F3-H: AIT Ground Truth Effort

Abstract—X-ray backscatter (XBS) and millimeter wave (mmW) whole body imaging are currently deployed worldwide to screen individuals at security screening checkpoints. The detection of impermissible objects carried by passengers via XBS and mmW screening shares many commonalities with medical radiological screening and diagnosis. A primary objective of this project is to leverage radiological imaging science onto the whole-body screening problem. Further, application of radiological imaging science may also lead to metrics which could be applied to all XBS and mmW equipment.

An assumption which underlies this study is that improvements in the ease with which objects of interest can be visualized in XBS and mmW datasets will lead to a corresponding increase in the probability of detection and a decrease in the probability of false alarm. For this study we utilized XBS and mmW datasets acquired by Sandia National Laboratory for this purpose. The dataset was specifically acquired to permit comparison of coupons present on given individuals as imaged by equipment from four vendors. The dataset consists of ~1500 distinct combinations of (1) person, (2) coupon (i.e. object from a defined list), and (3) placement of coupon on body (again, from a defined list of placements); each combination is designated as a “case” (neither the coupon list nor coupon placement list is contained in this report.) Each case was acquired in prompt sequence on each of four TSA-certified systems (two XBS-type and two mmW-type) that were employed in the manufacturer-recommended workflow. Within this overall dataset, we examined 30 selected cases in detail, which contained 30 separate placements of 74 total distinct coupons. This data was examined by four independent research scientists in Massachusetts General Hospital’s Department of Radiology. This work was performed at the SSI level, and an SSI final report was generated and sent to the ALERT Program Manager.

I. PARTICIPANTS

<table>
<thead>
<tr>
<th>Faculty/Staff</th>
<th>Name</th>
<th>Title</th>
<th>Institution</th>
<th>Email</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Richard Moore</td>
<td>Dir. MGHBI</td>
<td>MGH</td>
<td><a href="mailto:rhmoore@partners.org">rhmoore@partners.org</a></td>
<td>508.572.9317</td>
</tr>
<tr>
<td></td>
<td>Alyssa White</td>
<td>Staff</td>
<td>MGH</td>
<td><a href="mailto:arwhite@partners.org">arwhite@partners.org</a></td>
<td></td>
</tr>
</tbody>
</table>

II. PROJECT OVERVIEW AND SIGNIFICANCE

The primary objective of this project is to leverage radiological imaging science onto the whole-body screening problem. Further, application of radiological imaging science may also lead to metrics which could be applied to all XBS and mmW equipment. An assumption underlying this project is that improvements in the ease with which objects of interest can be visualized in XBS and mmW datasets will lead to a corresponding increase in the probability of detection and a decrease in the probability of false alarm.

For this second conspicuity study (a follow-on to a previous study) we utilized XBS and mmW SSI datasets acquired by Sandia National Laboratory. The datasets was specifically acquired to permit comparison of
coupons present on given individuals as imaged by equipment from four vendors. The dataset consists of ~1500 distinct combinations of (1) person, (2) coupon (i.e. object from a defined list), and (3) placement of coupon on body (again, from a defined list of placements); each combination is designated as a “case” (neither the coupon list nor coupon placement list is contained in this report.) Each case was acquired in prompt sequence on each of four TSA-certified systems (two XBS-type and two mmW-type) that were employed in the manufacturer-recommended workflow. Within this overall dataset, we examined 30 selected cases in detail, which contained 30 separate placements of 74 total distinct coupons. This data was examined by four independent research scientists in Massachusetts General Hospital’s Department of Radiology.

The following were accomplished during the course of the 2nd conspicuity study:

- Representative cases were selected from a 1500 case repository purpose-collected for assessment of coupons located on persons.
- A robust methodology was applied for subjectively assessing the difficulty with which coupons can be detected on XBS images; this methodology involved defining a subjective conspicuity scale.
- Four independent readers examined 30 cases involving 74 separate coupon placements from the Sandia whole-body screening dataset and provided conspicuity measures.
- Based on these conspicuity assessments, conclusions were reached about the consistency of reader assessments, and the relative conspicuity of coupons as visualized by 4 vendor systems

Based on the data examined and the methodology used, the following conclusions were reached:

- Subjective conspicuity measurements
  - A reproducible process for assessing the visual conspicuity of coupons was established.
  - Subjective assessments from the four readers were found to be highly consistent (p < 0.001).
  - Coupon conspicuity was divergent between modality (some better, some worse, some similar), suggesting the potential for multimodality applications.

Study recommendations for further developments and areas of focus:

1. Several development efforts may be undertaken to increase the detection rate of objects in XBS data:
   a. Algorithms may be developed which can automatically pre-screen an image to identify locations of potentially anomalous objects for the TSO to further interrogate, either via greater attention in the XBS screening process, or through another process within the chain of checkpoint procedures.
   b. Coupon material, coupon placement, and subject poses have an impact on conspicuity. Requiring more poses will have a negative impact on the screening process. An alternative solution may be the use of multiple detectors, concurrently - either of the same modality or of different modalities (e.g., mmW combined with XBS). Such sensor fusion may result in a more comprehensive view of the passenger being screened.
   c. Pre-processing algorithms seem highly promising. With greater optimization and testing, pre-processing algorithms may realistically improve conspicuity. Human factors issues associated with the deployment of pre-processing algorithms will also need to be examined.

2. This study was limited in scope, duration, funding, and access to other resources. Potential improvements in study design include:
   a. This study was conducted by four readers who have no prior experience in XBS or mmW screening. A comparison of trained TSO personnel performance on the same dataset would inform the generalizability of the conclusions.
   b. Further, if properly conducted, an assessment of the reading relationship between the medically-experienced readers and the trained TSO personnel might permit future “calibration” of medically-
experienced readers

c. To control for the lack of checkpoint experience of the readers, the conspicuity experiments were conducted in such a way that the readers were prompted to examine specific anatomic regions for coupon detection (by using the coupon placement configuration cartoon which came with the Sandia XBS dataset). For this and other reasons this study could not obtain probability of detection/probability of false alarm (PD/PFA) statistics. It would be of interest to perform a blinded study in which the readers were not prompted to particular locations, and coupon locations were known with high accuracy, so that robust PD/PFA statistics may be obtained.

d. A more systematic and exhaustive assessment over a broader range of coupons will likely provide greater insights into the detection capabilities of XBS and mmW screening systems. This might include not only coupons of the same material but of different sizes, and should also look at the impact of articles of clothing for masking objects, or whether the presence of moisture (moist towel or shirt with perspiration mimicking the backscattering characteristics of tissue) has an impact.

e. The methodology developed in this study generalizes to other screening modalities. This study may be repeated with mmW data in order to assess conspicuity of coupons in such datasets.

III. RESEARCH AND EDUCATION ACTIVITY

A. State-of-the-Art and Technical Approach

Using XBS, we previously demonstrated the ability to produce subjective conspicuity assessments with good inter-reader correlation. Conspicuity was assessed on a 5-point scale. We use the same assessment paradigm to test the null hypothesis that coupons cannot be observed with greater conspicuity when both XBS and MMW are used than when only one modality is used.

The 5-point conspicuity scale is defined by a subjectively score rendered by each reader using the following definition:

- Score = 1: cannot be separated from background
- Score = 2: marginal contrast difference from background
- Score = 3: average contrast difference from background
- Score = 4: clear contrast difference from background
- Score = 5: maximal contrast difference from background

We define a “case” of the data as:

A subject with the same coupons in the same coupon placements imaged by both XBS and mmW in approximately the same pose.

Overall design of current study

Four readers from the Department of Radiology at MGH, all with medical imaging-reading backgrounds, independently tabulated subjective conspicuity assessments of each coupon for each case (thereby for both mmW and XBS). Combination ratings were made post-hoc to keep the measures independent. Forgetting was marginally operative, as the cases were read over a shortened timeframe due to programmatic constraints. Nevertheless, this pilot was intended to measure conspicuity “unblended,” with physical coupon pre-designation. This is in no way a detection study and the study design purposely diverges from the TSA detection task.

To test the operating null hypothesis, the frequency of occurrence of single-modality low-conspicuity coupons would be compared to that of the joint-modality XBS/mmW. A statistically significant decrease in the
occurrence of low-conspicuity coupons in the joint data, relative to each modality separately, will indicate that the null hypothesis is incorrect, and that there is value to fusing XBS with MMW. Since readers are prompted by the “ground truth” cartoon showing the configuration of the coupons, a false alarm assessment will not be possible.

Data set characteristic

The 4-vendor, 2 modality Dataset was acquired by Sandia entitled “AIT data Collection on Mock Passengers for characterization Studies” by Jeff Jortner, Jerry Friesen, John Didlake and Cheryl Lari was referred to by us as “Sandia dataset # 3” This was due to receiving an incorrect dataset, “Sandia Dataset#2,” which was replaced when the error was detected, but for which the chosen labeling “Sandia DataSet #2” had to be retained for housekeeping and SSI-tracking purposes.

The Sandia dataset #3 which was used for this study involved two modalities: X-ray Backscatter (XBS) and Millimeter-wave (mmW). We employed two XBS systems, from vendors AS&E and Rapiscan, and two mmW systems, from vendors L3 and Smiths.

Homer Pien, PhD (MGH) chose a sub-set of 30 cases from the total 1536 provided. A “case” is one subject with a particular coupon configuration imaged by all four scanners.

Case Selection

30 cases were examined in total, each of which presented all 4 of the imaging systems (V1, V2, V3, and V4) with the same particular coupon configuration. All configurations employed 1-3 coupons. This traverse encompassed 74 coupons total. The resulting sample size on which conspicuity was independently scored was 296 (74 coupons x 4 modalities).

The cases were approximately selected from the overall dataset with attention to the following details:

- Looked for a physical and positional distribution of coupons
- A distribution of Body Mass Indices (BMIs)
- Looked for a distribution of imaging dates
- Preference for configurations with multiple coupon placements for high reading yield

Reader Backgrounds

For this project we selected a group of readers who are all trained in analyzing radiological images, but with a variety of experiences. Four individuals were recruited for this experiment, 3 PhD’s, all associated with the Department of Radiology at Massachusetts General Hospital. One of the data subsets had to be re-read due to presentation limitations. Reader #3a was replaced for that portion of the readings with an alternate, Reader #3b. The backgrounds of the individuals are as follows:

- Reader 1
  - Ph.D. in Computer Science and Engineering
  - Expertise in pattern recognition, machine learning, and image processing
  - Seven years experience in medical imaging and image processing
  - Currently a 3rd-year post-doctoral research fellow at MGH.
- Reader 2
  - Ph.D. in Biomedical Engineering
  - Expertise in CT physics, ultrasound, and image processing
  - Ten years experience in medical imaging and image processing
  - Currently research scientist at MGH and Instructor at Harvard Medical School
• Reader 3a
  ◦ Ph.D. in Computer Science
  ◦ Expertise in image processing, computer vision, and imaging systems
  ◦ 18 years experience in medical imaging
  ◦ Research Director at MGH and Assistant Professor at Harvard Medical School

• Reader 3b
  ◦ Expertise in image processing, computer vision, and medical imaging systems
  ◦ 35 years experience in medical imaging
  ◦ Breast Imaging Research Director at MGH Breast Imaging Lab (25 years)

Workflow and Work Area
An office containing a PC-based computer system, with multiple large-screen monitors, was selected for this study. During the image assessment sessions, the office door was closed to prevent interruptions, the lights were dimmed, and the window shades were drawn. All data resided on an encrypted USB drive, and the computer was disconnected from the local area network during the sessions. A timer was set so that the readers took a 5-minute break every 60 - 90 minutes.

Two computer monitors were used - the cartoon .gif file of the coupon placement configuration was shown on the first monitor, and the preview jpeg images (two or four, depending on the experimental trial and the number of poses being examined) were displayed on the second monitor (See Figure 1).

The cases to be examined were selected beforehand by Dr. Pien. Prior to the reading sessions, a Microsoft Excel spreadsheet was created for each reader, in which the columns represented the 21 distinct coupons, and each row represented the configuration number to be assessed. Since not every coupon was placed on every subject for every run, in order to avoid mislabeling the subjective assessments, for each run, the coupons which occurred in that run had their corresponding cells in the spreadsheet colored in gray.

The work area was chosen so as to ensure all three readers could gather around the computer screens concurrently. Each reader had a laptop computer, and the spreadsheet corresponding to that experimental trial was loaded. The individual readers navigated their own response-recording spreadsheet. The sheet was saved periodically and before breaks. The coupon placement configuration cartoon and the corresponding XBS preview images were displayed in the order dictated by the Excel spreadsheet (in case-number order), and real-time assessments were performed and denoted within the spreadsheets. Each reader called out that he or she has completed the assessment for that run when assessments for all coupons were entered into the spreadsheet. When the last reader indicated that he or she had completed the data entry, the next run was displayed and the process repeated. No time limit was imposed on the readers. Manipulation of the images (changing windowing levels, zooming) was not permitted during these subjective assessment sessions, after the initial windowing was completed.

Figure 1: Office area used for study.
Reading Workflow

All Images were read concurrently by 3 of 4 readers, all from the department of radiology at Massachusetts General Hospital. Reader #3 (Dr. Pien) left MGH, before the Smiths’ data re-read. For the rerun of the Smiths’ data only, Mr. Moore was substituted for Homer Pien. Mr. Moore was verbally trained to establish examples of scoring for consistency of the Reader 3 sample datasets.

The configuration cartoons provided with the datasets were used as ground truth to direct visual assessment. All readers performed independently.

Each reader held a laptop computer displaying a tabulation spreadsheet designed to record the reader’s scores without comparison to the other 2 viewers. No discussion was permitted until the score for a given presentation was locked.

This pilot was intended to measure conspicuity “unblended,” with physical coupon designation. Therefore, this study is not a classic “unknown” detection study. Its study design purposely diverges from the TSA detection task to factor out the TSO skillset.

Images were initially read in their raw format using standard open-source visualization software called Image-J (Image-J Version 1.44, National Institutes of Health.) This software automatically makes a simple measure of the numerical dynamic range of the intended file and sets the lookup table (LUT) to present this range as black-to-white. For fixed frames this is adequate. For video datasets made of multiple frames, the brightness range assigned to the first frame may not be appropriate for the whole series of frames. When this inadequacy was detected for the Smiths’ video datasets, an additional step was taken to move the video to the first frame containing actual data. The LUT was then set based upon this first “real” frame by Image-J’s standard automatic algorithm.

B. Major Contributions

The following were accomplished during the course of this five-month study:

- A robust methodology was applied for subjectively assessing the difficulty with which coupons can be detected on XBS images; this methodology involved defining a subjective conspicuity scale.
- Four independent readers examined 30 cases involving 74 separate coupon placements from the Sandia whole-body screening dataset and provided conspicuity measures.
- Based on these conspicuity assessments, conclusions were reached about the consistency of reader assessments, and the relative conspicuity of coupons as visualized by 4 vendor systems.
- An assessment was made of the extent to which changes in window levels may improve conspicuity.

Based on the data examined and the methodology used, the following conclusions were reached:

- Subjective conspicuity measurements
  - A reproducible process for assessing the visual conspicuity of coupons was established.
  - Subjective assessments from the three readers were found to be highly consistent \((p < 0.001)\).
  - Coupon conspicuity was divergent between modality (some better, some worse, some similar), suggesting the potential for multimodality applications.
- We examined several pre-processing techniques in an attempt to improve image contrast.

Numerous other conclusions can be found in the Final Report delivered to ALERT. These conclusions include factors which impact coupon conspicuity, the effect of windowing on conspicuity, and the effect of contrast-enhancing pre-processing.
IV. RELEVANCE AND TRANSITION

The use of conspicuity metrics allows a quantitation of the level of difficulty with which coupons may be detected on passengers. This methodology readily extends to other AIT modalities which are screened by TSO's. Whole-body imaging scanners which purport to improve image quality, or processing algorithms which purport to enhance visibility, can be assessed using the methodology developed.

This work moved from the SSI domain into the secret domain, at which time David Castanon visited Harry Martz at LLNL to review the results and generate a classified final report.

V. LEVERAGING OF RESOURCES

Previous conspicuity study. Dataset from Sandia.

VII. PROJECT DOCUMENTATION AND DELIVERABLES

N/A

VIII. REFERENCES


