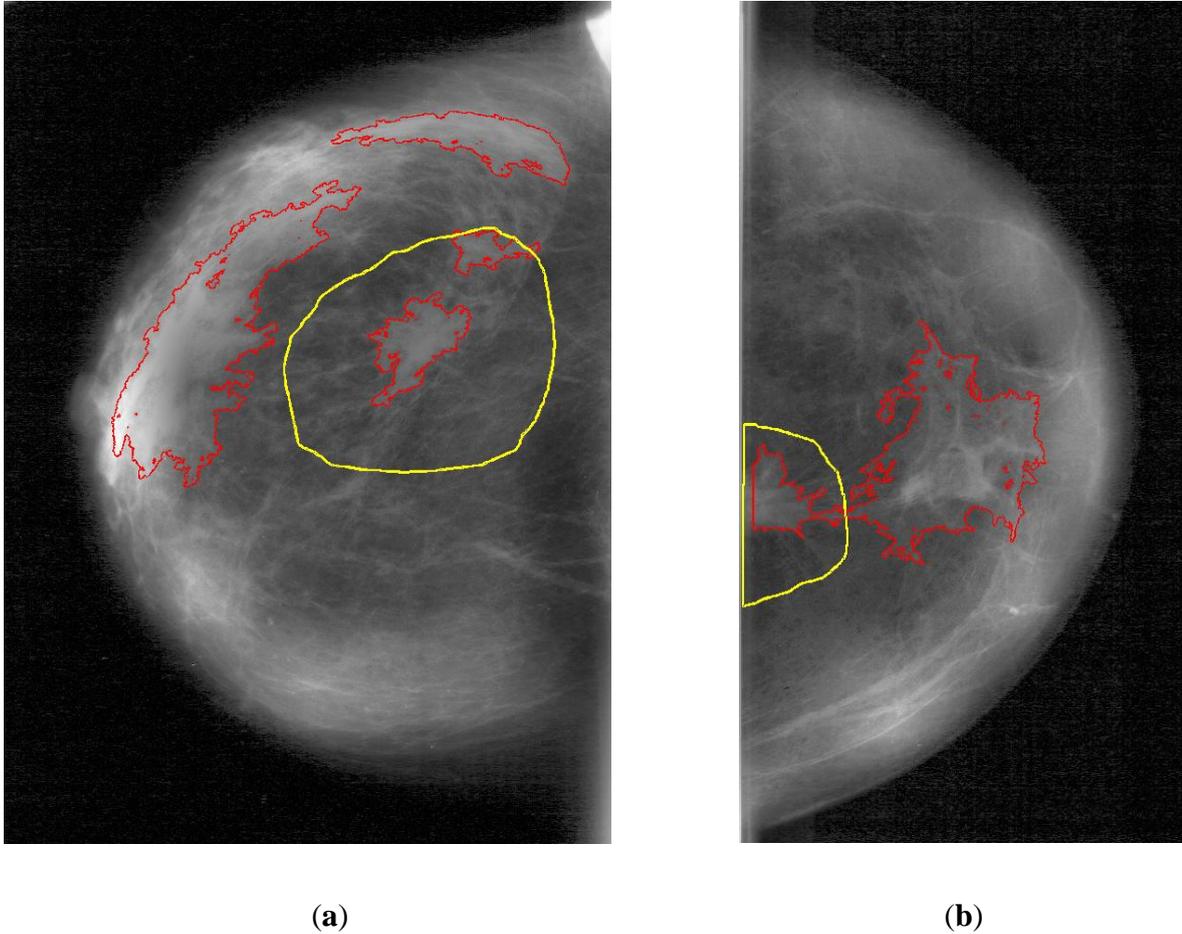
3.3. ROI Segmentation

By contrast, *significantly* increasing the density threshold permits us to discard the breast itself as background and segment out regions of interest within the breast. As the density requirements at a high threshold far exceed the density of background noise, morphological noise removal is no longer necessary for this procedure. This is best thought of as an algorithm for candidate selection or feature-space segmentation, as WaveCluster decides which regions are significant only through intensity and density information, and does not learn from other images in the dataset.

We re-ran the WaveCluster analysis with a density threshold of 85% on the malignant images. We then applied the mean absolute density threshold from the malignant images (a cell density of 14) to the 50 normal images in our dataset.

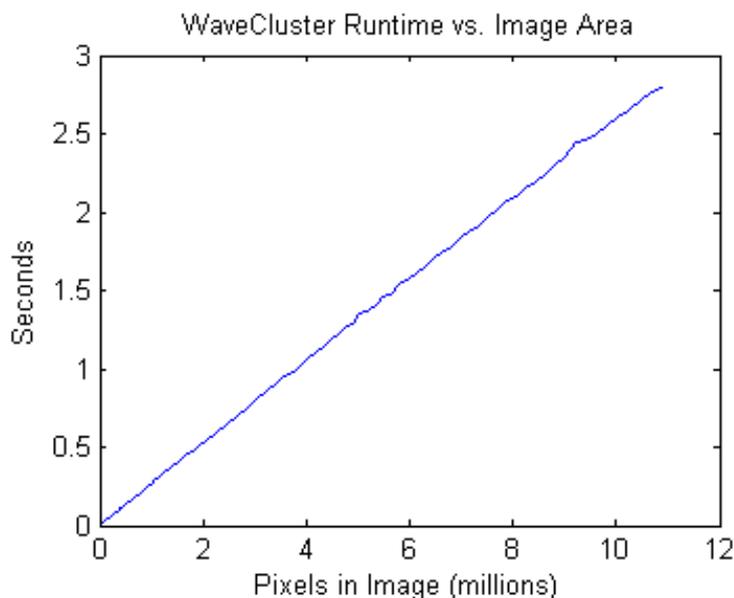
Ground truth data was available for all lesions, though margins were not precisely drawn. True positives were considered clusters which fell at least halfway within the bounds of a ground truth ROI. Because the ground truth data provided with the dataset was not precisely segmented, clusters smaller than the ROI were not penalized. Any ROI not overlapped by at least one cluster was considered a false negative. This resulted in a sensitivity of 81% with 0.96 false positives per image (89% and 0.88/image, respectively, if any overlap at all was considered sufficient for a true positive). The most common false positive by far was the pectoral muscle; with this region excluded, the FPR dropped to 0.7 per image. See Figure 4 for examples of sample false positives. Though the high sensitivity indicates that the majority of lesions were isolated and flagged significant by the clustering algorithm, it does not offer further guidance as to which cluster contains the malignancy.

Figure 4. Example false positive WaveCluster results overlaid on original images in red, ground truth in yellow. WaveCluster does not have a spherical clustering restriction; consequently, it can follow the contours of architectural distortions and spiculated lesions. **(a)** has two “false” positives along the contours of an architectural distortion, while **(b)** has one false positive at the bottom of the image and one FP resulting from oversegmentation of an actual lesion.



3.4. Performance

The mean image size was 4571×2339 pixels and the mean time to segment a single image (with morphological filtering) was 5.2 s. Consistent with WaveCluster’s theoretical complexity, the runtime of our approach (without morphological filtering) increased linearly with image area, as shown in Figure 5:

Figure 5. Runtime of WaveCluster as a function of image size.

3.5. Comparison

Similar segmentation approaches used in the literature include fuzzy c-means [2,3] and watershed segmentation [5]. These techniques were compared in accuracy and runtime, using the following experimental setup:

- WaveCluster was run with an 85% density threshold and the 2.2 biorthogonal wavelet.
- WaveCluster and watershed segmentation were evaluated by considering a cluster a true positive if and only if at least half of the cluster fell within the bounds of a ground truth ROI.
- Naïve watershed segmentation is prone to severe oversegmentation. In practice, watershed segmentation requires an h-minima transform as a preprocessing step. The h-minima transform removes all minima in the image of depth less than h . $h = 0.075$ was chosen empirically.
- Fuzzy c-means was run with $c = 5$. Since each pixel can have multiple cluster memberships, a cluster was considered a true positive if at least 50% of the *sum* of cluster weights came from pixels within the ground truth ROI.
- All methods were run with the morphological post-filtering described above. This particularly improved performance on the watershed segmentation approach.

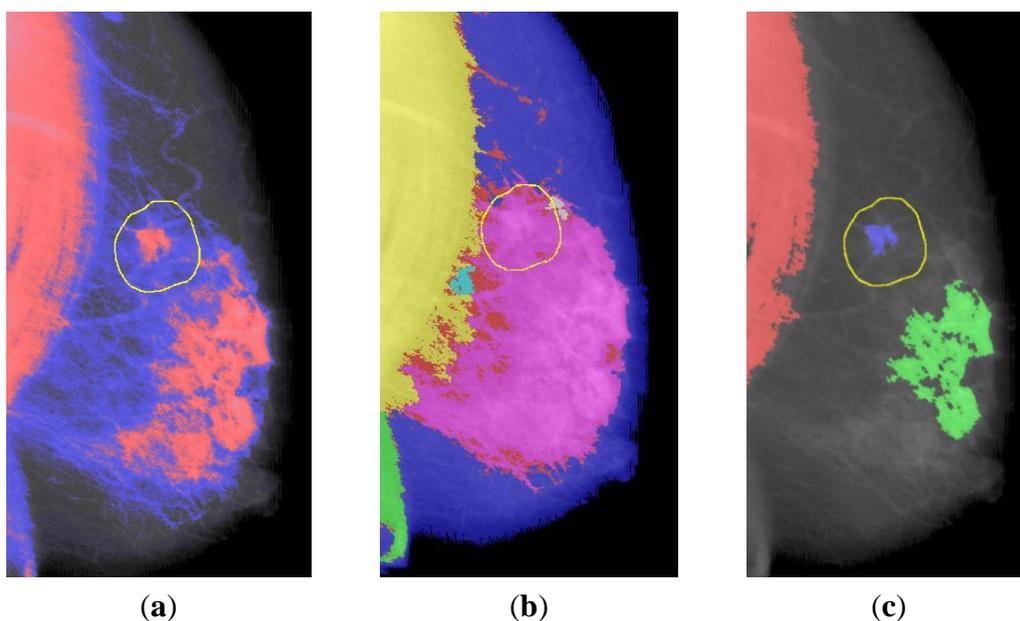
Comparative results on the ROI detection task are shown in Table 1:

Table 1. Comparison of sensitivity, false positive rate, and runtime for the WaveCluster, fuzzy c-means ($c = 5$), and Watershed segmentation approaches.

	WaveCluster	Fuzzy 5-means	Watershed
Sensitivity:	81%	23%/78%	33%
False Positives:	0.96/image	N/A	1.17/image
Average Runtime:	5.2 s/image	325 s/image	7.3 s/image

While fuzzy c-means performed excellently at detecting dense tissue (the application originally proposed in [2]), correctly stratifying regions of the breast by density and discovering a significant cluster containing an ROI in 78% of positive cases, it was found unsuitable for differentiating between normal and abnormal density because it did not incorporate locality. This is shown in Figure 6, and is also visible in Figure 1 (as k-means). When compared using a similar 50% (weighted) coverage metric to the metric used to assess other algorithms, it only correctly isolated 23% of true positive ROIs. This could be corrected in many cases by adding the thresholding and connected component steps inherent in WaveCluster to the fuzzy c-means approach, imposing an additional partitioning on the existing clusters. However, since this is not part of the original fuzzy c-means approach, it was not performed. Fuzzy c-means was also orders of magnitude slower than other approaches, due to its nature as an iterative algorithm with superlinear computational complexity.

Figure 6. Clustering a mammogram with high tissue density. (a) Fuzzy 5-means clusters 1–3 map to red, green, and blue (clusters 4 and 5 merely outlined the breast and are not shown). Cluster weights were combined additively to form RGB colors. While the lesion was identified, it was not distinguished from other dense tissue by the clustering algorithm; (b) 7 clusters were discovered by watershed segmentation, which fails to distinguish the ROI due to heavy oversegmentation; (c) WaveCluster also falsely identifies two clusters of dense tissue, but correctly isolates the malignancy because it takes locality into account.



4. Experimental Section

With the exception of variations in clustering parameters, our methods for segmenting the breast and the ROIs are identical, indicating the applicability of WaveCluster to more general mammographic segmentation problems.

4.1. Breast Profile Segmentation

Image intensities are first normalized to double precision values in the range (0,1) to mitigate inter-scanner differences. To reduce the substantial acquired noise present in the background region of the image, we apply a morphological opening with a 10-pixel disk structuring element followed by a 3×3 median filter.

We then map the image to a grid of (x, y, intensity) triples and quantize these to 10×10 cells to form the WaveCluster grid. Grid mappings are stored for later unmapping of clusters. We assign each grid cell a single value corresponding to a sum of the pixel intensities which mapped to it. We then perform a 2-dimensional level-1 2.2 biorthogonal stationary wavelet transformation on the grid cells, thresholding with a density threshold corresponding to the 15th percentile of all values in the grid. We chose a small threshold in this application to exclude the background while preserving the entire breast profile: as the breast is much denser than the surrounding background regions of the image, it is considered “more significant” by the algorithm. Thresholding against this 15th percentile threshold yields a “binary image” of grid cell values, which we analyze using a connected component algorithm (Matlab’s built-in `bwlabeln` algorithm) to yield distinct cluster labels for regions which are not connected. Finally, we propagate the cluster labels of each grid cell to all points which fell into that cell (using the original mapping we built up when constructing the grid) and return one cluster label per point. The largest significant cluster is assumed to be the breast, a reasonable assumption in our case. However, this assumption is not true for methods such as k-means, which do not have intrinsic cluster significance metrics and in which the largest cluster is usually the background.

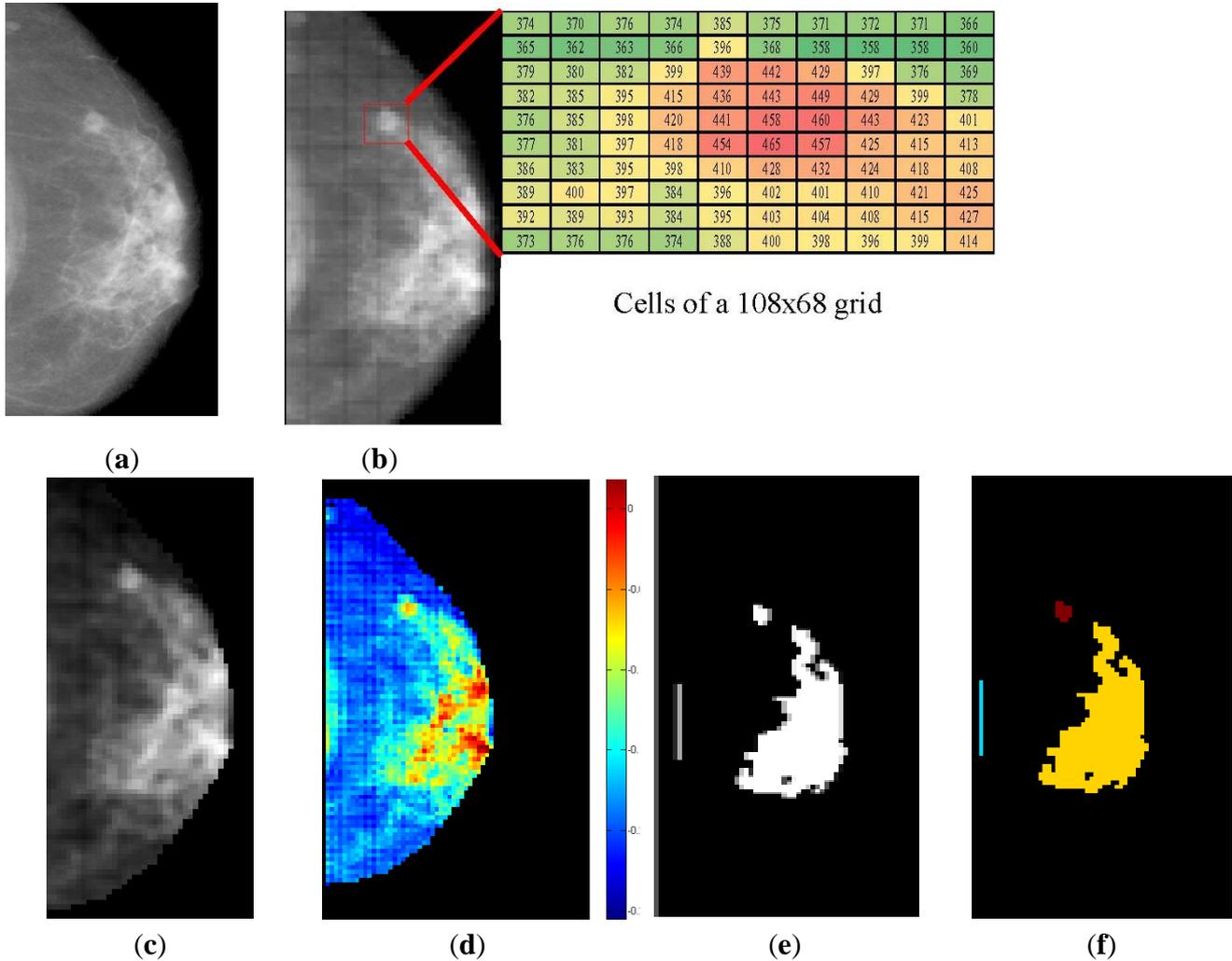
4.2. ROI Segmentation

The approach used to segment ROIs is identical with the exception of a few details:

- The density threshold is raised to the 85th percentile. This places the entire background and most of the breast beneath the significance threshold, yielding a clustering in which only the most dense regions *within* the breast are considered significant.
- Morphological noise filtering was not performed prior to clustering, as the low-power noise did not affect the significantly more dense regions of interest.
- Clusters <30 pixels in radius were discarded using morphological opening by reconstruction, which preserves the shape of the remaining clusters.

The steps of the WaveCluster algorithm used for ROI detection, shown theoretically in Figure 2, are illustrated on a mammogram in Figure 7:

Figure 7. (a) a 5386×3391 mammogram (b) quantized to a 108×68 grid by summing local regions of pixels and (c) run through a 1-level stationary wavelet transform using the 2.2 biorthogonal wavelet; (d) A difference map between the transformed grid and original grid; (e) An 80th percentile density threshold is applied, tiny clusters are eliminated through morphological opening; and (f) connected components are merged to form three distinct clusters.



5. Conclusions

We presented a novel mammographic segmentation methodology based on an extension of WaveCluster, demonstrating WaveCluster’s ability to segment the breast profile from a highly noisy dataset of film mammograms as well as to extract individual regions of interest from these images, with the application determined by the density threshold. This paper also represents one of the first practical demonstrations of WaveCluster’s linear computational complexity. Our results demonstrate the efficacy and speed of this algorithm for the first time in the domain of mammography, and emphasize the need for further study of underutilized clustering approaches in image segmentation.

References

1. Lee, S.; Lo, C.; Wang, C.; Chung, P.; Chang, C.; Yang, C.; Hsu, P. A computer-aided design mammography screening system for detection and classification of microcalcifications. *Int. J. Med. Inform.* **2000**, *60*, 29–57.
2. Oliver, A.; Freixenet, J.; Martí R.; Pont, J.; Pérez, E.; Denton, E.; Zwiggelaar, R. A novel breast tissue density classification methodology. *IEEE Trans. Inform. Technol. Biomed.* **2008**, *12*, 55–65.
3. Torrent, A.; Bardera, A.; Oliver, A.; Freixenet, J.; Boada, I.; Feixes, M.; Martí R.; Lladó, X.; Pont, J.; Pérez, E.; *et al.* Breast Density Segmentation: A Comparison of Clustering and Region Based Techniques. In *Proceedings of the 9th International Workshop on Digital Mammography*, Tucson, AZ, USA, 20–23 July 2008; pp. 9–16.
4. Khuzi, A.M.; Besar, R.; Zaki, W.W.; Ahmad, N.N. Identification of masses in digital mammogram using gray level co-occurrence matrices. *Biomed. Imaging Interv. J.* **2009**, *5*, doi:10.2349/bij.5.3.e17.
5. Alhadidi, B.; Zu'bi, M.H.; Suleiman, H.N. Mammogram breast cancer image detection using image processing functions. *Inf. Technol. J.* **2007**, *6*, 217–221.
6. Sheikholeslami, G.; Chatterjee, S.; Zhang, A. WaveCluster: A wavelet-Based clustering approach for spatial data. *VLDB J.* **2000**, *8*, 289–304.
7. Barnathan, M. Mining Complex High-Order Datasets. Dissertation. Ph.D. Thesis, Temple University, Philadelphia, PA, USA, April 2010.
8. Oliver, A. Automatic Mass Segmentation in Mammographic Images. Ph.D. Thesis, Universitat de Girona, Girona, Spain, July 2007.
9. Sholl, D.A. Dendritic organization in the neurons of the visual and motor cortices of the cat. *J. Anat.* **1953**, *87*, 387–406.
10. Guliato, D.; Rangayyan, R.M.; Carvalho, J.D.; Santiago, S.A. Polygonal modeling of contours of breast tumors with the preservation of spicules. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 14–20.
11. Nagi, J.; Abdul Kareem, S.; Nagi, F.; Khaleel Ahmed, S. Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms. In *Proceedings of 2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, Kuala Lumpur, Malaysia, 30 November–2 December 2010; pp. 87–92.
12. Heath, M.; Bowyer, K.; Kopans, D.; Moore, R.; Kegelmeyer, W.P. The Digital Database for Screening Mammography. In *Proceedings of the Fifth International Workshop on Digital Mammography*, Toronto, Canada, June 2000; Medical Physics Publishing: Madison, WI, USA, pp. 212–218.