EBA STUDY OF ALGORITHMS FOR THE ISOLATION OFNEOPLASMS ON MAMMOGRAPHIC IMAGES

Y.A.Bulanova¹, S.S.Sadykov¹ ,I.R.Samandarov² ,N.T.Dushatov², Z.M.Miratoev²

Candidate of Technical Sciences, Associate Professor, Doctor of Technical Sciences, Professor, Candidate of Technical Sciences, Associate Professor, ¹Murom Institute of Vladimir State University, Murom, Russia,

²Almalyk branch of Tashkent State Technical University, Almalyk, Uzbekistan

yuliyabulanovad@yandex.ru, sadykovss@yandex.ru

<u>samadarov.2003@gmail.com</u>, <u>n_dushatov@rambler.ru</u>, <u>miratoyev2014@mail.ru</u> Abstract: More than 10 methods have been investigated for the isolation of neoplasms on mammograms

Keywords: Neoplasm, image, mammogram, segmentation, texture, histogram.

INTRODUCTION

Algorithms for selecting objects in an image are based on segmentation methods.

Segmentation is a division of an image into regions based on some similarity

criterion (features). These features can be natural (brightness, texture, etc.) or

artificial (brightness distribution histogram, spectrum, etc.). Before the stage of

segmentation, pre-processing of images is carried out [8-11].

Below are the results of more than 10 algorithms for neoplasm detection on mammograms.

1. Threshold segmentation

The simplest segmentation algorithm is the segmentation by brightness threshold.

$$s(x, y) = \begin{cases} s1, & npu \quad g(x, y) \le t, \\ s2, & npu \quad g(x, y) > t, \end{cases}$$
(1)

where g(x,y) is the source brightness at point (x,y), t is the threshold value, s1, s2 are brightness values taken under certain conditions, s(x,y) is the output brightness value at point (x,y).

We can say that this method is most often used for binary segmentation (when s1 and s2 take values 0 or 1).

The main problem in this type of segmentation is the choice of threshold, because choosing too small a threshold will lead to over-segmentation, when the objects in the image are divided into a large number of segments, which are difficult to recognize. If you choose too high a threshold, then a large amount of information about the objects is lost, moreover, some objects may be mistakenly combined with others or lost altogether.

The most common algorithms for selecting a threshold are:

1. Choice of segmentation threshold by brightness distribution histogram on the image. The simplest of the thresholding methods consists in dividing the image histogram into two parts using a single global threshold. The image segmentation is then performed by scanning the image item by item, with each pixel being marked as belonging to an object or background, depending on whether the brightness of that pixel exceeds the value of threshold t or not. The success of this method depends entirely on how well the histogram of the image lends itself to separation. Determining the threshold using a luminance histogram is a simple method that achieves "clean" segmentation if the image histogram is clearly bimodal. This shape of the histogram means that two types of relatively common pixels - bright and dark - can be distinguished in the image. And the histogram is easily separated by a single global threshold t located in the trough between the histogram peaks.

Fig. 1 shows an example of segmentation by the global threshold of the original image (Fig. 1a).



Fig. 1 Example of segmentation by global threshold: a) original image, b) result of threshold segmentation of the original image

2. The Burnsen algorithm [1]. This algorithm is used for schematic and cartographic images. The essence of the method is as follows: the image is divided into blocks of size $r \times r$ (r is an odd number) with the center at (m,n). In each block threshold calculated by formula (2) is used.

$$t(m,n) = \frac{\max(f(x,y)) + \min(f(x,y))}{2},$$
(2)

where $\max(f(x,y))$, $\min(f(x,y))$ - maximum and minimum brightness values in the block, respectively, t(m,n) - threshold value in the block.

Bernsen method has a number of disadvantages: after processing monotone brightness areas strong parasitic disturbances are formed, in some cases leads to the appearance of false black spots [1].



Fig. 2 - Example of segmentation by threshold calculated by the Burnsen algorithm: a) original image, b) result of threshold segmentation of the original image

2. Otsu's method [1]. This method is the most effective method for determining the threshold. The essence of the method: the histogram of brightness distribution in the image is divided into 2 classes c_0 and c_1 using threshold value t ($t \in [0, 255]$). The probability of each of these two classes and their average brightness values are described as follows:

$$P_{0} = \sum_{i=0}^{t} p_{i} = P_{t},$$

$$P_{1} = 1 - P_{t}$$

$$\mu_{0} = \sum \frac{i \times p_{i}}{P_{0}} = \frac{\mu_{t}}{P_{t}}$$

$$\mu_{1} = \frac{f_{cp} - \mu_{t}}{1 - P_{t}}$$
(3)

where p_i is the probability of occurrence of the *i*-th brightness in the image, μ_0 and μ_1 are the average brightness in each class, respectively, f_{cp} is the average brightness **over the whole image**.



Fig. 3 - Example of segmentation by threshold calculated by Otsu method: a) original image, b) result of threshold segmentation of the original image

2. Segmentation by area building algorithm

The region-building algorithm is one of the simple brightness segmentation algorithms: neighboring elements with the same brightness are grouped together and form a region.

Researchers Bryce and Fennema [2] proposed a method of region building based on simple growth rules. In the beginning, all points of the image are considered in a four-connected neighborhood, in which the brightnesses of neighbors are checked, the points with the same brightnesses are combined. In the next step weak boundaries between areas are removed. After the initial steps of buildup it may turn out that some region contains previously merged subareas with different brightness values.

The heuristic rule of merging smaller areas with larger ones is as follows:

$$D/\min\{P_1, P_2\} > \varepsilon 1, \qquad (4)$$

where $\varepsilon 1=0,5$, *D* is the length of the smaller boundary of the fusion region, P_1 , P_2 are the fusion regions.

This rule prevents the merging of adjacent regions of approximately equal size, but allows the absorption of smaller regions by larger ones.

The second rule ensures the elimination of common weak boundaries remaining after the application of the first rule. Adjacent regions merge if:

$$D/C > \varepsilon 2, \tag{5}$$

where $\varepsilon 2=3/4$, *D* is the length of the smaller boundary of the fusion region, *C* is the length of the common boundary.

Applying only the second rule leads to excessive merging of regions.

This algorithm shows the best segmentation results for images with few objects and no texture.



Fig.4 - Example of region building algorithm for the original image in Fig.1a)

3. Segmentation using hierarchical Markov Random Field (MRF) model

The grayscale image is described by the function p(i,j), where $i \in [1, N]$, $j \in [1,M]$, N^*M - image size. The brightness of the image point is $y_s \in [0,255]$. A four-connected neighborhood is used as a system of neighbors for the point.

The clique V is the basic structural unit, a set of neighboring pixels, one of which - analyzed, the others belong to its system of neighbors.

According to Hammersley-Clifford's theorem, an image can be described by a hidden Markov random field model if and only if the brightness distribution of the image is described by a discrete Gibbs distribution that has the following form [3]:

$$P(A) = \frac{1}{Z} \times \exp^{-U(a_s)},$$

$$U(a_s) = \sum_{c \in C} \gamma_c \times V_c(a_s),$$
(6)

where a_s is a quantitative characteristic of a point equal to brightness y_s , Z is a normalizing constant, $U(a_s)$ is an energy function determining the generalized

interaction of the point with the system of neighbors, V_c is a clique function C describing the statistical relation of the point with the system of neighbors.

In the framework of the model [3], an ordered sequence of images is generated from the original grayscale image according to the principle: each subsequent image is obtained by changing the contrast of the previous one according to the scheme, shown in Fig. 5.



Fig. 5 - Hierarchical structure used in the work: OC - original image, HC - image with high contrast, LC - image with low contrast

When classifying the point *s*, it is necessary to analyze only the transition probability of the point from the state r_i , characterized by the *i*-th layer, to the state r_{i-1} or r_{i+1} , choose the state, the probability of being in which is maximal.

$$V^{i\pm 1}(a_s) = y_s - y_k^{i\pm 1}, \, k \in \eta_s.$$
(7)

The energy function is equal to:

$$U_{s}^{i} = \frac{1}{2} \times \left(\sum_{k} y_{s} - y_{k}^{i-1} + \sum_{k} y_{s} - y_{k}^{i+1} \right).$$
(8)

The classification of point *s* can be represented by the following expression:

$$r_s = \arg \max_i (\exp^{-U_s^i}), i \in [1, R],$$
 (9)

where r_s is the label of pixel $s, r_s \in [1, R]$.

Thus, the segmentation process decides which state (segmentation class) is most likely for each pixel.



Fig.6 - Example of texture segmentation using MRF: a) original image, b) result of texture segmentation of the original image

The advantage of using the MRF model is in singling out small and lowcontrast objects.

The disadvantage of segmentation using MRF is the a priori setting of the number of hierarchy layers, and hence the number of segmentation classes, which is an indicator of segmentation "with a teacher".

4. Marker Watershed Algorithms

The notion of watershed is based on the representation of the image as a three-dimensional surface, given by two rostrate coordinates and the brightness level as the height of the surface (relief) [4].

Let M_1 , M_2 , ..., M_R be the sets of points on the coordinate plane corresponding to local minima of the surface g(x, y); g(x, y) is usually a gradient image. Let us denote by $C(M_i)$ the set of points of the basin corresponding to local minima of M_i (recall that the points of any basin form a component of connectivity). We will use the notation min and max to indicate the smallest and the largest values of the image g(x, y). Finally, the notation T[n] means the set of points (s, t) for which g(s,t) < n, i.e.

$$T[n] = \{(s,t) \mid g(s,t) < n\},$$
(10)

Geometrically, T[n] is the set of points where the surface g(x, y) lies below the plane g(x, y) = n. When water fills the terrain, the level rises as integer discrete increments from $n=\min+1$ to $n=\max+1$. In the process of water rise at any step n, the algorithm needs to know the number of points lying below the water level. Imagine that all points of the set T[n] (i.e. which lie below the plane g(x,y)=n) are marked black, and all other points are marked white. Then at arbitrary (n-th) step of water level rise, the considered three-dimensional surface in projection on plane XY can be represented by binary image, in which black points correspond to the points of initial function lying below the plane g(x,y)=n. This interpretation is very useful for understanding the following exposition.

Let $C_n(M_i)$ denote the set of points of the pool with local minimum M_i , which were flooded at step *n*. Given the above, $C_n(M_i)$ can be considered as a binary image given by the relation.

$$C_n(M_i) = C(M_i) \cap T[n], \tag{11}$$

Let now C[n] be the union of the flooded parts of all pools at step n:

$$C[n] = \bigcup_{i=1}^{R} C_n(M_i), \qquad (12)$$

Then $C[\max+1]$ is the union of all available pools:

$$C[\max+1] = \bigcup_{i=1}^{R} C(M_i), \qquad (13)$$

It can be shown that the algorithm never removes elements from sets $C_n(M_i)$ and T[n]; thus, as n increases, the number of elements of these sets either increases or remains unchanged. Hence, C[n-1] is a subset of C[n]. C[n] is also a subset of T[n], so C[n-1] is also a subset of T[n]. Hence an important result follows: each connectivity component of set C[n-1] is contained in exactly one connectivity component of set T[n].

The algorithm for finding watershed lines starts with initialization $C[\min+1]=T[\min+1]$. After that the algorithm is executed recurrently, assuming at the *n*-th step the set C[n-1] is already constructed. To get set C[n] from set C[n-1] the following procedure is used. Let Q[n] be the set of connectivity components

of set T[n]. Then for each connected component $q \in Q[n]$ there are three possibilities:

a. $q \cap C[n-1]$ is an empty set;

- b. $q \cap C[n-1]$ contains a single connectivity component of the set C[n-1];
- c. $q \cap C[n-1]$ contains more than one connectivity component of the set C[n-1].

The way to construct C[n] by C[n-1] depends on which of these three conditions takes place.

Condition (a) means that a new local minimum is encountered (a new pool begins to fill); in this case, to construct C[n] the component q is added to C[n-1]. Condition (b) takes place when q lies inside the pool of some local minimum; in this case the component q is also added to C[n-1] to construct the set C[n].

Condition (c) arises when the points of a ridge separating two or more basins meet. In this case, further water rise would lead to the merging of these pools, so a partition (or partitions, if more than two pools are merged) must be built inside the connected component q to prevent the pools from merging together. As explained in the previous section, a partition of one pixel thickness, if necessary, can be constructed by applying a dilatation operation on the set $q \cap C[n-1]$ by a 3x3 primitive filled with units, and then bounding the result of the dilatation to the points of the set q.

One of the most important applications of watershed segmentation is the selection of brightness-homogeneous objects (in the form of spots) on the background of the image. Areas characterized by small brightness variations have small gradient values. That is why in practice we often encounter the situation when the watershed segmentation method is applied not to the image itself, but to the gradient of this image. In this formulation, the local minima of the basins agree well with the small values of the gradient, which usually correspond to the objects of interest.

This algorithm is described in detail in [4].



Fig. 7. Segmentation result and overlaying of watershed contours: a) original mammogram, b) binarization, c) foreground object markers, d) result of morphological operation (erosion), overlaying contours of selected objects on the original image

5. Binary homogeneity augmentation algorithm

This algorithm can be represented as a block diagram [5] shown in Fig. 8.

1) Image orientation

If a mammogram taken in the mediolateral projection is used, the image should be inverted so that the pectoral muscle is in the upper right corner. This requires reducing the intensity of the image, the tissue in this area is evaluated with minimal cross-entropy [5]. After locating the chest wall, we extract the vertical center of gravity of the image, and then assume that the asymmetric area closest to the right side of the vertical center is the border of the mammary gland. The image is flipped horizontally if you want to place the asymmetric region below the vertical center of gravity, as a result of these transformations the image will be located correctly (the pectoral muscle will be in the upper right corner) [5].

2) Noise reduction.

There are many types of noise. High intensity noise is characterized by high optical densities, such as marks or scanning artifacts. These noises must be replaced by black pixels.



Fig. 8. Schematic of the algorithm

3) Gaussian Blur

Gaussian blur is used to smooth out sharp changes in brightness.

$$G(x, y) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2 + y^2}{2\sigma^2}},$$
 (14)

where G(x,y) is the filtered image, σ is the standard deviation of the filtering area, *x*, *y* are the coordinates of the center of the filtering window core.

4) Binary homogeneity increase algorithm

This method uses the maximum value of the difference threshold, which is a constant threshold determined by experience. We start comparing this value with the image by starting to scan the array horizontally from the upper left corner. If the result of the difference is greater than the threshold, the array will be divided into 2 equal subsets, the middle position and the last and first positions of both subsets will be stacked. Otherwise, the values of the subsets will affect the other positions after the value changes, using a single color quantization method in the

color space, dividing it into 16 levels. This process continues recursively until the value of the stack and the right pixel in the mammogram converge. The same process is repeated to analyze the image vertically [5].

5) Algorithm of contours extraction

At the first stage a horizontal sweep is carried out. If a change in brightness is observed, these points are marked in black, indicating the horizontal contour point. The process continues for all points of the image to obtain a horizontal contour map. Further actions are carried out vertically to obtain a vertical contour map. Then we combine the horizontal and vertical maps by performing a logical OR operation for the two images to obtain a single map [4].

6) Algorithm of breast contour extraction [5]

Step 1. We start scanning from the left side of the image to find the leftmost point of the breast area (BF).

Step 2. Draw a vertical line along this point from top to bottom, which will indicate the left border of the breast area.

Step 3. Scan the edge of the map from right to left, starting from the first row.

Step 4. Find the point that calls the end of the contour. Go through all the points of the edge, considering all the surrounding points in priority order clockwise, the goal is the point with the highest priority.

Step 5. The points that have a lower priority are stored in the stack in case of a deadlock. The search continues for all points until the last point of the chest contour is reached.

Step 6. If a deadlock is reached, pull the lower priority point from the stack and continue the traversal process.

Step 7. Save the points that intersect the list of edge points, which will be used to draw the boundary of the ML.

Step 8. Consider each point to be traversed until we reach the base line or the bottom of the image.

Step 9. If the bottom of the image or the left baseline is not reached, the entire outline is removed and the list is filled in from Step 4.

Step 10. We draw the first point.

Step 11. Draw a simple curve between two neighboring points from the list of edge points.

7) Segmentation

The image is divided into regions. Each region has a different intensity value. Fatty tissue, glands, lobules, and ducts have different intensities, and can be highlighted in different regions. The neoplasms have a high intensity compared to normal breast tissue. In this step, you need to find the average brightness in each region and assign that color to the region. For each region to have a closed contour, the following method is used: color the initial pixel in the image, then check the 4link neighborhood of the point to find out if these points are boundary points or already painted over.

If the point is not bordered or painted over, it is the center of the further search.



Fig.9 - Segmentation result with the binary homogeneity augmentation algorithm

6. Algorithm "k-means"

The k-means method (k-means) is the most popular clustering method. The algorithm seeks to minimize the total quadratic deviation of cluster points from the centers of these clusters [6]:

$$\min V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_j)^2, \qquad (15)$$

where k is the number of clusters, S_i are obtained clusters, i=1,2,...,k and μ_i are centers of clusters, x_i is the cluster point, $x_i \in S_i$.

$$\mu_i = \frac{1}{S_i} \cdot \sum_{x_j \in S_i} x_j , \qquad (16)$$

The centers of clusters are called main points, and the method itself is called the method of main points and is included in the general theory of main objects providing the best approximation of the data.

The algorithm is a version of the EM algorithm also used to separate a mixture of Gaussians. It is an iterative method. It partitions the set of vector space elements into a predetermined number of clusters k.

The basic algorithm is given below:

1. Select *k* cluster centers, either randomly or based on some heuristic.

2. Place each pixel of the image into the cluster whose center is closest to that pixel.

3. Recalculate the cluster centers by averaging all the pixels in the cluster.

4. Repeat steps 2 and 3 until convergence (for example, when pixels remain in the same cluster).

Here, the sum of the squares or absolute values of the differences between a pixel and the center of the cluster is usually taken as the distance. The difference is usually based on the color, brightness, texture, and location of the pixel, or a weighted sum of these factors; k can be chosen manually, randomly, or heuristically.

This algorithm is guaranteed to converge, but it may not lead to an optimal solution. The quality of the solution depends on the initial cluster set and the value of k.

The basic idea is that at each iteration the center of mass is recalculated for each cluster obtained at the previous step, then the vectors are partitioned into clusters again according to which of the new centers was closer according to the chosen metric.

The algorithm terminates when at some iteration there is no change in the center of mass of the clusters. This happens in a finite number of iterations, since the number of possible partitions of finite set is finite, and at each step the total quadratic deviation *V* decreases, so looping is impossible.

The disadvantages of this method are as follows:

It is not guaranteed to reach a global minimum of the total quadratic deviation *V*, but only one of the local minima.

The result depends on the choice of initial cluster centers, their optimal choice is unknown.

The number of clusters needs to be known in advance.

The advantages of this method are the speed of execution and simplicity of implementation.



Fig. 10 - The result of k-means segmentation of the original image Fig. 1a)

7. Fuzzy C-mean segmentation

Fuzzy *C*-mean segmentation [7] is an iterative method of partitioning the image into clusters. It boils down to minimizing the following function:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{M} u_{ij}^{m} \times d(x_{i}, c_{k}), \qquad (17)$$

Choose the number of clusters *M*, the fuzzy measure $1 \le m \le \infty$, the distance function d(x,c), the search termination criterion $0 \le e \le 1$.

A matrix of weights for each point belonging to a cluster with a certain center is set:

$$U = \{u_{ij}(x_i, c_j) : x_i \in X, c_j \in C\},$$
(18)

The cluster membership can be calculated randomly:

$$0 < u_{ij} < 1, \ 0 < \sum_{i=1}^{N} u_{ij} < N ,$$
 (19)

The next step is to calculate centroids:

$$c_{j} = \frac{\sum_{i=1}^{N} u_{ij}^{m} \times x_{i}}{\sum_{i=1}^{N} u_{ij}^{m}},$$
(10)

Weights are then recalculated:

$$u_{ij}^{m} = \frac{1}{\sum_{k=1}^{M} \left(\frac{d(x_{i}, c_{j})}{d(x_{i}, c_{k})}\right)^{\frac{2}{m-1}}}.$$
(11)

The next step checks if condition (12) holds, and if it does, the algorithm terminates, but if not, the iteration repeats.

$$\left| U^{k} - U^{k-1} \right| < e$$

$$\max_{ij} \left| u_{ij}^{k} - u_{ij}^{k-1} \right| < e .$$

$$(12)$$



Fig.11 - The result of fuzzy c-mean segmentation of the original image in Fig. 1a)

CONCLUSION

When comparing the texture segmentation algorithms from steps p.p.1-7 with the participation of an expert (radiologist), the algorithms from steps p.4 and p.7 turned out to be the most sensitive and accurate in picking out neoplasm areas. The threshold processing algorithms correctly picked out only the breast area in the mammogram, but did not pick out neoplasm areas.

Segmentation by *k*-means method also cannot be applied to neoplasms extraction, as it depends on the number of clusters that are set manually. The algorithm of binary homogeneity augmentation singles out together with neoplasms a part of fibrous formations, which can further introduce an error in classification.

Thus, it has been experimentally established that:

- the use of threshold segmentation algorithms is effective in highlighting only the breast area, their use in highlighting neoplasms is not feasible;

- since the mammogram contains a large number of textures, the area augmentation algorithm is also not acceptable for segmentation;

- the binary homogeneity augmentation algorithm can quite successfully highlight the marginal region of the breast, which allows detecting some signs of breast cancer, such as inverted nipple, but segmentation highlights areas of fibrosis, which makes correct diagnosis of neoplasia more difficult;

- the *K*-means clustering algorithm is not at all suitable for segmentation of mammographic images, as it is better suited for classification;

- the marker watershed and Fuzzy *C*-mean algorithms most accurately distinguish areas of neoplasia.

REFERENCES

- Fedrov A. Binarization of black and white images: status and prospects for development. Retrived from: http://itclaim.ru/Library/Books/ITS/wwwbook/ist4b/its4/fyodorov.htm (accessed December 1, 2013).
- Pratt W. Digital Image Processing: Per.s Engl. M: Mir, 1982, book 1-312p.; book 2. - p. 493.
- Akhmetshin A.M., Fedorenko A.E. Segmentation of low-contrast images by means of hierarchical Markov's random field model // Artificial Intelligence, No. 4, 2002. – pp. 414-421

- Indra Kanta Maitra et al. Automated Digital Mammogram Segmentation for Detection of Abnormal Masses Using Binary Homogeneity Enhancement Algorithm // Indian Journal of Computer Science and Engineering (IJCSE), 2011. - Vol. 2 No. 3. – pp. 416 - 427
- Patel B.C., Sinha G.R.. An Adaptive K-means Clustering Algorithm for Breast Image Segmentation // International Journal of Computer Applications (0975 - 8887) Vol.10, N.4, 2010. - pp. 35-38
- Sadykov S.S., Bulanova Yu., Romanov Yu.A. Texture segmentation algorithm to detect cyst areas on mammograms // Algorithms, methods and data processing systems. 2013. №19. – pp. 50-55.
- Durand B., Odell P. Cluster analysis. Translated from English by E.Z. Demidenko. Ed. by A.Ya. Moscow: Statistics, 1977. – p. 128.
- Orlov A.D., Kulkov Y.Y., Sadykov S.S., Samandarov I.R. Algorithm of Determination of the Side of a Flat Object of Symmetric Form. Opticselectronic devices in image recognition and image processing systems. Kursk, 2021. – pp. 281-283.
- 9. Sadykov S.S., Savicheva S.V., Samandarov I.R. Recognition of separate and superimposed real flat objects by curvature of points of their binary images contours. 2022, Gospodarka i Innowacje, 22. pp. 383-398.
- 10.Буланова Ю.А., Садыков С.С., Самандаров И.Р., Душатов Н.Т., Миратоев З.М. Исследования методов повышения контраста маммографических снимков. Oriental renaissance: Innovative, educational, natural and social sciences. 2022. Vol.2. No. 10. pp. 304-315.
- 11.Буланова Ю.А., Садыков С.С., Самандаров И.Р., Душатов Н.Т., Миратоев З.М. Исследование методов фильтрации шума на маммографических снимках Oriental renaissance: Innovative, educational, natural and social sciences. 2022. Vol.2. No. 10. pp. 177-191.