BREAST CANCER DETECTION USING IMAGE PROCESSING TECHNIQUES

Tobias Christian Cahoon[A]
Melanie A. Sutton[B]
James C. Bezdek
Department of Computer Science
University of West Florida
Pensacola, FL 32514

Abstract – We describe the use of segmentation with fuzzy models and classification by the crisp *k-nearest neighbor* (k-nn) algorithm for assisting breast cancer detection in digital mammograms. Our research utilizes images from the *Digital Database for Screening Mammography* (DDSM). We show that supervised and unsupervised methods of segmentation, such as k-nn and *fuzzy c-means* (FCM), in digital mammograms will have high misclassification rates when only intensity is used as the discriminating feature. Adding window means and standard deviations to the feature suite (visually) improves segmentations produced by the k-nn rule. While our results are encouraging, other methods are needed to detect smaller pathologies such as microcalcifications.

Key Words: breast cancer, classification, computer assisted diagnosis, digital mammography, FCM, fuzzy models, k-nn, segmentation, supervised, unsupervised.

1. INTRODUCTION

The only current means of early detection of breast cancer is through regular mammography screening. Through mammogram analysis radiologists have a detection rate of 76%-94%, which is considerably higher than the 57%-70% detection rate for a clinical breast examination [1]. Additionally, through mammography radiologists are able to correctly identify that a woman does not have breast cancer in over 90% of all cases [1]. This practice has greatly increased the diagnosis of early-stage tumors that previously eluded physicians until the cancer had reached deadlier stages. Despite this, mammographic screening may miss 10%-15% of breast cancers [2]. Through CAD (*computer assisted diagnosis*) we hope to reduce the error rate for the detection of cancers, by providing a tireless, consistent, and undistracted second observer. According to Kopans, double reading systems such as the one we are developing may help reduce the error rate by 5%-15% [3]. At least one rudimentary CAD system developed by Kegelmeyer et al. for mammography screening is already commercially available, (see [4]).

Many electronic databases specific to research on mammographic images are available. However, images from these databases differ in their resolution, sensor basis, ground truth information, and quality. Our research group has chosen to work with images from one such database: the Digital Database for Screening Mammography (DDSM) [5]. The DDSM database is the largest and most comprehensive mammographic database, currently containing approximately 2500 cases as of September, 1999. DDSM images have grey-level resolutions of 12 or 16 bits, as well as spatial resolutions as high as 5000 x 2000 pixels, depending on the scanning device.

2. TWO APPROACHES TO SEGMENTATION

Fig. 1 depicts the two approaches used in our system. Traditionally, methods for image segmentation are classified into one of two groups, supervised or unsupervised. The labels of the data (and their use during training) are the distinguishing factor. Supervised methods such as the crisp *k-nearest neighbor* (k-nn) algorithm [6] use physical labels of tissue classes selected by an operator from regions within the image prior to segmentation. These *training pixels* (X_{tr}) are then used to classify the remaining unlabeled *test pixels* (X_{te}). Unsupervised methods such as *fuzzy c-means* (FCM) [7] rely on the human to label each tissue class appropriately once the algorithm has clustered *all* of the pixels in the image ($X=X_{tr} \cup X_{te}$) into c tissue classes. See Bezdek and Sutton [8] for an extensive survey of many fuzzy models used in one or both of these approaches. Notice that both paths to the segmented image require human intervention for *each* input image. The ultimate goal of mammographic CAD systems is to eliminate dependence of the second-reader system on clinicians, which is costly, time-consuming, and subjective.

The unsupervised track in our system segments digital mammograms with FCM. FCM segments the breast tissue into c homogeneous (but possibly disconnected) regions. An example from the DDSM data set and its FCM segmented image based on a single feature (intensity) is shown in Figs. 2 and 3 respectively. Fig. 2 is a 214 x 143 8-bit image taken from the DDSM data set. Therefore,

[A] Research supported by a Council on Undergraduate Research Summer Research Fellowship.
[B] Research supported by a Whitaker Foundation Biomedical Engineering Research Grant.

0-7803-5877-5/00/$10.00 © 2000 IEEE

$|X|$=34,463. X is *not* subdivided into training (X_{tr}) and test (X_{te}) sets before unsupervised segmentation. Parameters of FCM for this and subsequent runs are: weighting exponent m=2; Euclidean norm for J_m; termination criterion $\|\mathbf{V}^{new} - \mathbf{V}^{old}\| \leq \varepsilon = 0.5$, where **V** represents the set of c cluster centers in R^p found by FCM.

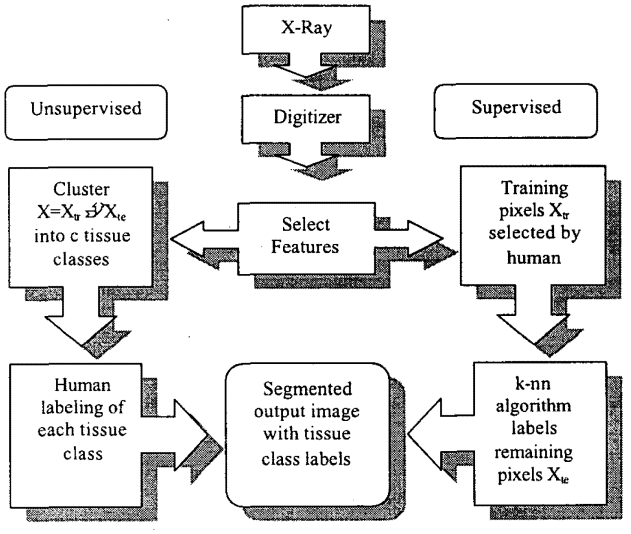


Fig. 1. Two approaches to image segmentation

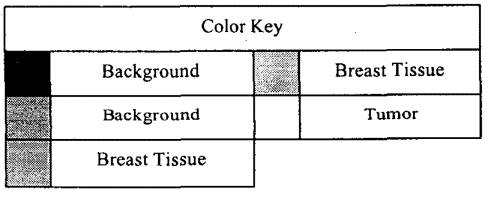


Fig. 2. Original Image

Fig. 3. FCM segmented Image with c=5

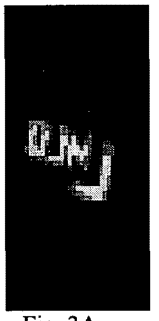


Fig. 3A. Artifact

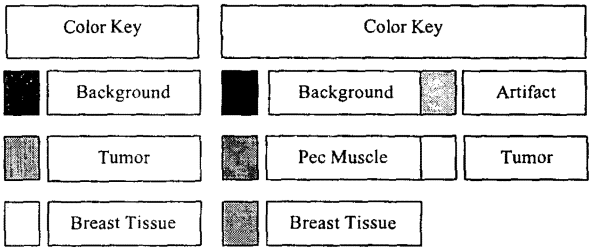
Figure 3 shows a relatively "good" FCM segmentation in terms of separating background from breast, and the large tumor area is clearly defined. However, FCM also labeled the artifact (Figure 3A) in the upper left corner as breast tissue and the pectoral muscle in the upper right corner as tumor. Sometimes errors such as these can be eliminated by changing the number of tissue classes, but this is an ineffective "trial and error" approach. Finding the most useful number of tissue classes (the *cluster validity* problem) is a serious handicap for segmentation by *any* clustering algorithm.

The supervised approach to image segmentation receives human intervention in a different phase of system processing. As shown in Figure 1, a human chooses those pixels that best represent various regions of the breast, background and artifacts and then labels them appropriately prior to segmentation. Figure 4 shows the result of processing the original image in Figure 2 with the crisp k-nn algorithm using k=5 and the Euclidean norm as the distance. In these experiments, c=3 subsets of labeled training pixels have been chosen to represent background, tumor, and non-tumor breast tissues. Each subset was chosen from a 3 x 3 window, resulting in 9 pixels of training data per window. Three windows each were sampled from the background, non-tumor breast tissue, and tumor regions. This produces a total training set where $|X_{tr}| = 81$. As seen in Figure 4, the artifact and the pectoral muscle are again mislabeled. The artifact is labeled non-tumor breast, while the breast tumor and pectoral muscle tissues are grouped together (gray regions). In an attempt to alleviate these problems we sampled a small portion of the original image that best represented the artifact and pectoral muscle with two windows each, and used these additional pixels as training data for c=5 tissue classes. This resulted in a total training set of $|X_{tr}| = 117$ points. Figure 5 shows the resulting segmented image with training data obtained from the artifact and pectoral muscle regions[D]. Again, we observe that the supervised method will also misclassify points when only intensity is used, since tumor and pectoral muscle are still grouped together. In addition, we see that increasing the value of c results in over-segmentation of the tumor and tissue areas.

We also experimented with a fuzzy k-nn rule due to Keller and Gray [9] that weights each vote for a class label by the inverse of the (intensity) distance between the voter and the pixel to be labeled. This classifier is called the *inverse distance weighted* (IDW) design. The results of the limited experiments we conducted with this algorithm were disappointing in that segmentations of the image in

[D] For presentation purposes, output images are shown with uniform cropping of 7 pixels per edge. Figures 4-9 have an effective image size of (m-M+1) x (n-N+1), where the original image is m x n and the window size is M x N. Intensity of border pixels in Figures 4-9 are original intensities.

Fig. 2 were (visually) much less appealing than those shown in Figs. 4 and 5.



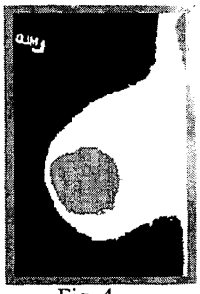


Fig. 4.
5-nn 3-class
training data

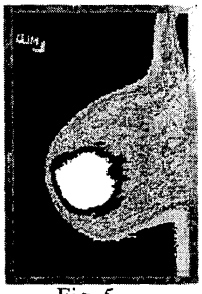


Fig. 5.
5-nn 5-class
training data

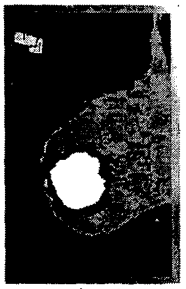
3. DIFFERENT FEATURES

The processing described thus far has been based on a single feature, viz., the measured intensity x_{ij} at pixel (i,j) in X. The artifact problems shown in Figures 3-5 can be alleviated somewhat by introducing more complex features. To this end, we define the pixel vector $\mathbf{x}_{ij}$ associated with spatial address (i,j) in an image as the p-vector $(x_{1,ij}, \ldots, x_{p,ij})$, $p \geq 1$. Since the desired output of segmentation is a set of c homogeneous regions, we augmented the intensity x_{ij} at (i,j) by region-based features. The most convenient regions surrounding (i,j) are M x N windows (M,N odd), and the features we have chosen to experiment with are the mean intensity and the standard deviation within a window. Specifically, we define the *mean* m^w_{ij} and *standard deviation* s^w_{ij} of the pixel intensities in an M x N window W centered at (i,j) as

$$m_{ij}^W = \sum_{(s,t)\in W} \Sigma x_{s,t}/|W| \quad \text{;and} \qquad (1)$$

$$s_{ij}^W = \sum_{(s,t)\in W} \Sigma (x_{s,t} - m_{ij}^W)^2/|W| \quad . \qquad (2)$$

Figures 6 and 7 show the results of processing the original image in Figure 2 again with the k-nn algorithm using k=5, c=5, and the Euclidean norm as the distance. These outputs are based on p=3 features per pixel: $\mathbf{x}_{ij} = (x_{ij}, m^W_{ij}, s^W_{ij})$ for W=3 x 3 (Figure 6) and W=15 x 15 (Figure 7). Figures 6 and 7 are noticeably better than Figure 5 in terms of improved segmentation of the breast tissues. However, smaller regions of the tumor and pectoral muscle are still incorrectly mislabeled as artifact. Additionally, regions of the pectoral muscle and artifact have either been mislabeled as artifact or pectoral muscle respectively. While Figures 6 and 7 seem somewhat better than Figures 4 and 5, complete elimination of the label artifact and useful isolation of the tumor region require a final step: thresholding the segmented image.



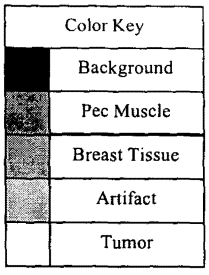


Fig. 6.
5-nn with 5-class
training data and
3x3 window features

Fig. 7.
5-nn with 5-class
training data and
15x15 window
features

Figures 8 and 9 show the results of processing the original image in Figure 2 with the k-nn algorithm using k=1, and two thresholded output classes, and the Euclidean norm as the distance. The training set for Figures 8-11 remained the same as in the previous experiments, having c=5 physically labeled training classes. These outputs are based on p=1 (intensity) or 5 features per pixel, respectively, where $\mathbf{x}_{ij} = (x_{ij}, m^{W3x3}_{ij}, s^{W3x3}_{ij}, m^{W15x15}_{ij}, s^{W15x15}_{ij})$. The black and white images in Figures 8 and 9 are created by thresholding 1-nn segmentations of the image in Figure 2 using the following rule:

> If unlabeled pixel z is labeled tumor by 1-nn, set intensity of z = 255; otherwise, set intensity z = 0.

Notice that the boundary of the tumor in Figure 9 is considerably smoother than it is in Figure 8. This is the expected result of using the five feature data, which has four window based features that create the smoothing effect.

Figures 10 and 11 show the results of applying this same approach with 3 output classes instead of 2, still based on c=5 training classes with $|X_{tr}| = 117$ points in R (Figure 10) or R^5 (Figure 11). Figures 10 and 11 are thresholded similarly to Figures 8 and 9 with the following rules:

If unlabeled pixel z is labeled tumor by 1-nn, set intensity of z = 255, else if unlabeled pixel z is labeled non-tumor breast tissue by 1-nn, set intensity of z = 128; otherwise, set intensity z = 0.

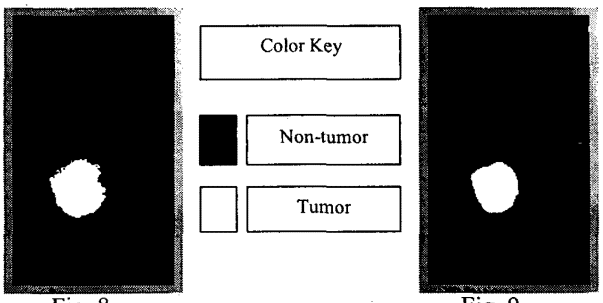


Fig. 8.
1-nn with 2 output classes and 1 feature

Fig. 9.
1-nn with 2 output classes and 5 features

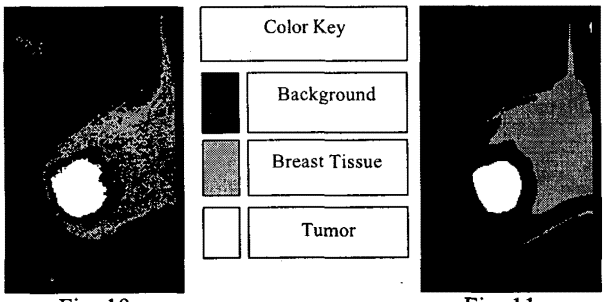


Fig. 10.
1-nn with 3 output classes and 1 feature

Fig. 11.
1-nn with 3-output classes and 5 features

Figure 10 shows that the artifact is still being mislabeled as breast tissue. In Figure 11 the artifact is no longer mislabeled, but a few pixels within the pectoral region have been mislabeled as tumor. Figure 11 shows an appreciable improvement over Figures 3 – 7 and Figure 10 in terms of enhancement of medically relevant areas without significant extraneous clutter.

4. CONCLUSIONS

We have shown that supervised and unsupervised methods of segmentation in digital mammograms will have higher misclassification rates when only intensity is used as the discriminating feature. However, with additional features such as window means and standard deviations, methods such as the k-nn algorithm are able to significantly reduce the number of mislabeled pixels with respect to certain regions within the image.

In the future we plan to study the incorporation of additional region-based features and speed-up techniques to improve the final segmentation. In addition, we plan to include performance ratings from practicing clinicians and comparison to DDSM ground truth to evaluate the efficacy of our final system outputs.

REFERENCES

[1] American Cancer Society (1999). *Breast Cancer Facts & Figures 1997-1998.*

[2] Basset L.W., Manjikian V. III, Gold R.H. (1990). Mammography and breast cancer screening (Review). *Surg Clinics of North America*, 775-800.

[3] Kopans, D. B. (1996). The potential impact of computer-aided diagnosis on clinical mammography. *Proceedings of the Third International Workshop on Digital Mammography*, 35.

[4] Kegelmeyer, P. (1995). Breast cancer detection software licensed. News release of Sandia National Laboratories, August.

[5] Heath, M., Bowyer, K., Kopans, D., Kegelmeyer Jr, P., Moore, R., Chang, K., and Munishkumaran, S. (1997). Current status of the Digital Database Screening Mammography. *Proceedings of the Fourth International Workshop on Digital Mammography*, 457-460.

[6] Devijver, P. A.and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall International, Inc.

[7] Bezdek, J. C., Hall, L. O., Clark, M., Goldof, D. and Clarke, L. P. (1997). Segmenting medical images with fuzzy models: An update. In *Fuzzy Information Engineering*, ed. Dubois, D., Prade, H. and Yager, R., Wiley, NY, 69-92.

[8] Bezdek, J. and Sutton, M.A. (1999). Image processing in medicine. To appear in *Handbook of Fuzzy Sets, 7: Applications*, H.-J. Zimmermann, ed., Kluwer Publishing Company.

[9] Keller, J.M., Gray, M., and Givens, J. (1985). A fuzzy k-nearest neighbor algorithm, *IEEE Trans. Syst., Man and Cyberns.*, 15, 580-585.