Comparison of AATR Scoring Metrics based on Image Similarity versus Object Overlap

David Paglieroni

Lawrence Livermore National Laboratory

October 17, 2018

Lawrence Livermore National Laboratory

NCI Nondestructive Characterization Institute

LEDP Livermore Explosives Detection Program

LLNL-PRES-759800

Contacts:
martz2@llnl.gov
paglieroni1@llnl.gov

So What? Who Cares?

- Space:
  - Scoring metrics are needed for comparing performance of different adaptive automatic threat recognition (AATR) algorithms on CT images of baggage.[1-3]

- Problem:
  - The traditional scoring metric compares image segments. It computes overlap between AATR and ground truth objects. However, the score depends on the specified overlap threshold, and over-segmented objects often contribute to false alarms but not to detections.

- Solution:
  - The proposed scoring metric compares image voxels. It computes the similarity between images of AATR alarm voxels and images of ground truth voxels. P_D reflects the fraction of ground truth volume alarmed on. P_{FA} reflects the fraction of background volume alarmed on (excluding air and low attenuating voxels).

- Results:
  - For the proposed metric, there are no thresholds on object overlap, and objects can still be counted as detections even if they are over-segmented. Also, the metric can be applied no matter how the alarm value of a voxel is defined.

- Contact: paglieroni1@llnl.gov, martz2@llnl.gov

Segment-Based AATR Scoring Metric (based on Segment Similarity)

- Used in ALERT's ATR and AATR projects.

- Attempts to mimic TSL's subjective scoring methodology.

ground truth object
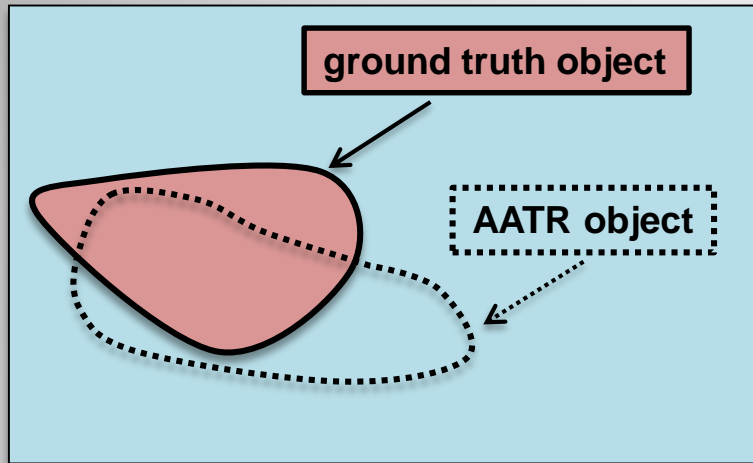
AATR object

$$precision \; = \; \frac{overlap\ volume}{AATR\ object\ volume}$$

$$recall \; = \; \frac{overlap\ volume}{ground\ truth\ object\ volume}$$

positive: an extracted object with alarm value > decision threshold

true positive: a positive for which precision and recall > prescribed threshold

false positive: a positive for which precision or recall $\leq$ prescribed threshold

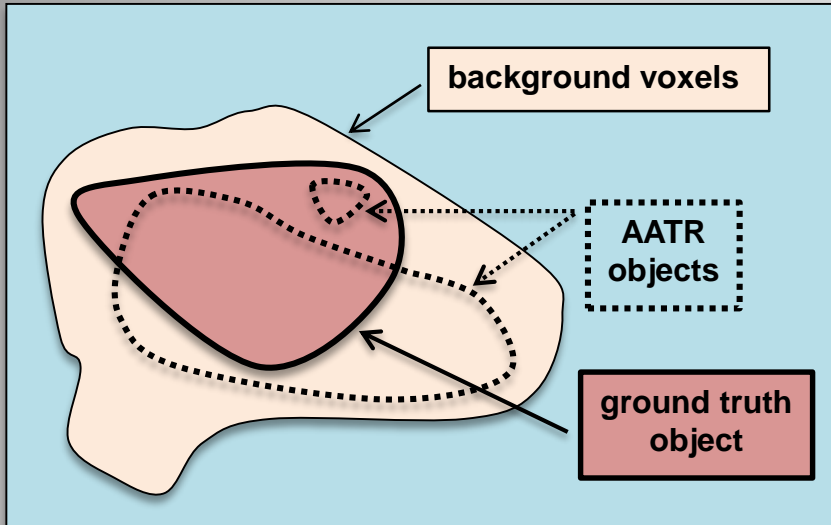*Thresholds must be prescribed for precision and recall.*

number of true positives

$P_D \; = \; N_{TP} / N^+$ N^+ = number of ground truth objects packed into bags

$P_{FA} \; = \; min \, (\, 1, N_{FP} / N^-)$ N^- = number of non-ground truth objects packed into bags

number of false positives

Proposed Voxel-Based AATR Scoring Metric (based on Image Similarity)

Background Voxels:

- Do not belong to ground truth objects.
- Have $\mu > \mu_{background}$

  (excludes air and low attenuating voxels)

Background Voxels: typically 5-10 % of total voxels
Ground Truth Voxels: typically <1% of total voxels
Air & Low Attenuating Voxels: typically > 90% of total voxels, excluded from the (P_D, P_{FA}) calculation

positive: a voxel <u>v</u> for which alarm value > decision threshold

true positive: a positive voxel <u>v</u> that belongs to a ground truth object

false positive: a positive voxel <u>v</u> that does not belong to a ground truth object

There are no thresholds on precision and recall.
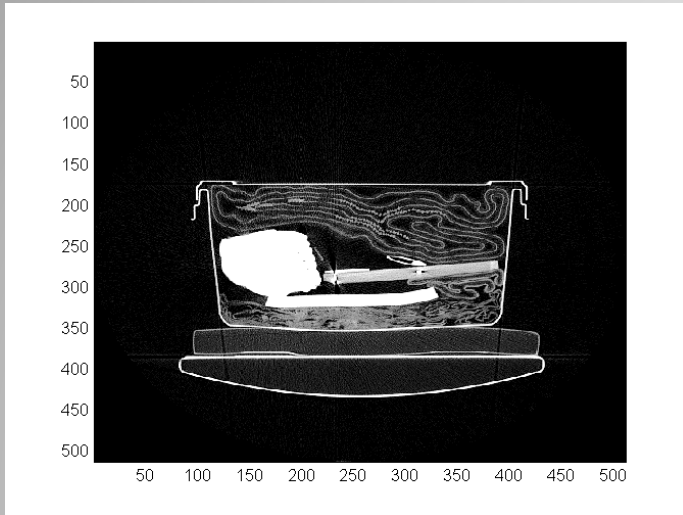
$P_D = N_{TP} / N^+$ N^+ = number of voxels that belong to a ground truth object
(P_D reflects fraction of ground truth volume alarmed on)

$P_{FA} = N_{FP} / N^-$ N^- = number of background voxels
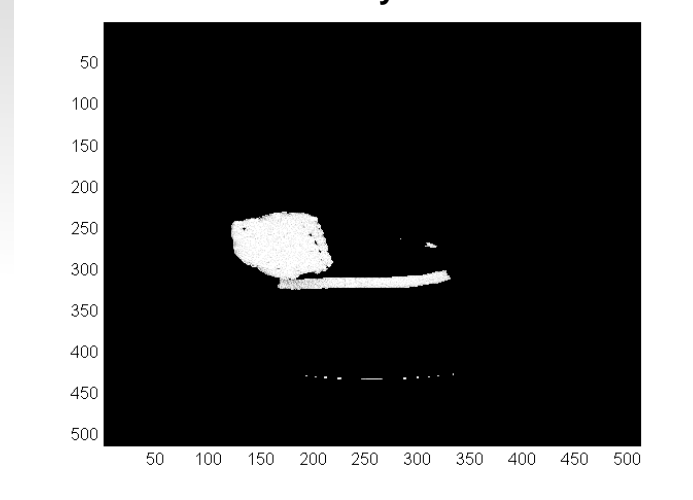(P_{FA} reflects fraction of background volume alarmed on)

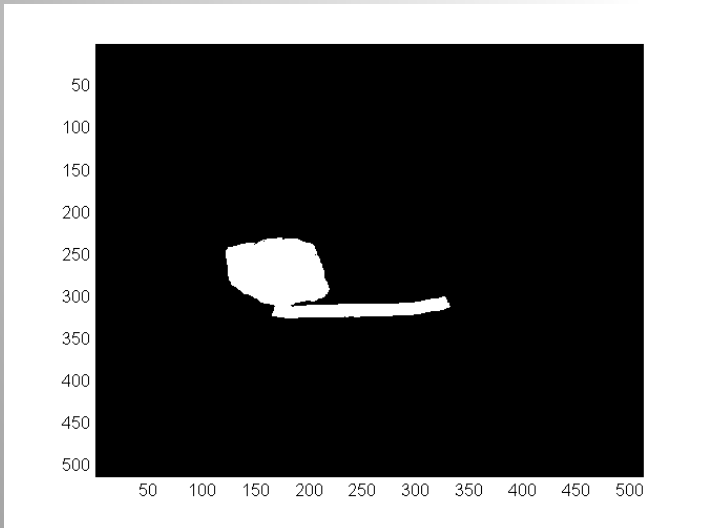# Examples of 2 Different Alarm Values for Voxels

TO4 image 127, slice 122 [4-5]

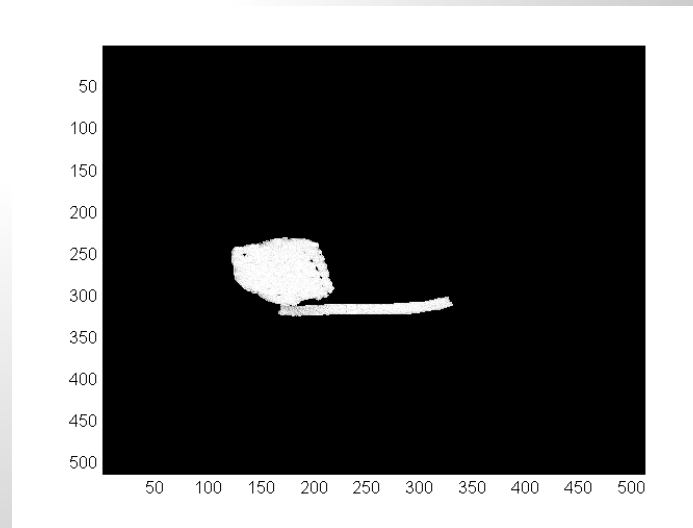<u>**Images with an Alarm Value Assigned to each Voxel**</u>

Voxel alarm values based solely on voxel material composition

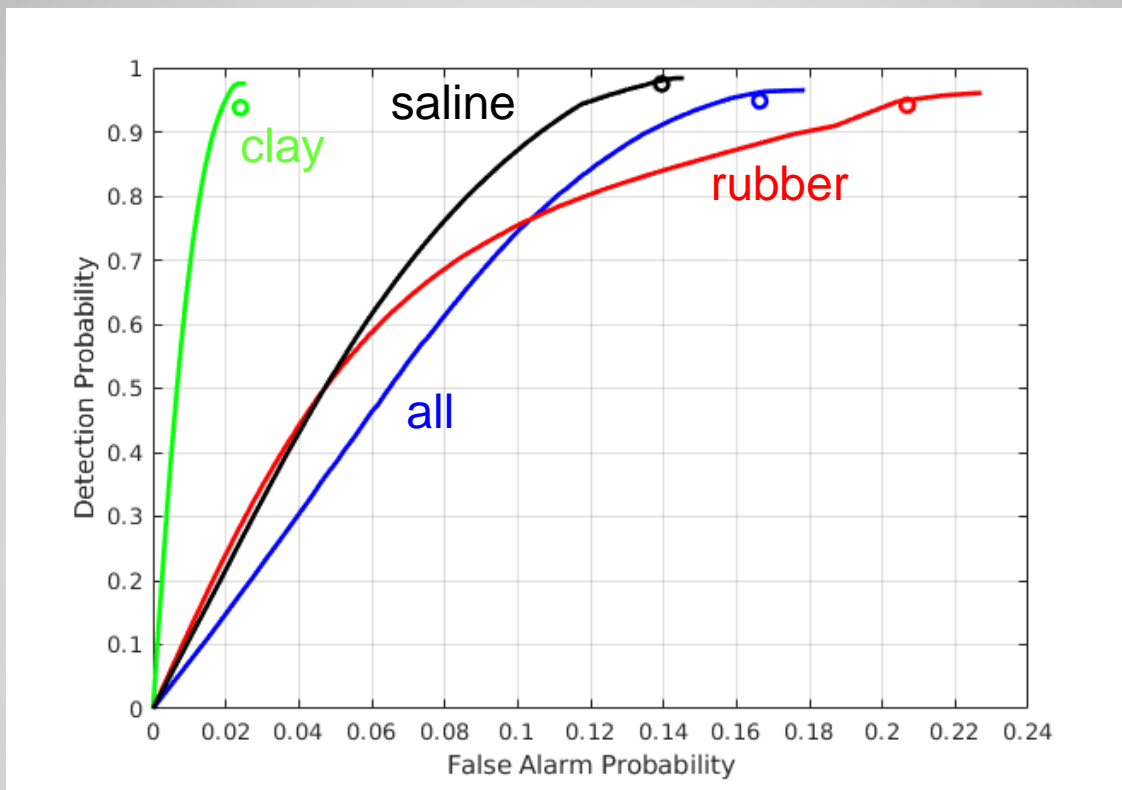**TO4 ground truth image 127, slice 122**

Voxel alarm values based on voxel material composition and mass / thickness of the object that contains the voxel

ROC Curves when Voxel Alarm Values are based Solely on Voxel Material Composition

Based on TO4 [4-5] images and ground truth

○ (P_D, P_{FA}) score at computed alarm threshold

Object Requirement Specification

material of interest
saline
clay
rubber

- Voxel alarm value varies from 0 to 1. ROC curves generated by varying the alarm threshold on voxel alarm values.

- (P_D, P_{FA}) score is derived from computed alarm thresholds, which vary between bags and materials-of-interest.

ROC Curves when Voxel Alarm Values are based on Voxel Material Composition and the Mass / Thickness of the Object that Contains It

Based on TO4 [4-5] images and ground truth

O (P_D, P_{FA}) score at computed alarm threshold

Object Requirement Specification

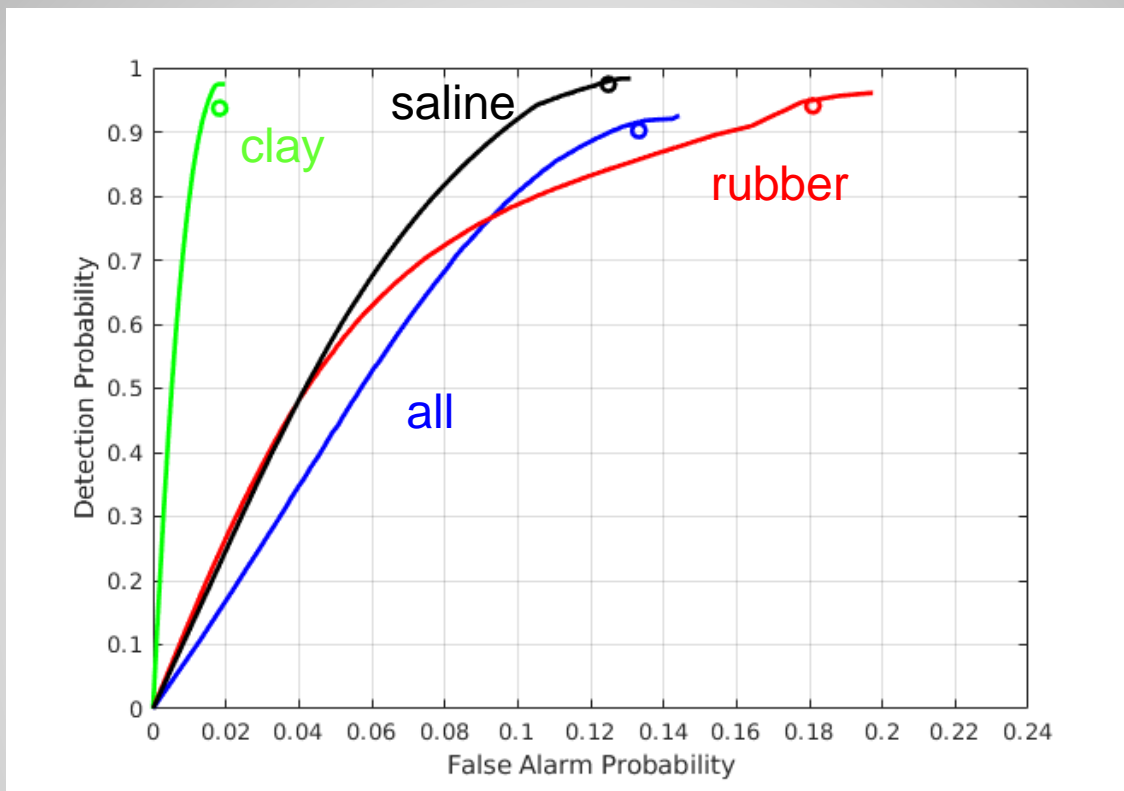| material of interest	min mass	min thickness
saline	50 g	0.7 cm
clay	50 g	0.7 cm
rubber	50 g	0.7 cm

- Voxel alarm value varies from 0 to 1. ROC curves generated by varying the alarm threshold on voxel alarm values.

- (P_D, P_{FA}) score is derived from computed alarm thresholds, which vary between bags and materials-of-interest.

Backup

Voxel-Based AATR Scoring Metrics
can use Voxel Weighting to Cope with Objects of Different Sizes

larger ground truth object
(potentially over-represented in the AATR score)

smaller ground truth object
(potentially under-represented in the AATR score)
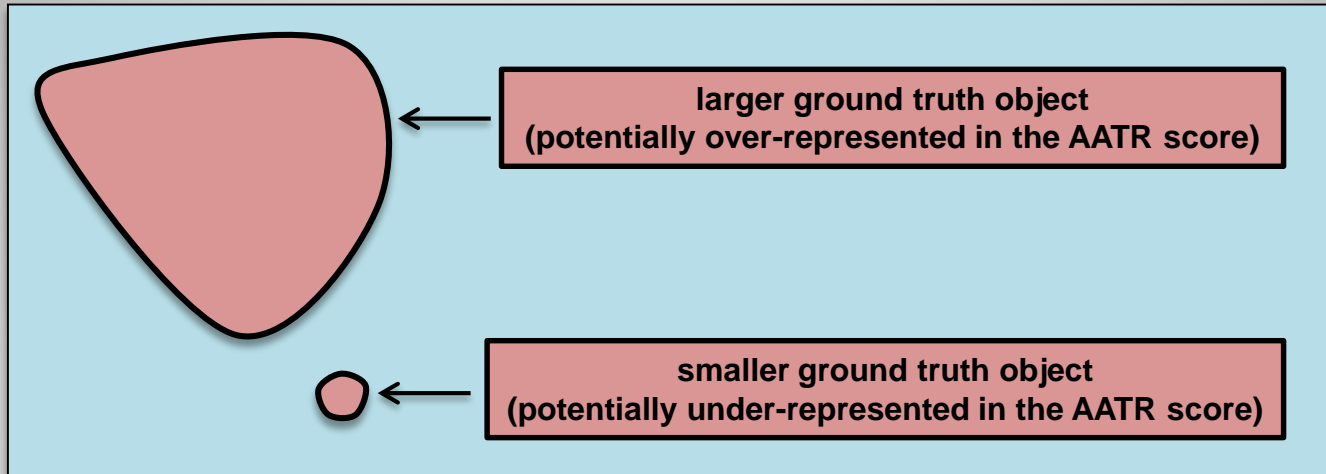
- Since small objects contain fewer voxels, they will be under-represented by voxel-based AATR scoring metrics.

- All objects could be equally represented if voxels in smaller ground truth objects were weighted more heavily.

- For equal representation, each ground truth voxel $\underline{v}$ is weighted by $1/N(\underline{v})$, where $N(\underline{v})$ is the number of voxels in the ground truth object that contains $\underline{v}$. In this case, voxel $\underline{v}$ is counted as $1/N(\underline{v})$ voxels.

Manually Generated Ground Truth needs to be Filtered for Accurate AATR Scoring

Ground Truth Filter Parameter	Remarks
Filter out the objects that do not contain materials of interest.	Only those ground truth objects that contain materials-of-interest can impact the AATR score.
Filter out the objects that do not have the desired physical features (e.g., mass and thickness within a specified range).	Only those ground truth objects whose physical features satisfy the stated constraints can impact the AATR score.
Filter out the object voxels whose values are inconsistent with object material composition.	Voxels in manually generated ground truth objects do not always correspond to the materials-of-interest. If not filtered out, the P_D values reported in the ROC curves will be too low. This type of filtering has potentially more impact on voxel based scoring results than on segment based scoring results.

Summary

- Scoring metrics are needed for comparing performance of different AATR algorithms.

- Two types of AATR scoring metrics were discussed:

  - Traditional metric <u>compares image segments</u>.

    — **Thresholds on overlap (precision & recall) between ground truth and AATR objects must be pre-specified.**

    — **These thresholds can impact the computed AATR score.**

  - Our proposed metric <u>compares image voxels</u>.

    — **P_D reflects the fraction of ground truth volume alarmed on.**

    — **P_{FA} reflects the fraction of background volume alarmed on (excluding air and low attenuating voxels).**

    — **Voxel based metrics do not use thresholds on precision & recall.**

    — **Voxel based metrics apply no matter how the alarm value of a voxel is defined, but the computed AATR score might change.**

    — **All objects can be equally represented by a voxel based metric if voxels in smaller ground truth objects are weighted more heavily.**

Footnotes

[1] **AATR for security emerged from several discussions amongst DHS S&T EXD, ALERT DHS Center of Excellence at Northeastern University, LLNL and others.**

[2] **https://myfiles.neu.edu/groups/ALERT/strategic_studies/ADSA15_final_report.pdf which contains the presentation and the PPT itself at https://myfiles.neu.edu/groups/ALERT/strategic_studies/ADSA15_Presentations/24_martz.pdf**

[3] **Awareness and Localization of Explosives-Related Threats http://www.northeastern.edu/alert/**

[4] **ALERT TO4 was funded by DHS EXD.**

[5] **TO4: https://myfiles.neu.edu/groups/ALERT/strategic_studies/TO4_FinalReport.pdf**