COMPUTER SCIENCE AND DATA ANALYTICS

INTERIM REPORT
COURSE: GUIDED RESEARCH I

Arzu Fatullayeva
arzu.fatullayeva@gwmail.gwu.edu

Baku 2023

Topic : Finding informative regions in grayscale mammogram images.

1. The description of the problem

The problem addressed in this research is the identification and localization of informative regions in grayscale mammogram images. Mammograms are essential diagnostic tools used for breast cancer screening and detection. However, the interpretation of mammogram images can be challenging due to their inherent characteristics, such as low contrast, noise, and the presence of various anatomical structures.

The current approach to mammogram analysis mainly involves manual inspection by radiologists, who visually examine the entire image to identify suspicious areas indicative of potential breast abnormalities. This process is time-consuming, subjective, and can lead to variations in interpretation, potentially affecting the accuracy and reliability of breast cancer diagnosis.

To address these limitations, the proposed research aims to develop a methodology that automates the process of identifying informative regions in mammogram images. Informative regions refer to specific areas within the images that contain crucial information relevant to breast cancer detection and diagnosis. By identifying these regions accurately, radiologists can focus their attention on the most relevant areas, potentially improving the efficiency and effectiveness of their analysis.

To tackle this problem, the research will explore the use of keypoint detection algorithms combined with feature extraction and machine learning techniques. Keypoint detection algorithms will identify salient points or regions within the mammogram images, which are likely to contain significant information. Feature extraction methods will capture the local characteristics surrounding these keypoints, enabling the representation of informative regions. Machine learning algorithms will be applied to classify or segment the informative regions based on labeled data or unsupervised clustering techniques.

2. The standard ways of approach

1. Region-based Approaches:

Region-based methods focus on segmenting the mammogram images into regions of interest. Various segmentation algorithms such as thresholding, region growing, and active contour models have been used to extract potential informative regions. After segmentation, features such as texture, shape, and intensity can be extracted from the segmented regions. Machine learning algorithms, including SVM, random forests, and neural networks, are then employed to classify the regions as informative or non-informative.

In research work by G. R. Jothilakshmi , to segment the affected area in a mammogram efficiently, an automated approach is employed using a split and merge technique based on region-based segmentation. This method involves identifying a seed point to initiate the process. The proposed algorithm utilizes morphological operations to digitally eliminate noise and applies region split and merge technique to separate the affected region from the background in the image.

2. Texture Analysis Approaches:

Texture analysis methods aim to capture the textural patterns within the mammogram images to identify informative regions. Techniques like co-occurrence matrices, Gabor filters, and local binary patterns have been used to extract texture features.

Classification algorithms, including k-nearest neighbors, decision trees, and ensemble methods, are often applied to classify the extracted texture features into informative and non-informative regions

3. Feature-based Approaches:

Feature-based methods involve extracting specific features that are indicative of informative regions in mammogram images. Examples include shape features, such as circularity or irregularity, and edge-based features, such as the presence of spiculated edges.

4. Deep Learning Approaches:

Deep learning techniques, particularly convolutional neural networks (CNNs), have gained popularity in recent years for various medical image analysis tasks, including mammogram analysis.

5. Ensemble Approaches:

Ensemble methods combine multiple classifiers or approaches to enhance the accuracy and robustness of informative region detection. Ensemble techniques, such as bagging, boosting, or stacking, have been employed to fuse the outputs of multiple classifiers or feature extraction methods, leading to improved performance in identifying informative regions.

3. Dataset Collection, Cleansing and Statistics

The data is collected from the RSNA Screening Mammography, which is Kaggle dataset. The mammograms there in **dicom** format. The way that DICOM organizes information into data sets makes it different from other image formats. A header and image data sets are combined into one file to form a DICOM file. The header's content is arranged using a consistent, standardized series of tags. One can acquire vital information about the patient demographics, research parameters, etc. by extracting data from these tags.

It can be expected roughly 11,900 patients in the train set. There are usually but not always 4 images per patient. It should be mentioned that the RSNA dataset has been provided without any noise and contained no patient information.

However, using all this images for research purpose requires computational and storage power. So for this research it will be used 1000 images, 500 positive and 500 negative ones. For this aim the train.csv file is used, and the column “cancer”, which shows whether or not the breast was positive for malignant cancer, helped to extract 1000 images (Code is in github)

When it comes to the next step, images were converted from dicom format to png. Although it results in larger file sizes, its lossless compression allows for superior image quality.

In computer vision, we frequently work with medical images, and nearly all of our databases feature DICOM images. The Dicom image is more than just an image; it contains the pixels information, the patient information, and other things.

In this situation, all that is required of us is to view the image or save it as a PNG file so that we can access it using any software we like.

We require extra Python libraries to accomplish that. The first one is Pillow, which I prefer to use for showing and saving JPG images, and the second one is Pydicom, a library designed specifically for Dicom images. The third option is Numpy, which manipulates arrays.

When it comes to data statistics, the number of negative ones is 53548 , positive ones- 1158, which shows how the dataset imbalanced.

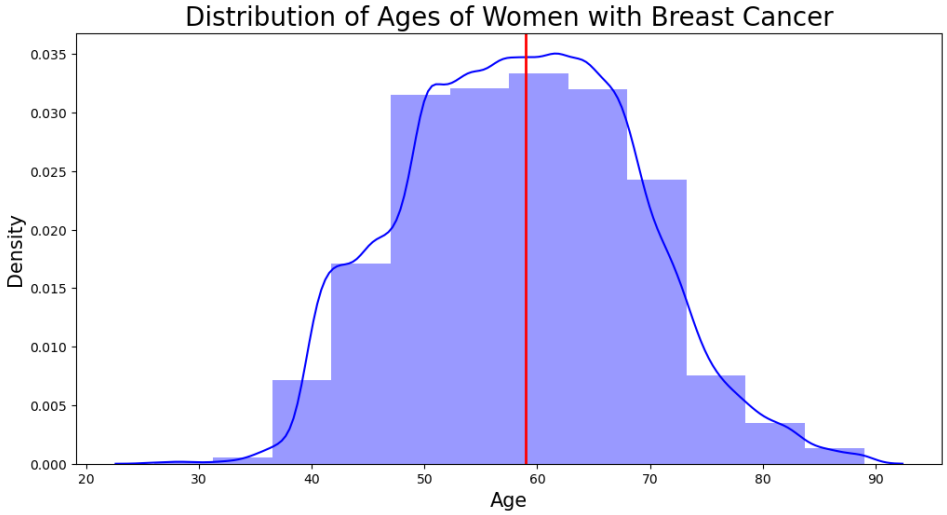

Fig. 1. distribution of women's ages with cancer

People of ages between 48 to 73 are most suffer from Breast Cancer as its shown on Fig.1.

4. Feature extraction (about ORB)

In 2011, Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski created Oriented FAST and Rotated BRIEF (ORB) as an effective and practical replacement for SIFT and SURF. ORB was developed mostly due to the patented nature of SIFT and SURF. ORB performs as well as SIFT on the task of feature detection (and is better than SURF) while being almost two orders of magnitude faster. ORB builds on the well-known FAST keypoint detector and the BRIEF descriptor. Both of these techniques are attractive because of their good performance and low cost.

The FAST algorithm analyzes a pixel, p, by comparing its brightness to the surrounding 16 pixels arranged in a circular pattern. These pixels are classified into three categories: lighter, darker, or similar to p. If more than 8 pixels in the circle are either brighter or darker than p, p is identified as a keypoint. Therefore, the keypoints detected by FAST provide us with information about the locations where edges are determined in an image.

Despite lacking an orientation component and multiscale features, the FAST algorithm is enhanced in the ORB algorithm by utilizing a multiscale image pyramid. This pyramid represents the image at different resolutions through a series of downsampled versions. ORB applies the FAST algorithm to detect keypoints in each level of the pyramid, effectively identifying keypoints at various scales. This approach allows ORB to achieve partial scale invariance by locating keypoints across different scales in the image.

Once keypoints are located, ORB assigns an orientation to each keypoint based on the directional change of intensity levels around it. This is determined by using the intensity centroid, which assumes that the intensity of a corner deviates from its center. By calculating the vector of this deviation, an orientation can be inferred and assigned to the keypoint.

Once keypoints are located, ORB assigns an orientation to each keypoint based on the directional change of intensity levels around it. This is determined by using the intensity centroid, which assumes that the intensity of a corner deviates from its center. By calculating the vector of this deviation, an orientation can be inferred and assigned to the keypoint.

After applying ORB keypoints with size and without size are shown on Fig. 2 for one image:

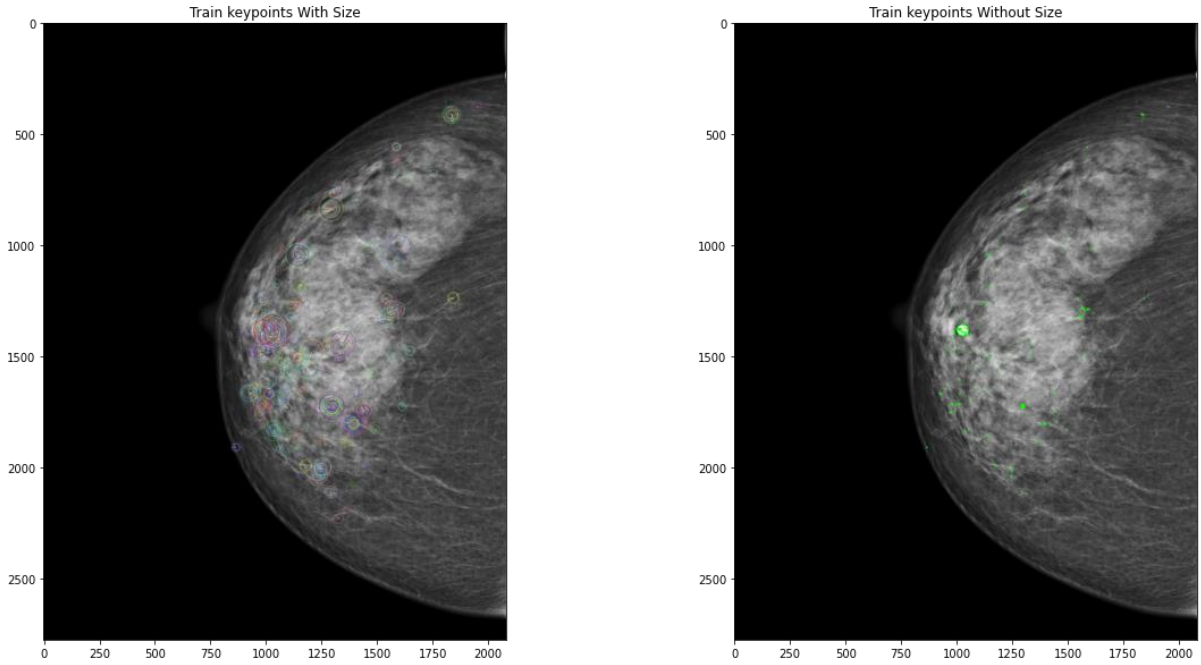

Fig. 2. Keypoints with and without size

5. Further work

Further step will be application of AI/ML model for the classification (input: ORB, output yes/no) and showing the efficiency of the approaches by calculating accuracy, recall, precision, f1-score .

[1] Jothilakshmi, G. & Sharmila, P. & Raaza, Arun. (2016). Mammogram Segmentation using Region based Method with Split and Merge Technique. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i40/99589.

[2] Justaniah, Eman & Alhothali, Areej & Aldabbagh, Ghadah. (2021). Mammogram Segmentation Techniques: A Review. International Journal of Advanced Computer Science and Applications. 12. 520-529. 10.14569/IJACSA.2021.0120564.

[3] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2564-2571, doi: 10.1109/ICCV.2011.6126544.

[4] Kim YJ, Kim KG. Detection and Weak Segmentation of Masses in Gray-Scale Breast Mammogram Images Using Deep Learning. Yonsei Med J. 2022 Jan;63(Suppl):S63-S73. doi: 10.3349/ymj.2022.63.S63. PMID: 35040607; PMCID: PMC8790585.

[5] https://www.kaggle.com/competitions/rsna-breast-cancer-detection