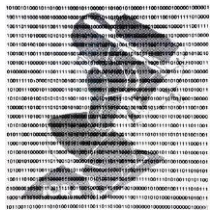
G.Y.N.A.S.

Gynaecology Analysis System

Digital Image Processing for the Detection of Micro-Calcium Deposits in Very Early Forms of Breast Cancer with an Artificial Intelligence Approach

Version 3.1, November 2002



ARTHUR G. SUTSCH
I.C.R.A.

Abstract

There is consensus among gynaecologists that breast cancer is advancing at an alarming rate. With the capability of detecting very early forms of breast cancer, an added dimension for diagnosis can be reached.

This paper discusses image processing applications for the detection of very early forms of breast cancer through micro-calcifications from mammography X-ray films or digital radiography. The films are scanned into digital values at high spatial resolution and then processed with special image processing techniques taken from astronomy as well as general digital image processing techniques. Instead of scanning film, digital radiography is recommended for its considerably lower radiation dose and already digital information content.

The results show that digital image processing techniques can detect much smaller micro-calcium deposits than the human eye can see.

Interpretation of these finds is suggested via an artificial intelligence engine in order to present the data to the gynaecologist in degrees of varying probability for breast cancer.

Key Words

BREAST CANCER, MAMMOGRAPHY, MICRO-CALCIFICATIONS, DIGITAL IMAGE PROCESSING, DIGITAL RADIOGRAPHY, X-RAYS, ARTIFICIAL INTELLIGENCE, FULL-TEXT RETRIEVAL SYSTEMS, DOCUMENT ARCHIVE SYSTEMS, LASER DISK STORAGE, DATA SECURITY, DATA COMPRESSION**.**

1. Introduction

Through discussions with experts in the field of early breast cancer research it is a well known fact that breast cancer is spreading at an alarming rate within the Western woman throughout all ages. With more knowledge accumulated on the subject, the old myth of breast cancer developing during the menopause and possibly from prolonged use of oral contraceptives gives way to a rather alarming situation: breast cancer can be found in any age group and increases dramatically within the Western civilization.

A known origin of breast cancer is the formation of micro-calcium deposits (micro-calcifications) within the breast tissue, around which degenerate cells accumulate which in turn cause breast cancer to form. Identifying those early and extremely tiny micro-calcium deposits can hold a key to possible cure of this type of breast cancer in its earliest form through various methods.

The human factor in breast cancer identification by the gynaecologist is particularly important: on many occasions, he or she has not spotted the micro-calcium deposits in a mammography; the patient comes back one year later and a lump has evolved. The tragic part is that in many cases re-inspection of the previous film does show very small micro-calcium deposits. They have been simply overlooked due to various reasons because of their extremely small size.

The work underlying this paper stems from applications in the digital processing of images in astrophysics, in detail the search for new stars in early forms (prototype stars) within gaseous nebulae. The author describes a method how to analyse the information with an image processing computer presenting results to the gynaecologist to look more closely into certain areas. Degrees of probability are attached to the finds by the computer. An artificial intelligence (AI) engine driven system is proposed for this task.

2. Expertise

The GYNAS - Gynaecology Analysis System represents an application of image processing and image analysis. It consists of a workstation platform concept of powerful computers linked through a LAN (Local Area Network) or WAN (Wide Area Network) with distributed intelligence in the various stations: server, analysis station, work station, scanning station, etc. The system contains a database of textual description and support information as well as an image database for raw and processed images with tools for the gynaecologist to analyse the images.

The author's expertise over the past 20 years spans areas such as image processing in astronomy and space applications, with archiving and intelligent retrieval concepts utilizing advanced indexing and information extraction techniques on very large volumes of data (as in Laser disk archives of computer data, text, or image form or public online information systems - Intranets). The base technology is a proprietary patented data compression process and data security; from this technology, intelligent optical quality control systems for the manufacturing industry have also been derived.

2.1 Full-Text Retrieval System

Images are always accompanied by descriptive textual databases; large amounts of stored textual data are only as usable as the retrieval language that can find relevant information. As an addition to relational database systems, a full text retrieval system with advanced queries such as proximity and phrase searching - not only on descriptors but in case of text search *on each individual word or element* with its logical surrounding in a large textual database will be implemented. An individual who does not know how to search for the term he is looking for directly can enter the search domain within a thesaurus functions and the machine will deliver relevant information to such words contained within the database.

2.2 Data Compression and Security

For reduction of data storage space and for security reasons, the data is stored in a special symbolized and encrypted form; reconstruction is only possible with proprietary software. Data compression plays an important factor in access time and cost of material (storage, LAN capacity), and communication link time, especially in the case of large volumes of image data.

The compression ratios achieved with ***loss-less*** techniques vary with source material from 1:2 to 1:40, with the higher ratios typically found in structured database applications. The compressed information contains a high degree of security against third party deciphering. For maximum data security requirements an *encryption system* can be applied and added to

the data sent or stored. This results in virtually non-decipherable data strings and efficient data security.

The application of ***lossy*** image compression algorithms is not automatically advocated in this first step of the system. Certainly, algorithms like DCT using Fourier analysis (as in JPEG image compression) are not automatically recommended due to their "chopping" of high and low frequency elements. The application of image compression algorithms such as Wavelet based algorithms has to studied in depth as to their applicability in this task.

3. Image Analysis Approach

Analysis of image and textual data requires sophisticated mathematical algorithms in order to extract relevant information. Applications span from document content analysis (after possible character conversion from bit raster data) to image analysis with Artificial Intelligence (AI) rules to get away from mere template and model matching procedures. Concepts like "fuzzy logic" (approximation of image morphology to a given reference image structure) are applied.

Interim stages for this image analysis are 2- and 3-dimensional area of interest (AOI) representations with various image processing operators. Special filtering algorithms after contrast stretching and line following AI techniques are a common feature extraction method. Image restitution from unclear images can provide surprising results on text and image data.

The basis of this analysis can be applied to a variety of applications; the computer can offer associative and logical assertions - most of these yield good results when applied correctly. In industrial applications, the decision process can be simplified and automated in most cases to arrive at on-line optical quality control systems, etc. since the surrounding data structure is well known and the case can be defined by image form and logical definitions: dimension, tolerance, size, location, etc. of a specific piece to be checked.

In medical surroundings these definitions often are not fully known or cannot be defined readily; therefore, an interactive learning process of the system with user interaction is mandatory.

The approach to follow is the creation of a known surrounding through exhaustive analysis of known situations; in the case of breast cancer, the analysis of already known micro-calcifications and their use as a *reference library* is a first step in this analysis process. From this, *models* are extracted and applied as a first step for testing the knowledge acquired. After a sufficient "hit rate" is achieved, the understanding of the process can be transferred into a more logical abstraction of these models with the help of artificial intelligence procedures (rules). In many cases, the system rule structure will change with time and experience influenced by both, the machine and the gynaecologist.

The application of the "fuzzy-logic" concept in artificial intelligence is well suited for this approach since it does not rely on mere template matching concepts and simple rule structures but allows approximations with probabilities for a hit rate. In this application, the gynaecologist will be informed of such a probability and his attention drawn to the specific image for further analysis. Depending on his degree of proficiency he can either revert to visual inspection on the computer screen or operate the computer image processing

algorithms presented to him in a format he can use without special image processing knowledge.

Neural self-learning can be combined with the above feature sets in order to arrive at an advanced, self-learning system. This is seen as a second step of implementation, however.

4. Digital Analysis

In a digital analysis approach with an image processing computer, the first steps are a manual interaction of the system with the medical specialists. It is the aim to arrive at an *automated system* that pre-processes much of the information in order to arrive at only critical decision points for the gynaecologist to intervene.

4.1 Resolution

In the enclosed GYNAS examples, radiographic data (films or digital radiography) from mammography is stored at high spatial resolution on optical disks such as CD-ROM or Worms, or DVDs. Features <=0.07 mm have been resolved optically with the image examples supplied, image resolution can go far beyond this through special weighting algorithms and sub-pixelation techniques. Resolution in these sample images is 255 *grey* levels (and then represented in false colour for contrast to the viewer) but can be extended to higher levels if so required (12 bit resolution or 4096 shades of grey/colour is used in many applications).

The major difference is the requirement for *visual representation of a radiography* or *image processing techniques by the computer*: in general the radiologist is used to seeing many fine levels of grey when he judges from a plate or its representation on a computer screen. Image processing does not require these elevated grey level structures since the mathematical algorithms operate on the same structure of data. As an example, the jaggedness of a digitised curve can be represented with many steps (high optical resolution) or fewer steps (lower optical resolution). The *slope of the line* within a region with a sufficiently large amount of pixels can be approximated to the same curve with higher and lower resolution.

4.2 Reference Catalogue

A *"reference catalogue"* of early breast cancer features is stored in an image database with associated textual description (the text database). Mathematical analysis is performed on these reference images in order to arrive at an image topology surrounding such micro-calcium features (<=0.1 mm). The resulting image is represented in a 3-dimensional form to extract information on the morphology of such an image. Reflection of X-rays at given wave-lengths from an uneven shape of calcium deposit (the possible origin of one form of early breast cancer) with the added effect of diffuse and/or direct refraction and reflection with attenuation due to tissue depth inside the breast yields a certain morphology of darkening on an X-ray film or its representation of grey-level distribution in a given *area of interest* within the image.

The various *gradients of contrast change* (mostly second derivatives of the topology equation) form a typical morphological structure when normalized for tissue depth

attenuation and thus provide an indication of size and structure of the micro-calcium deposit. Comparisons to other known features show very clear distinctions (plate defects, etc.).

4.3 Artificial Intelligence

It is this topology of the *3-dimensional grey level structures* surrounding the micro-calcium that is compared with imaging AI techniques from a reference catalogue to the actual image in an advanced stage of the system. A decision as to the actual percentage of match (*hit rate*) can be given and thus aid the diagnosis and shorten the associative thinking process of the gynaecologist to arrive at a decision point for unclear images. Possible solutions with their varying degree of match (in percent) are presented.

Upon identification by the gynaecologist the database is augmented by his judgement: the result is flagged positive for the algorithm to increase the degree of assurance for a later matching process. In a negative case, the decision is flagged negative: an input by the gynaecologist as to the reasons is required. The AI kernel functions extract the relevant elements from the comment and react the next time a similar decision arises with "I have been trained (not 'programmed'!) to decide for solution X, a negative comment has been entered causing rules U, V, W, to be re-trained. Issue further comment."

Once a negative training process has been initiated on several finds, a logical "retraining" of the rule structure with weighting algorithms has to be carried out indicating a faulty algorithm, wrong comparison structure (weighting), or just new experience has been gained on the subject. The system can thus be considered "self-learning" with an interactive user interface.

5. Analysis on Selected Images

The following images describe the approach how to isolate individual micro-calcium features in known cases of breast cancer in X-ray films and how to derive the model for detecting similar and smaller features that are not detectable with the naked eye. Information is hidden inside the image that is brought about through image processing techniques. An example of a surprising find is presented.

The language used in the description of the images is not from the gynaecologist's mammography dictionary; the author works in digital image processing.

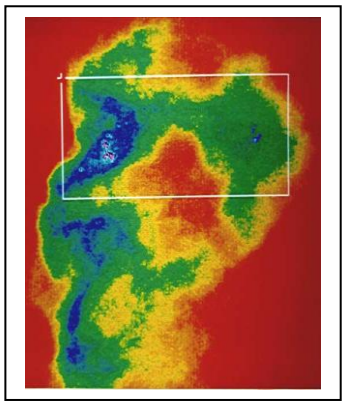
Image 1

The original X-ray film (reproduced here in Black and White) image from which all subsequent image transformations were taken. It serves as "*Master*" to exemplify how information can be extracted from the picture. Which information is visible to the eye and how additional information can be extracted through computer image processing techniques, not visible to the naked eye.

The large white central feature with clear white points, imbedded in a large light greyish neighbourhood are known micro-calcium deposits that have subsequently caused breast cancer. Such white dots can be seen also towards the right side of the picture veiled in a darker grey matter (tissue).

All subsequent colour images use *false colour rendition* to show contrast and do not have any true colour meaning.

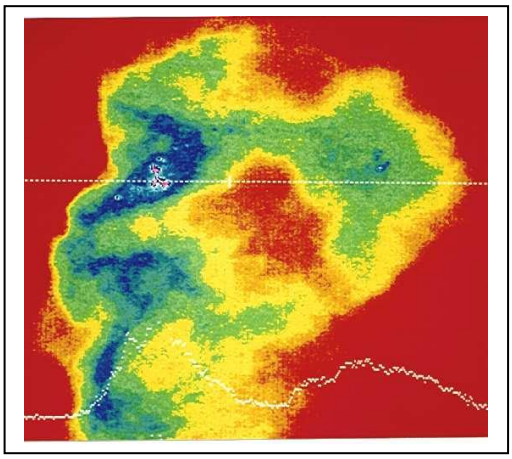
Image 2



The selection of an *area of interest (AOI)* shows the white rectangle in the picture. It is used to identify the known features and to extract their properties in three dimensional form. The image is digitised (here) with 255 grey levels (represented as colours) and subsequently treated with image processing functions. Smallest features visible in the image are 0.07 mm. The background is shown in dark red, contrast is enhanced to merge from dark red to yellow, green, and blue for the non-critical areas. The colour green inside the AOI is contrasted with the blue colour and the known lumps of calcium are shown in reddish and black against a light blue background for contrast reasons only.

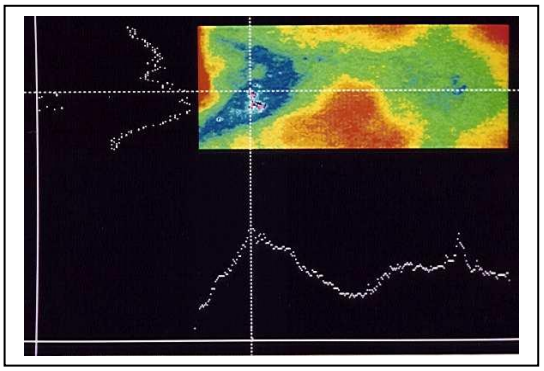
A small feature can also be seen at the right hand side against a darker green background.

Image 3



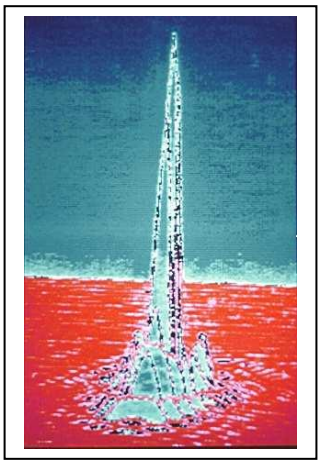
At each line within the picture, a brightness and contrast distribution curve is recorded; the dotted line intersecting the known lumps of calcium is shown and its brightness profile plotted at the bottom of the image. This process is repeated for every line in the digitised image. The same way, a *contrast gradient* is recorded. The lines form a 2-dimensional "mountain" and "valley" representation of each point along a line in the image, the mountains representing the high intensity or contrast points, respectively with values ranging from 0 to 255 for the 256 grey level digitisation process.

Image 4



The AOI is pictured with the line analysis showing the two-dimensional intersection of the contrast and brightness functions for one known point. This is recorded for every picture element (=pixel) to form a pair of lines within the picture. The 'mountains' and 'valleys' show the distribution curves for the individual image points. The reference (0) lines are the solid white lines intersecting at the lower left corner of the picture.

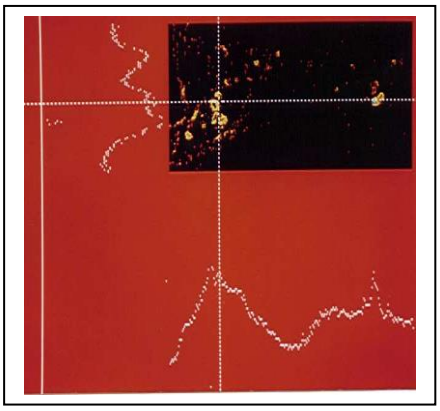
Image 5



Combining all image points along a given horizontal and vertical line and plotted against a three-dimensional background (in red) yields the well-known "Mexican hat" distribution with the high peak representing the strongest gradient of change. A *steep peak* represents a large *brightness (or contrast) change* in the image. The surrounding pattern of how the 'peak' is created gives a measure for the reflection of the X-ray on a calcium deposit. The distribution pattern in a 3-D shape also indicates the difference between any other feature in the image so that a catalogue of *common feature patterns* can be extracted from known elements on the source material. In this way, patterns for plate effects, arteries, attenuation due to plate darkening, etc. can be established after normalization for a given feature.

The comparisons of these different 3-D shapes with AI techniques yields probabilities for different classes of calcium deposits. Normalizing the procedures for the mentioned tissue depths brings about a cleaned up function for gradient changes in a given pixel neighbourhood.

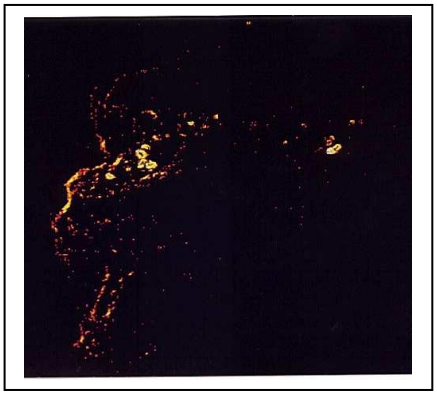
Image 6



The known elements of brightness/contrast distribution are still present. The image has been processed to bring out strong features only: the background has been normalized to black. Only the lumps have been extracted and are shown colour inverted into greens and blues in their centres.

It now becomes apparent that there are *many more tiny little calcium deposits* visible through the image processing algorithms that were not visible to the eye before. As mentioned earlier, features as small as 0.07 mm show up.

Image 7



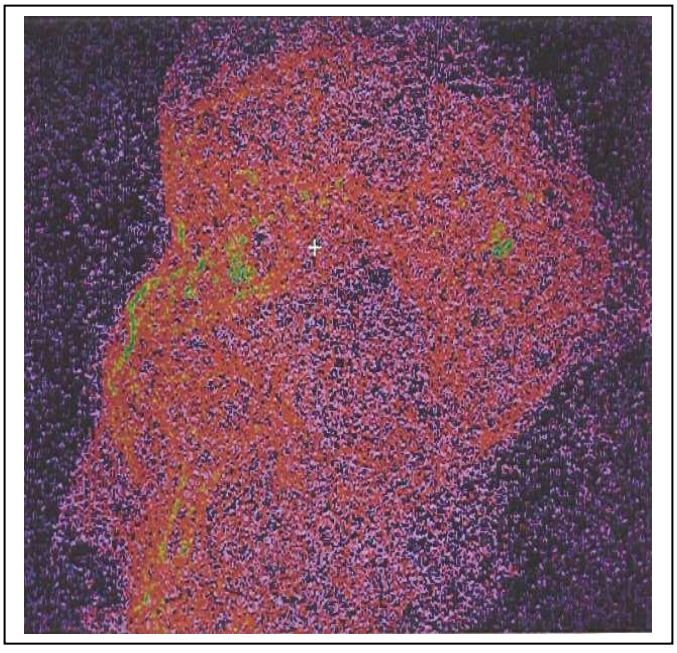
The isolation technique is applied to the entire image (ref. image 2 and 3) in order to show more micro-calcium deposits. Besides the known features, a 'ridge' appears towards the left side of the image with very pronounced features of calcium deposits. The lower part of the image also exhibits a series of very small micro-calcium deposits invisible to the eye (in the form of individual single points after the normalization and feature extraction process).

Image 8

The same image as image 7 but in black and white and inverted shows the features but they lack the information content of the colour: the added colours of image 7 also indicate area information due to the colour inversion technique. It is, therefore, recommended to carry out this analysis on a colour monitor. Features have been normalized with a lesser degree in order to show small areas of possible interest to the gynaecologist.

Image 9

The processing techniques were changed completely after isolated points have been extracted in the image. Interested features were connections between probable candidates for micro-calcium deposits. The background in this processed image is not 'cleaned up', the dark is speckled with blue/red dots; their density distribution is a measure for the transgressing of rose into deep red.



This image is the most interesting and yet most intriguing: it is quite obvious that **micro-calcium deposits pictured in green /yellow are interwoven in a cocoon of bright red filaments**. The stronger the red filaments are grouped together, the higher the degree of micro-calcium deposits.

The inverse question is: if such a (red) cocoon can be detected, is there also a micro-calcium deposit region? Questioning for the cause of such a cocoon structure has not brought about any direct results from mammography specialists. This shows that information lies hidden in images that can be brought about by digital image processing techniques.

6. Conclusion

A method for extracting relevant information from mammography X-ray film or digital radiography for very early forms of breast cancer with micro-calcium deposits has been presented through the use of digital image processing techniques taken from astrophysics. The results show that a possible origin of breast cancer, micro-calcification features smaller than the human eye can detect, can be extracted through digital image processing techniques.

Not only does this proposed method show very small micro-calcium deposits in the breast tissue, but it also generates new questions for the gynaecologist to resolve, like a "cocoon" structure that seems to embed micro-calcium deposits; it requires further research as to the importance of such a feature.

The combination of image processing with Artificial Intelligence interpretation techniques allows automated pre-processing of X-ray data to alert the gynaecologist to areas of interest that he might miss due to the small detail contained in the information.

The image part is combined with powerful full text retrieval methods to link images to text and vice versa. A database with self-learning features forms the basis for accumulation of knowledge within the system. For security and fast access to the data, no-loss compression is applied for storage, LAN and WAN transport.

Table of Contents