Annotation Parsing in Clinical, Real-World Breast Ultrasound Imaging Data

Kailee Hung[1], Arianna Bunnell[12], Peter Sadowski[1], John Shepherd[2]
University of Hawaii, Honolulu, HI, USA[1], University of Hawaii Cancer Center, Honolulu, HI, USA[2]

Introduction

- Ultrasound (US) is a viable imaging modality to mammography for the detection of breast cancer in resource-limited settings.
- Clinical US images often contain annotations from examining sonographers that contain information about scanning protocol and conditions.
- Identification of sonographer text annotations may aid in data cleaning for Artificial Intelligence (AI).
- Text annotations in clinical US images is often partially cut off due to Protected Health Information (PHI) removal protocol.
- The goal of this research is to establish a pipeline for identifying and parsing annotations in clinical breast US scans.

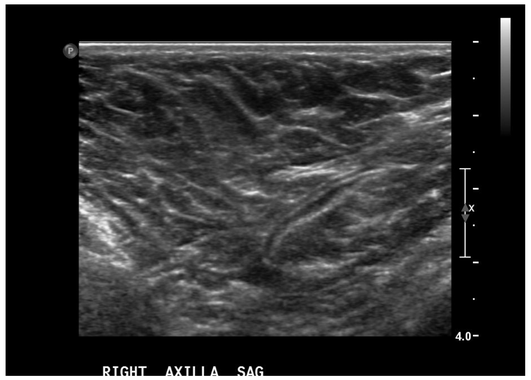
Methods

Text extraction methods were developed through observation of a set of over 100,000 breast US images from the Hawaii & Pacific Islands Mammography Registry (HIPIMR).

1. Black padding of 70 pixels was added to the bottom of the image to aid in identifying cut-off characters.
2. The EasyOCR (Jaided AI; Bangkok, Thailand) optical character recognition tool was applied to each scan.

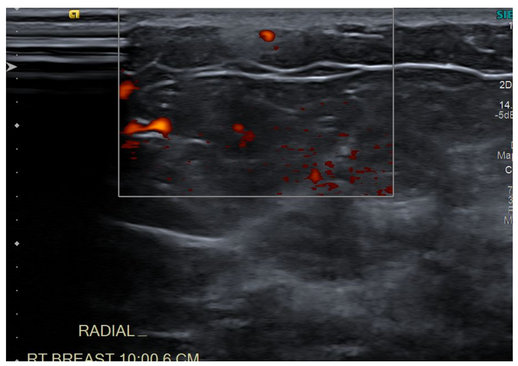
Sonographer annotations describing the structured scanning protocol was categorized into **5 types**: **laterality**, **axilla presence**, **transducer orientation**, **distance from nipple (CMFN)**, **clock position**. A variety of regex patterns were also employed to account for the variety of expected text and incomplete text.

1. For each of the 5 categories: check for the pattern in the detected text from the image. See **Figure 2** for reference.
2. If there is a match, reformat the text as needed. See **Figure 1** for reference.
3. Remove the matched pattern.
4. After checking for all the patterns, the remaining text is classified as miscellaneous.



4.0 - RtGHT AXTI 4 SAG
[RIGHT, SAGITTAL, AXILLA, , , 4.0 -]

4.5. 241 cmeft Rreast 1.00 4 CM FN 4 Rad
[LEFT, ANTIRADIAL, , 4 CMFN, 1:00, 4.5. 241]



9 SII 2D 14. ~SdE Mal C 1 RADIAL RT RRFAST 10*nn 6 CM
[RIGHT, RADIAL, , , , 14. ~sde mal c 10*nn]

AL RT BR 9 : 0 0 8 . 5 CM FN ARADI
[RIGHT, , , 5 CMFN, 9:00, 8 . aradi]

Figure 2: Visualization of how text was read and parsed into meaningful categories. Under each ultrasound image, the first line is the raw string read by EasyOCR and color coded depending on type of text. The second line is the return array with the text formatted and with the same color coding

Figure 1: ACR Labeling and Measurement Standards For Ultrasound [1]

Labeling for breast US images may contain the following descriptive fields:

- **Laterality**: Designation of left or right breast being examined
  - Rt, Right ➤ RIGHT Lt, Left ➤ LEFT
- **Axilla**: Refers to the armpit region and indicates examination of lymph nodes
  - Axilla, Axillary ➤ AXILLA
- **Transducer Orientation**: Angle that the ultrasound transducer is positioned
  - Rad ➤ RADIAL Arad, Antirad ➤ ANTIRADIAL Sag ➤ SAGITTAL
  - Trans, Trns, Trv ➤ TRANSVERSE
- **CMFN**: Distance from the nipple to the abnormality or the area being scanned in cm
  - 8 cmfn, 8cm fn ➤ 8 CMFN 7-8cmfn ➤ 7-8 CMFN
- **Clock Position**: Anatomic location using clock-face notation
  - 7:00, 7o'clock ➤ 7:00

Results

Table 1. Sensitivity and Specificity of Fine-Tuned EasyOCR on Held-Out Test Set

	Laterality	Axilla	Transducer Orientation	CMFN	Clock Position
Sensitivity (# True Positives)	96.85% (1,632)	97.44% (190)	97.05% (1,351)	93.99% (720)	93.58% (1,167)
Specificity (# False Positives)	100.00% (0)	100.00% (0)	99.34% (4)	97.24% (34)	99.87% (1)
Image Count (% of Images)	1,685 (84.25%)	195 (9.75%)	1,396 (69.80%)	800 (40.00%)	1,248 (62.40%)

The text annotation extraction pipeline was validated on a randomly-selected, hand-labeled subset of 2,000 breast US images from the HIPIMR dataset. Generally, some failures could be attributed to basic scanning errors both due to text cropping or text cursor presence. Other failures were due to oversights in the code for circumstances unaccounted for (cm/n, ftn, fn). Additionally, clock position was excluded if there were two instances leading to false negatives.

Conclusion

These results show the efficacy of our domain-specific text recognition pipeline and may improve breast US data for AI model development.

Improvements

- Further refinement of the pipeline, namely in CMFN, and clock position when it comes to handling multiple values (6:00-7:00) and cm/n, ftn, fn.
- For laterality, significant improvement would be seen after accounting for the "+ breast" pattern.

Future Developments

- Bounding box coordinates were returned, so a system needs to be developed to crop text out of the images.
- Detection and removal of lesion annotations.
- Plans to release code via an open-source license for research use.

References

1. CJ DO, EA S, EB M, Morris EA, al. e. ACR BI-RADS ® Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology; 2013.