Snakules for Automatic Classification of Candidate Spiculated Mass Locations on Mammography

Gautam S. Muralidhar, Mia K. Markey
The University of Texas Department of Biomedical Engineering, Austin, USA
{gautam, mia.markey}@mail.utexas.edu

Alan C. Bovik
Department of Electrical and Computer Engineering
The University of Texas at Austin, USA
bovik@ece.utexas.edu

***Abstract*—In this paper, we describe a novel approach for the automatic classification of candidate spiculated mass locations on mammography. Our approach is based on "*Snakules*" – an evidence-based active contour algorithm that we have recently developed for the annotation of spicules on mammography. We use *snakules* to extract features characteristic of spicules and spiculated masses, and use these features to classify whether a region of a mammogram contains a spiculated mass or not. The results from our initial classification experiment demonstrate the strong potential of *snakules* as an image analysis technique to extract features specific to spicules and spiculated masses, which can subsequently be used to distinguish true spiculated mass locations from non-lesion locations on a mammogram and improve the specificity of computer-aided detection (CADe) algorithms.**

Keywords- snakules; computer-aided detection; spiculated masses; snakes; active contours

I. INTRODUCTION

Breast cancer manifests as various findings on mammography – microcalcifications, masses (spiculated and non-spiculated), and architectural distortions. Spiculated masses are characterized by a pattern of radiating lines known as spicules that emanate from a central mass. Spiculated masses have a much higher risk of malignancy than non-spiculated masses and calcifications and, hence, it is crucial to detect spiculated masses [1]. CADe algorithms have been developed to assist radiologists in detecting signs of breast cancer [2]. Most CADe algorithms are comprised of two stages: a high sensitivity stage to detect suspect lesion locations on the mammogram and a high specificity stage to reduce the number of false positive (FP) candidates that do not correspond to actual lesions. The final outcome of a CADe algorithm is usually a set of marks (also referred to as prompts) on the mammogram identifying suspect lesion locations.

CADe algorithms designed to detect spiculated masses usually employ strategies to detect radial patterns of converging lines (e.g., [3, 4]). While these algorithms deliver high sensitivity in detection of spiculated masses, they invariably suffer from a high false positive (FP) rate, which reduces the specificity of the algorithm. The high FP rate is mainly attributed to the fact that a mammogram contains other normal linear structures that are superimposed on one another and resemble a pattern of converging lines. Such locations are routinely marked as suspect locations by the detection algorithms. Consequently, studies have shown that the detection performance of most CADe algorithms on spiculated masses is not optimal (e.g., [5]).

The performance of CADe systems on spiculated masses can be improved by developing sophisticated image analysis techniques to extract physical properties specific to spicules. These physical properties can then be used in the reliable classification of suspect locations identified on the image as lesion or non-lesion. Towards this goal, we have developed "*snakules*" [6], an algorithm that employs parametric open-ended snakes (active contours) to annotate spicules on mammography. The main contribution of this paper lies in demonstrating the potential of *snakules* to extract features characteristic of spicules and spiculated masses that can be used for automatic classification of candidate spiculated mass locations on mammography.

It is important to note that other groups have tried to extract properties specific to spicules to improve specificity of CADe algorithms. Zwiggelaar *et al.* demonstrated the use of cross-sectional intensity profiles as a basis for classifying linear structures seen on a mammogram with particular emphasis on correctly recognizing spicules and ducts [7]. However, Zwiggelaar *et al.* do not explicitly seek to capture the spicules; rather, they collect cross-sectional profile information from each linear structure detected on a mammogram using line detection operators and classify them into anatomical types by using a classifier trained on ground truth and cross-sectional information. A few other studies have attempted to deploy snake-like devices to segment solid masses and use feature extraction strategies to classify the masses (e.g., [8]); however, our approach is the first attempt to explicitly capture spicules in a bid to model their physical properties and use them for classification of putative spiculated masses.

The remainder of this paper is organized as follows: In Section II, we provide a brief description of the *snakules* algorithm. This is then followed by a detailed description of our feature extraction process and experimental methodology. In Section III, we discuss our results, and this is followed by conclusion and pointers to future work in Section IV.

II. PROPOSED METHOD

A. Snakules Algorithm

We provide a brief description of the *snakules* algorithm [6] in this section. *Snakules* is a set of parametric open-ended

snakes [9] that seek spicules on mammography. These snakes are automatically initialized in the region around a suspect spiculated mass location identified by a CADe algorithm or a radiologist. The points in this region from where the snakes originate represent the points in the region from where the spicules most likely originate. A set C of such candidate points is determined using the following equation –

$$C = \left\{ \begin{array}{l} (x,y) \ni (x,y) \in N, \left| \theta_{x,y} - \psi_{x,y} \right| < R / r_{x,y}, \psi_{x,y} \in k, \\ \forall (p,q) \in N, \left| \theta_{p,q} - \psi_{p,q} \right| < R / r_{p,q}, \psi_{p,q} \in k : r_{x,y} \le r_{p,q} \end{array} \right\} \quad (1)$$

where (x,y) is a candidate *snakule* point, N represents a neighborhood of pixels under consideration around the suspect spiculated mass location (x_c, y_c), $\theta_{x,y}$ is the dominant pixel orientation at location (x,y), $\psi_{x,y}$ is the direction of the location (x,y) with respect to (x_c, y_c) and is computed as $\psi_{x,y} = \tan^{-1}\left(\left(y_c - y \right) / \left(x_c - x \right) \right)$, R is the radius of a circular disk centered on (x_c, y_c), and towards which the pixel at location (x,y) is directed, $r_{x,y}$ is the Euclidean distance between (x,y) and (x_c, y_c), and k represents the k^{th} orientation bin. The condition $\left| \theta_{x,y} - \psi_{x,y} \right| < R / r_{x,y}$ is the same as defined in [3], in that we consider pixels in a neighborhood around the suspect spiculated mass location that are directed towards a circular disk of radius R centered on the suspect spiculated mass location. The condition on the second line of (1), ensures that of all the pixels that are directed towards the central mass region and whose directions with respect to the suspect spiculated mass location fall in the same orientation bin k, only the point that is closest to the suspect spiculated mass location will be selected as a candidate point.

The detection of candidate *snakule* points is carried out on *steerable filtered-Radon enhanced* regions of interest (ROIs) rather than on the ROIs cropped directly from the mammograms. Radon enhancement of spiculated lesions on mammograms is explained in detail in [4]. Radon enhancement is performed to enhance linear structures in the mammogram and to reduce the effects of noise and clutter caused due to overlapping out-of-plane tissue structures. The dominant orientation θ at each pixel of the Radon enhanced ROI is computed by filtering the Radon enhanced ROI with a set of steerable quadrature filter pairs comprised of the fourth derivative of a Gaussian and its Hilbert transform [10]. The choice of the parameters R and N, and the number of orientations bins considered around the suspect spiculated mass location are based on the average mass radius, range of spicule length, and the number of spicules respectively that have been computed from measurements of spiculated masses collected on mammograms [11].

Once the candidate points are identified, *snakules* are initialized at these points. We adopt an approach in which the snakes evolve and grow iteratively until a stopping criterion is met. This is motivated by the fact that the true length of a spicule is not known beforehand. The idea of growing snakes was first described by Berger [12] as a robust alternative to dropping long snakes to trace open-ended curvilinear structures in images and we adopt a similar strategy with a curvature-based stopping criterion. The evolution of the snakes is governed by the standard Euler-Lagrange force balance equation [13] and the vector field convolution (VFC) force [13] is used as the external force. The VFC force is computed as the convolution of a user-defined vector field kernel with a feature map (e.g., edge map) generated from the image and is shown to be robust to spurious edges and noise in the image and provides a large capture range. Instead of using a standard edge map as a feature map, we use the *Radon enhanced ROI* as the feature map. The Radon enhanced ROI has the property that it is non-negative and has a larger value near the enhanced curvilinear structures and hence these curvilinear structures contribute more to the VFC force than the homogeneous regions of the ROI. Fig. 1 illustrates an example of a spiculated mass ROI annotated using *snakules*.

B. Feature Extraction Process

The goal of the feature extraction process is to identify the best discriminatory features that characterize spicules and spiculated masses. Towards this goal, we have investigated the following features –

1) Average *Snakule* Contrast (f_1): We define *snakule* contrast C_S as follows –

$$C_S = \sqrt{\frac{1}{(N-1)} \sum_{\forall (x,y) \in S} \left(I_{(x,y)} - \mu \right)^2} \quad (2)$$

where S denotes a snakule, (x,y) denotes the coordinates of a point through which the snakule passes, $I_{(x,y)}$ denotes the interpolated image intensity at the point (x,y), N is the number of points through which the *snakule* passes, and μ is the average interpolated image intensity of the *snakule* trajectory (I_S) and two background trajectories (I_{B_1}, I_{B_2})

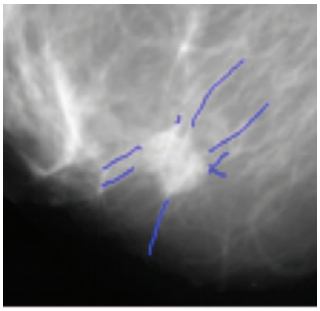

Figure 1. Example of a spiculated mass ROI annotated using *snakules*.

defined as $\mu = \frac{1}{3N}\left[I_S + I_{B_1} + I_{B_2} \right]$. The two background trajectories are computed at a distance marginally greater than (by approximately 4 pixels) half the width of the structure annotated by the *snakule* in the directions of the inward and outward normal at every point on the *snakule*. Fig. 2 illustrates the *snakule* and background trajectories. The average *snakule* contrast is computed as the average contrast over all *snakules* identified on the ROI. The intuition behind using average *snakule* contrast as a feature is that visually perceivable spicules are normally perceived as bright structures with good contrast relative to the background.

2) Median Distance from Point of Convergence (f_2): Convergence of spicules is a characteristic feature of spiculated masses. We seek to estimate a point of convergence (P_c) in the ROI space and subsequently compute the median distance of all linear structures annotated by *snakules* from P_c. Our hypothesis is that when an ROI actually contains a spiculated mass, then the median distance of linear structures from the point of convergence will be lower than the median distance from the point of convergence of linear structures annotated on an ROI not containing a spiculated mass. The problem of finding P_c can be posed as an optimization problem in which a point that is at a minimum median distance from all linear structures annotated in the ROI space is sought. We employ a greedy search strategy over the entire ROI space to find the point of convergence. Essentially, for every integer pixel location of the ROI, we compute the distance between the pixel location and a linear fit of each snakule. The point whose median distance from all lines is minimum is designated as the point of convergence and the minimum median distance is used as a feature for classification. It is important to note that the point of convergence can be computed when there are at least two linear structures annotated by *snakules*. If this is not the case, then the median distance is assigned a bookkeeping value of -1 for that particular ROI. Fig. 3 illustrates the point of convergence (denoted by a red '+') on a spiculated mass ROI annotated using *snakules*.

3) Histogram of Normalized Squared-Intensity Deviation (f_3): Features f_1 and f_2 are global features, in that a single value is assigned to a set of snakule annotations. We hypothesize that the discriminatory power of a classifier will be increased if the entire histogram of a feature is included along with the global features. This intuition is based on numerous studies in computer vision that utilize histograms of features of interest for the tasks of object recognition and classification (e.g.,[14]). Going by the intuition that visually perceivable spicules have good contrast relative to the background, we build a histogram of the normalized squared-deviation of the interpolated image intensity from the average intensity μ (defined earlier) at every point on the trajectory of the *snakule*. For every *snakule*, the squared-intensity deviation is normalized with respect to the maximum squared-intensity deviation for that snakule. We build the histogram using the k-means vector quantization technique (VQ) in which the centers of the VQ codebook are

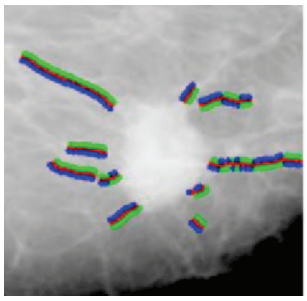

Figure 2. *Snakule* trajectory (red) and background trajectories (blue and green) on a spiculated mass ROI.

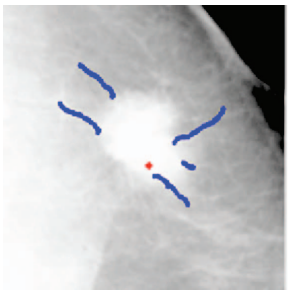

Figure 3. Point of convergence (red '+') illustrated on a spiculated mass ROI.

computed using k-means clustering of squared-intensity deviation values accumulated from a training set of ROIs containing both positive (ROIs with spiculated mass) and negative (ROIs with no spiculated mass) instances.

C. Experimental Methodology

The dataset for our study consisted of a total of 36 mediolateral view spiculated mass mammograms retrieved from the Digital Database for Screening Mammography (DDSM) [15]. A CADe algorithm previously developed in our group [4] was used to screen these mammograms and a set of true lesion and FP locations for each mammogram was output by the algorithm. Out of the true lesion locations, we only considered those locations deemed as most probable lesion locations by the CADe algorithm. An ROI was centered on each of these locations and was cropped from the mammogram. The final dataset consisted of a total of 312 ROIs with 36 positive instances and 276 negative instances. Out of the 36 mammograms, we reserved 18 randomly chosen mammograms for training the classifier and the remaining 18 mammograms for testing the classifier. This resulted in our training set comprising of 149 ROIs (18 positive instances, 131 negative instances) and testing set comprising of 163 ROIs (18 positive instances, 145 negative instances) to yield a total of 312 ROIs. *Snakules* were deployed and the features f_1, f_2, and f_3 were extracted from each ROI. For the feature f_3, the number of k-means clusters to build a histogram was varied between 3-10. We trained a

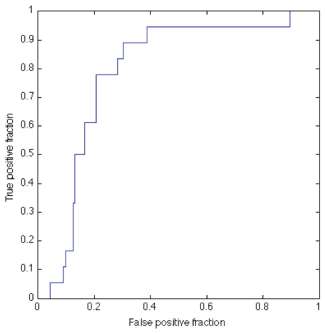

Figure 4. ROC curve of classification experiment.

a logistic regression classifier using the features extracted from each ROI in the training set. Since the ratio of the number of positive instances to the number of negative instances in our training set was much lower than 0.5, we sampled the positive instances with replacement to increase the ratio to 0.5 prior to training the classifier. The classifier was then tested using the features extracted from each ROI in the testing set and the performance of the classifier was evaluated using the area under curve (AUC) metric of the Receiver Operating Characteristic (ROC) curve.

III. Results

The best classification performance on the testing set was obtained when the number of k-means clusters to build the histogram (f_3) was set to 7. Fig. 4 illustrates the corresponding ROC curve obtained for the classification experiment ($AUC = 0.79 \pm 0.05$). The results of the classification experiment are promising and suggest that *snakules* can be used to extract features specific to spicules for the automatic classification of putative spiculated masses.

IV. Conclusion And Future Work

Reliable classification of spiculated masses is an important problem in the context of improving the specificity of CADe algorithms. To address this problem, we have developed a novel algorithm called *snakules* that explicitly tries to seek spicules on mammography. *Snakules* could prove to be a valuable device for extracting features specific to spicules for the classification of candidate spiculated mass locations. In this paper, we have presented our initial work on extraction of features specific to spicules for the classification of putative spiculated masses. Our results demonstrate the strong potential of *snakules* as a device that can be used for classifying candidate spiculated mass locations.

As part of future work we plan to investigate the use of s*nakules* to extract other physical properties specific to spicules such as spicule length and width. Our ultimate goal is to use *snakules* to construct reliable statistical models of spicule properties that can be used for classifying candidate spiculated mass locations. Finally, we intend to integrate *snakules* with a CADe algorithm previously developed in our group [4] and test the performance of the algorithm on a larger dataset of spiculated masses.

Acknowledgment

This work was supported in part by an Early Career Award from the Wallace H. Coulter Foundation and in part by an award from the Texas Ignition Fund.

References

[1] L. Liberman, A. F. Abramson, F. B. Squires, J. R. Glassman, E. A. Morris, and D. D. Dershaw, "The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories," AJR, vol. 171, pp. 35-40, 1998.

[2] M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-aided detection and diagnosis in mammography," in Handbook of Image and Video Processing, 2nd ed, A. C. Bovik, Ed.: Academic Press, 2005, pp. 1195-1217.

[3] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammograms," IEEE Trans. Medical Imaging, vol. 15, pp. 611-619, 1996.

[4] M. P. Sampat, A.C. Bovik, G.J. Whitman, M.K. Markey, "A model-based framework for the detection of spiculated masses on mammography," Med. Phys., vol. 35, pp. 2110-2123, 2008.

[5] S. K. Yang et al., "Screening mammography–detected cancers: sensitivity of a computer-aided detection system applied to full-field digital mammograms," Radiology, vol. 244, pp. 104-111, 2007.

[6] G. S. Muralidhar, A. C. Bovik et al., "Snakules: An evidence-based active contour algorithm for the annotation of spicules on mammography," unpublished.

[7] R. Zwiggelaar, S. M. Astley, C. R. M. Boggis, and C. J. Taylor, "Linear structures in mammographic images: detection and classification," IEEE Trans. Medical Imaging., vol. 23, p. 1077, 2004.

[8] Y. Yuan, M. L. Giger, H. Li, K. Suzuki, and C. Sennett, "A dual-stage method for lesion segmentation on digital mammograms," Med. Phys., vol. 34, pp. 4180-93, 2007.

[9] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," IJCV, vol. 1, pp. 321-331, 1987.

[10] W. T. Freeman and E. H. Adelson, "The Design and Use of Steerable Filters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, pp. 891-906, 1991.

[11] M. P. Sampat, G. J. Whitman, T. W. Stephens, L. D. Broemeling, N. A. Heger, A. C. Bovik, and M. K. Markey, "The reliability of measuring physical characteristics of spiculated masses on mammography," BJR, vol. 79, pp. S134-S140, 2006.

[12] M. O. Berger, "Snake growing," Proc. ECCV, Antibes, France, 1990, pp. 570-572.

[13] B. Li and S. T. Acton, "Active contour external force using vector field convolution for image segmentation," IEEE Trans. Image Processing, vol. 16, pp. 2096-2106, 2007.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in CVPR, 2005, pp. 886-893.

[15] M. Heath, K. W. Bowyer, and D. Kopans, "Current status of the Digital Database for Screening Mammography," in Digital Mammography, Kluwer Academic Publishers, 1998, pp. 457-460.