A MICROSCOPIC LOOK AT BREAST SKIN-LINE METRICS: A PERFORMANCE EVALUATION STRATEGY FOR FFDM/SCREEN-FILM PROJECTION MAMMOGRAMS

Yajie Sun, Jasjit Suri, Rangaraj M. Rangayyan*, Roman Janer
Fischer Imaging Corporation, Denver, CO, USA
* Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada

ABSTRACT: Performance evaluation of breast skin-line estimation algorithms is a crucial step for standardization of Computer-Aided Detection (CAD) techniques applied to mammograms. A good quantitative analysis for skin-line will benefit and facilitate (a) breast region segmentation algorithms and (b) the digital image acquisition chain including digital detectors. However, there has been no consensus on the metric to be used for evaluation of the skin-line boundaries.

This paper presents a close look at the metrics used for error measurement between the ground truth boundaries (traced by radiologists) and the automatic computer-estimated boundaries of breast skin-lines. We demonstrate the comparison of two major metrics: Suri’s Polyline Distance Measure (PDM) [1-4] and the Hausdorff Distance Measure (HDM) based on a distance transform and image algebra [5]. In addition, we also present a system which automatically (a) estimates normalized False Negative Fraction (FNF) and False Positive Fraction (FPF) measures, and (b) spots out those skin-line boundaries which are not within a radiologist’s accepted error threshold based on quartile measurement.

Our error metric techniques were applied to 83 images from the MIAS database [6], where the computer-estimated boundaries were computed using the Deformable Model developed by Ferrari et al. [7]. The PDM method yielded a mean error (µ) of 2.49 pixels with a standard deviation (σ) of 3.69 pixels. The HDM method yielded µ of 21.06 pixels and σ of 10.56 pixels. The normalized FNF was 0.57% and the normalized FPF was 1.27%.

KEY WORDS: Screening mammogram, breast skin-line, performance evaluation, Hausdorff distance

1. Introduction

Fischer is a pioneer in X-ray mammography and has established an industry standard in manufacturing mammography (SenoScan®) and biopsy (MammoTest) machines. The mammograms produced by these machines are of high resolution and quality. Nonetheless, with changes in the imaging chain of full-field digital mammography (FFDM) systems, corresponding changes in the software system need to take place, including Computer-Aided Detection (CAD) for lesion enhancement and detection. Different CAD systems give different results for breast region segmentation and skin-line estimation; as a result, there has been no consensus in the evaluation of skin-line detection algorithms. There is a lack of quantitative metrics for the evaluation of the estimated breast skin-lines. Most of the earlier breast boundary extraction algorithms used a subjective measure: the percentage of “acceptable” estimated breast boundaries. This method is subjective and also time consuming.

There has been considerable work on the development of skin-line extraction algorithms in the past decade [7-11, 14]. However, no serious work has been directed to evaluate the performance of the estimated skin-lines on mammograms. Bick et al. [8] used localized analysis based on modified histogram analysis, which consisted of global thresholding, region growing and morphological filtering. The skin-line was then extracted using contour tracing. On the whole, 97% of the results were “acceptable” when tested on a dataset of 740 mammograms.

Abdel-Mottaleb et al. [9] used multiple thresholding to get different breast masks in order to locate the final skin-line. Gradients were calculated from the mammogram, and the entire breast area was obtained from a union of two thresholded images. Overall, 98% of the results were “acceptable” when tested on a dataset of 500 mammograms.

Ojala et al. [10] developed a robust automatically adaptive thresholding method based on an analysis of image histogram that consisted of histogram thresholding, morphological filtering, and contour fitting. The initial segmentation was obtained using an automatic and adaptive threshold based on an analysis of the histogram. A final smooth boundary was obtained using contour fitting. The algorithm achieved satisfactory results over a range of screening mammograms digitized using different scanners.

McLoughlin et al. [11] used a greedy snake algorithm to locate the skin-line after initial segmentation using...
Otsu’s method [12]. The algorithm was tested on 40 mammograms from the DDSM database [13] and the results were “acceptable”.

Wirth et al. [14] developed a breast region segmentation method using active contours. A given mammogram was initially segmented using the threshold determined by Rosin’s method [15], and the initial boundary was extracted. A modified greedy active contour algorithm was then used to locate the final smooth boundary on the original mammogram. The algorithm achieved acceptable results on 25 mammograms from the MIAS database [6]. Evaluation of the segmentation was based on the percentage of false-positive and false-negative pixels; the false-negative fraction (FNF) and false-positive fraction (FPF) were determined by a quantitative comparison between the skin-line identified by a radiologist and the corresponding computer-estimated skin-line. The average FNF was 1.6%, and the average FPF was 1.1%.

Ferrari et al. [7] implemented a modified active contour model to obtain the skin-line. The initial segmentation was obtained using a threshold determined by the Lloyd-Max quantizer [16]. An initial boundary was extracted using a chain code. The final breast skin-line was obtained using an adaptive active deformable contour model. The method was tested on 84 mediolateral oblique (MLO) mammograms from the MIAS database. The method achieved FPF=0.41% and FNF=0.58%.

Most of the breast boundary extraction algorithms [8-11] lack quantitative metrics to compare the estimated breast boundary with the ground-truth breast boundary. In this work, we will investigate different metrics for error measurement between the ground-truth boundaries and the automatic computer-estimated skin-lines. We demonstrate a comparative analysis of the following metrics: Polyline Distance Measure (PDM), Shortest Distance Measure (SDM), and the Hausdorff Distance Measure (HDM) based on a distance transform. We also present the results of FNF and FPF analysis.

The paper is organized as follows: Section 2 describes several skin-line evaluation metrics. The results of our analysis are presented in Section 3, and finally, Section 4 gives the conclusions.

2. Skin-line Metrics

2.1 Polyline Distance Measure (PDM)

In order to compare a computer-extracted boundary with the ideal boundary, a quantitative error measure based on the average polyline distance of each boundary point was developed by Suri et al. [1-4]. The polyline distance is defined as the closest distance from the each estimated boundary point to the ideal/ground-truth breast region boundary. The closest distance of each estimated boundary point can be the perpendicular distance (shortest Euclidean distance) to one of the intervals derived from the successive boundary points of the ideal/ground-truth skin-line, or can be one of the end boundary points joining the points of the closest interval. In the following paragraphs, we will derive the PDM mathematically.

Let \( B_1 \) be the first boundary, and \( B_2 \) be the second boundary. Let the Cartesian coordinates of a point \( A \) on \( B_1 \) be \((x_0, y_0)\). Let there be two successive boundary points \( B \) and \( C \) given by coordinates \((x_1, y_1)\) and \((x_2, y_2)\) on \( B_2 \). Let \( \lambda \) be the free parameter for the equation of the line joining the points \( B \) and \( C \). Then, the line interval \( BC \) between \( B \) and \( C \) is given as:

\[
\begin{align*}
\begin{cases}
(x, y) = (x_1 + \lambda(x_2 - x_1), y_1 + \lambda(y_2 - y_1)) \\
0 \leq \lambda \leq 1
\end{cases}
\end{align*}
\]

where \((x, y)\) is the coordinate of a point on the line and \(\lambda \in [0, 1] \).

Now, let \( \mu \) be the parameter of the distance orthogonal to the line interval \( BC \). Then, the line segment between \((x_0, y_0)\) and \((x, y)\) is perpendicular to the line interval \( BC \). Therefore, we can express \((x_0, y_0)\) similar to equation (1):

\[
\begin{align*}
\begin{cases}
\frac{x_0}{x} = \frac{y_0}{y} = \frac{-(y_2 - y_1)}{x_2 - x_1} \\
\frac{x_0}{x} = \frac{y_0}{y} = \frac{-(y_2 - y_1)}{x_2 - x_1}
\end{cases}
\end{align*}
\]

Solving the above equation using the related determinants, the unknown parameters \( \lambda \) and \( \mu \) are obtained.

Let the two distance measures \( d_1 \) and \( d_2 \) between \( A \) on \( B_1 \) and \( B/C \) on \( B_2 \) be defined as Euclidean distances:

\[
\begin{align*}
d_1 &= \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2} \\
d_2 &= \sqrt{(x_0 - x_2)^2 + (y_0 - y_2)^2}
\end{align*}
\]

The polyline distance \( d_{poly}(A, BC) \) is then defined as:

\[
d_{poly}(A, BC) = \min\{d_1, d_2\}; \quad \lambda < 0, \text{ or } \lambda > 1 \quad 0 \leq \lambda \leq 1
\]

where,

\[
|\mu| = |\mu|_{BC} = \frac{(y_2 - y_1)(x_1 - x_2) + (x_2 - x_1)(y_0 - y_1)}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}
\]

A quantitative error measure between the ideal boundary and the computer-estimated boundary could then be defined using the polyline distance described in equation (6). The measure is defined as the average polyline distance of all boundary points of the estimated and ground-truth breast boundaries. We will denote the measure as \( d_{Error}^{poly} \), which is derived as follows:

\[
\begin{align*}
d_{Error}(A, B_2) &= \min_{S_{boundary}B_2} d(A, S) \\
d_{Error}(B_1, B_2) &= \sum_{verticesB_2} d_{poly}(A, B_2)
\end{align*}
\]
Figure #1 shows an overlay of an estimated boundary (GREEN) and the corresponding ideal boundary (RED). The PDM between the two boundaries is $d_{poly} = 0.963 \approx 1$ pixel.

2.2 Shortest Distance Measure (SDM)

The shortest distance measure represents the shortest distance between an estimated boundary and the ground-truth boundary. When deriving the shortest distance, the roles of the ideal and estimated boundaries can be interchanged. Hence, we denote one boundary as $B_1$, and the other boundary as $B_2$. Let the Cartesian coordinate of a point $A$ on $B_1$ be $(x_i, y_i)$. Let there be a boundary point $B$ given by the coordinates $(x_j, y_j)$ on $B_2$. Let $P_1$ be the number of boundary points of $B_1$, and $P_2$ be the number of points of $B_2$.

To derive SDM, we first compute the shortest distance $d_{ij}^1$ from the point $A$ on $B_1$ to the boundary $B_2$:

$$d_{ij}^1 = \min_{j=1,2,...,P_2} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

(9)

Then, we compute the average value of the shortest distance from the boundary $B_1$ to the boundary $B_2$, labeled as $\bar{d}^1$:

$$\bar{d}^1 = \frac{1}{P_1} \sum_{i=1}^{P_1} d_{ij}^1$$

(10)

Similarly, we obtain the average value of the shortest distance from the boundary $B_2$ to the boundary $B_1$, and label it as $\bar{d}^2$. The final shortest distance measure is

$$d_{shortest} = \frac{1}{2}(\bar{d}^1 + \bar{d}^2)$$

(11)

If each boundary point is in the 8-connected neighborhood of adjacent points, there will be little difference between PDM and SDM.

2.3 Hausdorff Distance Measure (HDM)

The Hausdorff distance is defined to be the maximum of the set of the shortest Euclidean distances between corresponding boundary points [18]. Given two boundary sets, $A$ and $B$, the HDM from $A$ to $B$ is defined as:

$$h(A,B) = \max \{ \min_{a \in A} \{d(a,b)\} \}$$

(12)

where $a$ and $b$ are points of the boundary set $A$ and set $B$, respectively, and $d(a,b)$ is the Euclidean distance between points $a$ and $b$.

The most-common implementation of the generic HDM is a brute-force method as follows:

1. Hausdorff = 0;
2. For each point $a_i \in A$,
   - Minimum = $\infty$;
   - For every point $b_j \in B$, if $d_{ij} <$ Minimum, then Minimum = $d_{ij}$;
   - If Minimum > Hausdorff, then Hausdorff = Minimum.

The Hausdorff distance is asymmetric, and the Hausdorff distance between $A$ and $B$ would be:

$$H(A,B) = \max \{h(A,B), h(B,A)\}$$

(13)

Figure #2. Illustration of the different steps of distance-transform-based HDM. Image size: 1024×1024 pixels. HDM: 38.4 pixels.

2.3.1 Generic HDM

The generic Hausdorff distance computation is straightforward. It is computed as the maximum of the set of the shortest Euclidean distances between corresponding boundary points [18]. Given two boundary sets, $A$ and $B$, the HDM from $A$ to $B$ is defined as:

$$h(A,B) = \max \{ \min_{a \in A} \{d(a,b)\} \}$$

(12)

where $a$ and $b$ are points of the boundary set $A$ and set $B$, respectively, and $d(a,b)$ is the Euclidean distance between points $a$ and $b$.

The most-common implementation of the generic HDM is a brute-force method as follows:

1. Hausdorff = 0;
2. For each point $a_i \in A$,
   - Minimum = $\infty$;
   - For every point $b_j \in B$,
     - $d_{ij} = d(a_i,b_j)$, if $d_{ij} <$ Minimum, then Minimum = $d_{ij}$;
   - If Minimum > Hausdorff, then Hausdorff = Minimum.

The Hausdorff distance is asymmetric, and the Hausdorff distance between $A$ and $B$ would be:

$$H(A,B) = \max \{h(A,B), h(B,A)\}$$

(13)

Figure #2. Illustration of the different steps of distance-transform-based HDM. Image size: 1024×1024 pixels. HDM: 38.4 pixels.

2.3.2 Distance-Transform-based HDM

The Hausdorff distance between the boundaries of two binary regions can be quickly obtained using the distance
transform [5]. An implementation of the Hausdorff distance using the distance transform is as follows:

1. Compute the distance transforms \( D_T \) of the two boundaries \( A \) and \( B \).
2. Compute the pixel-wise maximum of \( A_{DT} \) and \( B_{DT} : \ AB_{DT} = \max \{ A_{DT} , B_{DT} \} \).
3. Generate binary masks, \( A_R \) and \( B_R \), from the two boundaries, and obtain the XOR: \( AB_{XOR} = A_R \oplus B_R \).
4. Use \( AB_{XOR} \) to mask \( AB_{DT} \) to produce \( AB_{MASK} \).
5. Obtain HDM as the maximum value in \( AB_{MASK} \): \( HDM = \max ( AB_{MASK} ) \).

Figure #2 illustrates the above steps to obtain the Hausdorff distance measure.

### 2.4 FNF/FPF Measures

The FNF and FPF are often used to evaluate the performance of a classification procedure or a detection procedure in the biomedical fields. False-negative (FN) pixels are the breast pixels in the ground-truth mask \( AREA_{GT} \) that are not present in the region enclosed by the estimated boundary. False-positive (FP) pixels are the true background pixels in the ground-truth that are enclosed by the estimated boundary, as shown in Figure #3. FNF and FPF are then defined as:

\[
FNF = \frac{FN}{AREA_{GT}} \quad FPF = \frac{FP}{AREA_{GT}}
\]

Figure #3. Illustration of FPF and FNF: \( FNF=FN/AREA_{GT} \) and \( FPF=FP/AREA_{GT} \), where \( AREA_{GT} \) is the breast region enclosed by the ground-truth boundary.

### 2.5 Quartile Measure

Quartile analysis is a strategy to spot out those images in the database whose skin-line errors are unacceptable by radiologists. Having computed the chosen error metric between the computer-estimated boundary and the ground-truth boundary for all of the cases in the study, we arrange these errors in ascending order along with the case names. The list of ascending order errors is divided into four parts representing quartiles. The radiologist sets a threshold of acceptable error, and the computer automatically identifies the skin-lines that are unacceptable to the radiologist.

### 3. Results

We tested the metrics described above with a dataset of 83 mammograms selected from the MIAS database [6]. The ground-truth skin-lines were traced by a radiologist, and the estimated skin-lines were obtained using the method developed by Ferrari et al. [7]. Table 1 shows a comparative listing of the skin-line error metrics. The mean (\( \mu \)) and standard deviation (\( \sigma \)) values were computed over the 83 cases. PDM is a good measure for evaluating the average distance between the ground-truth and estimated skin-lines. FNF and FPF are good to evaluate the breast area and background area errors between the ground-truth and estimated skin-lines. HDM has large distance errors because it gives the maximum of various possible distances between two boundaries, instead of an average or the minimum.

Figure #4 shows a comparison of PDM, SDM, and HDM. Figure #5 shows a comparison of FNF, FPF, and FNF+FPF. From quartile analysis, we can automatically differentiate good results from marginally acceptable cases. Figure #6 shows the results of quartile analysis based upon the PDM over 83 cases. Figure #7 (a) and (b) illustrate two examples of overlays of the detected and ground-truth contours on the corresponding mammograms.

<table>
<thead>
<tr>
<th>Skin-line error Metric</th>
<th>Average (( \mu ))</th>
<th>Standard deviation (( \sigma ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDM (Pixels)</td>
<td>2.486</td>
<td>1.007</td>
</tr>
<tr>
<td>SDM (Pixels)</td>
<td>2.492</td>
<td>1.006</td>
</tr>
<tr>
<td>HDM (Pixels)</td>
<td>21.06</td>
<td>10.56</td>
</tr>
<tr>
<td>FNF</td>
<td>0.57%</td>
<td>0.83%</td>
</tr>
<tr>
<td>FPF</td>
<td>1.27%</td>
<td>0.41%</td>
</tr>
<tr>
<td>FNF+FPF</td>
<td>1.84%</td>
<td>0.75%</td>
</tr>
</tbody>
</table>

Figure #4. The distance error curves. PDM and SDM are almost identical over 83 cases.
Figure #5. The curves of FNF, FPF, and FNF+FPF. For most of images, the breast error (FNF) is smaller than the background error (FPF).

Figure #6. Quartile analysis of polyline distance errors.

Figure #7. From quartile analysis, we can order the 83 images tested into four quartiles. (a) An example overlay image of ground-truth and estimated skin-lines from the first quartile, PDM=1.96 pixels. (b) Overlay of an example from the fourth quartile, PDM=7.52 pixels.

4. Conclusions

We have tested several metrics to evaluate estimated breast skin-lines. The metrics described can be categorized into two groups. The metrics in the first group, which includes PDM, SDM, and HDM, measure the distance between two skin-lines. The metrics in the second group, which includes FNF, FPF, and FNF+FPF, measure the area of the difference between two skin-lines. PDM is one of the best distance error measures. FNF and FPF are usually computed together to evaluate the breast region and background region errors. In our database of 83 pairs of ground-truth and estimated skin-lines, the average FNF is smaller than the average FPF, which means the breast error is less than the background error.

Quartile analysis can be used to automatically categorize good results from marginal cases using a threshold set by a radiologist. From the cases in the fourth quartile, we can further investigate why a skin-line estimation algorithm performed poorly; the results may assist in developing approaches to improve the performance of the skin-line estimation method.

In the future, we will continue to investigate quantitative metrics for evaluating the performance of skin-line estimation algorithms. Skin-line estimation is an important step in CAD with mammograms. A good quantitative metric for skin-line error will benefit and facilitate the process of developing algorithms for computer processing of mammograms and the detection of breast cancer.

References:


