LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-754219

ALERT AATR Final Report: Lawrence Livermore National Laboratory

D. W. Paglieroni, H. Chandrasekaran, C. Pechard, H. E. Martz Jr.

July 3, 2018

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

ALERT AATR Final Report: Lawrence Livermore National Laboratory [1]

David Paglieroni, Hema Chandrasekaran, Christian Pechard and Harry Martz
Lawrence Livermore National Laboratory, P. O. Box 808, L-154, Livermore, CA, USA 94551

June 30, 2018

1. MOTIVATION FOR AATR AND BACKGROUND OF THE STUDY

This report addresses the problem of automatically recognizing objects of interest (OOIs) in x-ray computed tomography (CT) images of baggage (plastic bins in our case). The discussion draws heavily from the material in [1]. The material composition and basic physical features of interest (FOIs, such as mass and thickness) of the OOIs are defined in an object requirement specification (ORS). The ORS may also provide a detection – false alarm probability goal for each material of interest (MOI).

Automatic threat recognition (ATR) systems identify CT images of baggage that contain OOIs. ATRs must also identify locations of OOIs within baggage. The current ATR certification process used by DHS requires the ATR hardware and software to be certified on a specific ORS. When the ORS changes, the ATR hardware and software must be re-certified. To pass re-certification, the hardware, algorithms and codes may need to be modified. Because ATR-based certification and re-certification are so time-consuming, the current process cannot quickly adapt to changing requirements.

Unlike ATRs, adaptive automatic threat recognition (AATR) systems can quickly adapt to an ORS that changes or evolves over time [2]. The proposed AATR-based certification process would require the AATR hardware and software to be certified on a specific baseline ORS. Once certified, the same AATR hardware, algorithms and codes would be applied (by TSA or its delegates) to any ORS supplied in the future as input without going through lengthy re-certification. This proposed process would enable the AATR to quickly adapt to changing requirements.

This report describes the AATR developed at the Lawrence Livermore National Laboratory (LLNL) for x-ray CT images of baggage. The need and technical requirements for an AATR were developed in collaboration with DHS's Explosives Division and Northeastern University's Awareness and Localization of Explosives-Related Threats (ALERT) Center, a DHS Center of Excellence (http://www.northeastern.edu/alert/).

Approximately 180 CT images from task order 4 (TO4) were provided to facilitate AATR algorithm development [3]. The locations of voxels in the TO4 images that belong to OOIs were provided as "ground truth". While the MOIs in CT images from TO4 were limited to saline, rubber and clay [2], OOIs could in principle be defined for explosives, drugs or other contraband. The AATR was tested on CT images from task order 7 (TO7), and these images contained a variety of MOIs (not necessarily limited to saline, rubber and clay). Locations of the OOIs in the TO7 images were hidden from the AATR developers and known only to independent testers. While the MOIs are not called out by name, the ORS does supply a range of relevant x-ray attenuations for each MOI.

2. OVERVIEW OF THE LLNL AATR SYSTEM

The design of LLNL's AATR is guided by two basic philosophical points:

[1] LLNL Release number LLNL-TR-754219. This work was funded by the Science and Technology Directorate of the Department of Homeland Security (DHS). This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

1. It may not always make sense to train an AATR on the available CT images.

   Since the available CT images may not always adequately represent the newly prescribed ORS, it may not always make sense to train the AATR on those images. For example, there is no guarantee that the TO4 images will contain objects of desired mass and thickness composed of materials similar to the MOIs specified in a new ORS for TO7 images.

2. The goals of AATR can be achieved without image segmentation.

   One can conclude that a CT image contains OOIs and highlight their locations for an operator without extracting any segments. For example, one could highlight the centers of sliding window regions in 3D that contain enough relevant voxels (based, for example, on their x-ray attenuations). The set of all highlighted voxels would indicate the presence and highlighted localized occurrence of potential OOIs to an operator.

The first philosophical point led LLNL to develop an AATR that requires no classifier training, and thus no training images. Decision thresholds on voxel or segment relevance to the ORS must thus be estimated directly from the test images, and not by classifier training. Section 6 of [1] proposes a method that estimates a decision threshold for each MOI in each CT image based on probability density function (PDF) models for MOI attenuations and the PDF of CT image voxel attenuations. Because this method is still under development, the AATR performance results provided in Section 5 of this report were based instead on a single sub-optimal decision threshold estimated manually and applied universally to all images and MOIs.

The second philosophical point does not lead us to conclude that segmentation is useless. Image segmentation algorithms are important tools for many tasks in computer vision. However, it led us to focus less on sophisticated methods for image segmentation. Our AATR in fact uses a simple and computationally efficient segmentation method. However, as described in Section 5, the baseline (ALERT) performance measure assesses segmentation performance. By this measure, when the CT images are over segmented, the computed performance can suffer tremendously, even if the extracted segments, when considered collectively, cover the OOIs almost entirely. This led us to propose an alternative performance characterization measure (described in Section 6.1) based on matches, not between segments, but between binary ground truth images and images of relevance scores (from 0 for low relevance to the ORS to 1 for high relevance to the ORS). The proposed performance measure can cope with over and under segmented ground truth objects, and it produces results that are qualitatively consistent with intuition.

2.1 Top Level Description of the AATR

CT images of baggage contain voxels whose values represent x-ray attenuations at one or more energies (the COE data in our study is single energy and was acquired using an Imatron medical scanner). These attenuations provide indications of voxel material composition [4-6]. The material composition component of an ORS supplies a region of responsibility (ROR) for each MOI. An ROR can be represented by a range of values in linear attenuation coefficient or LAC (i.e., (μ_L,μ_H)) space, effective electron density – atomic number (i.e., (ρ_e,Z_e)) space, $(\mu_L/\mu_H, \mu_H)$ space, etc.

The joint PDF of LACs across one or more energies might also be supplied for certain MOIs. These joint PDFs might have been estimated as sample PDFs obtained from voxels across multiple bags associated with objects known to contain specific materials. How narrow these sample PDFs are is impacted by material homogeneity and/or the accuracies of values assigned to the voxels in tomographic image reconstructions. For example, the arrangement and material composition of objects in baggage can impact image reconstruction artifacts and the overall accuracy of the image reconstruction. Much work has been done over many years to improve the quality of CT image reconstructions (see, e.g., [7-9] and the references therein). This work is particularly relevant to us because CT image reconstructions are principal inputs to any AATR system. LLNL's AATR makes no attempt to address CT image artifacts. It is instead assumed that an attempt was already made to correct for artifacts in the CT images and that corrected CT images are supplied as input to the AATR.

The ability of an ATR to classify CT image voxels by material type is fundamentally limited by the widths, shapes and overlaps in PDFs associated with the various MOIs and benign material (particularly benign confuser material that can be mistaken for a MOI). Fig.1 shows sample and ROR-based PDFs for saline, rubber and clay in the single-energy case. Because the sample PDFs in Fig.1a for saline and rubber have significant overlap, one might expect saline and rubber to be difficult to discriminate. However, because the saline and clay PDFs overlap less, one might expect them to be easier to discriminate.

RORs contain less information than sample PDFs about the true shape of the PDF. For material mixtures, one might model the PDF associated with an ROR as uniform. For single materials, it might make more sense to model the PDF as unimodal with peak somewhere within the ROR (one naïve assumption being to assign the peak to the center of the ROR). Fig.1b shows ROR-based PDFs for saline, rubber and clay under this assumption. There is some basic resemblance between the ROR-based PDFs in Fig.1b and the sample PDFs in Fig.1a.

In a 3D image, an object manifests as an image volume composed of voxels. Such an object can be characterized not only by the material composition of the voxels it contains, but also by its physical features. By adding object physical features to the ORS, one might potentially be able to improve upon the performance of an ATR that finds OOIs (e.g., a rubber sheet or a saline bottle) based solely on voxel material composition (e.g., rubber or saline). For weapons such as guns and knives (which are not within the scope of the OOIs for this paper), specific object shape information is very important. Fuzzy K nearest neighbor (KNN) [10], support vector machine (SVM) [11] and convolutional neural network (CNN) classifiers [12] have all been applied to ATR of weapons in CT images of baggage. These classifiers are all trained on sets of positive and negative exemplars (CNN classifiers require the largest training sets). However, for OOIs such as explosives, drugs or other contraband (which *are* relevant to the scope of this paper), one would not want the object physical features to be too specific because the OOIs could have vastly different presentations (e.g., they could potentially come in all shapes and sizes). We thus currently limit the specification to prescribed ranges of very general physical features (in particular, mass and thickness) consistent with OOIs. Other features (namely texture and containment) have also been considered, but are not currently being used in our AATR implementation.

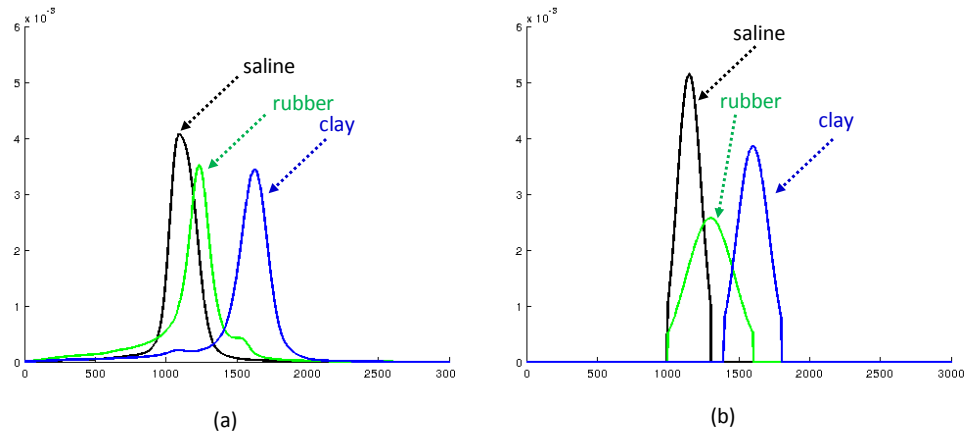


Fig.1 Examples of single-energy PDFs for saline, rubber and clay: (a) sample PDFs, (b) estimated ROR-based PDFs.

Fig.2 shows a top-level block diagram of LLNL's AATR. The most likely MOI composition is computed for each CT image voxel. Connected component segmentation is used to extract image volumes (i.e., segments or objects) from the material map (3D image of most likely MOI IDs). The voxels in an extracted object will all have the same most likely material composition. A score is computed for each voxel and for each extracted object. This score reflects degree of relevance to the ORS (it is a relevance score). Decision thresholds are estimated for each MOI within each bag (LLNL's algorithm for automatic decision threshold estimation is still under development). OOIs are identified by applying these decision thresholds to the extracted objects. The OOIs considered were limited to specific materials. However, OOIs could be defined for explosives, drugs or other contraband. Running on a single compute core in a desktop or laptop

computer with 24 GB of RAM, LLNL's AATR can process the single energy CT image of a typical plastic bin of size 512 x 512 x 400 voxels in ~15 – 30 seconds.

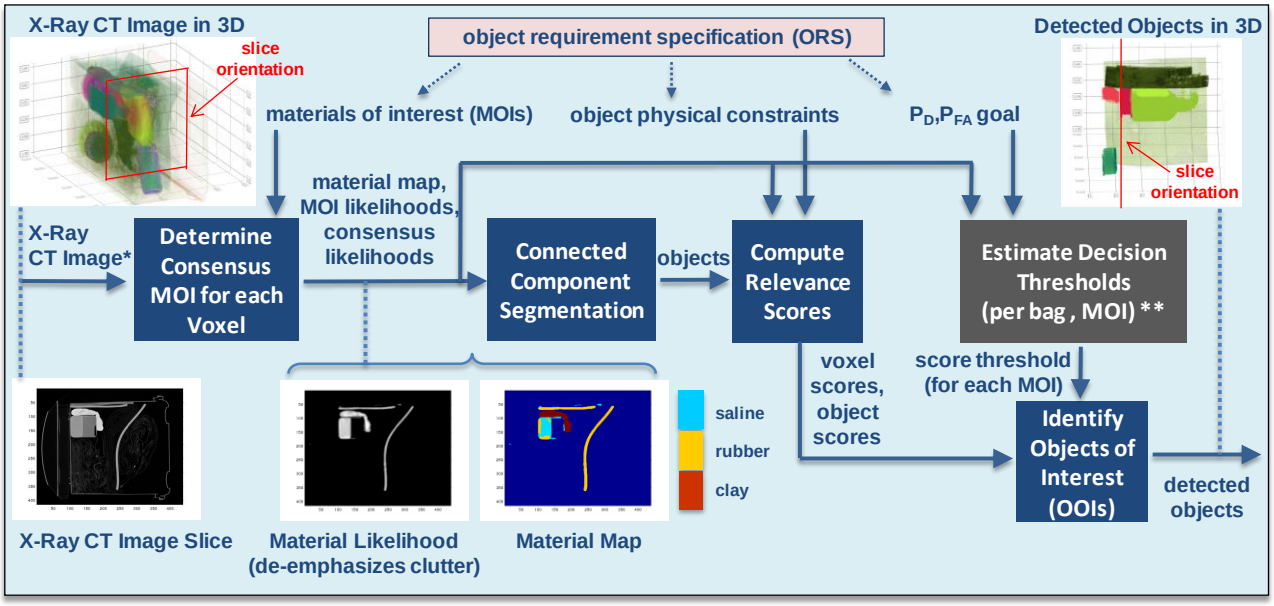


Fig.2 Top-level block diagram of LLNL's AATR.

3. CONSENSUS RELAXATION ON CT IMAGES

CT image cubes contain voxels $\underline{\mathbf{v}}(x,y,z)$ of linear attenuation coefficients (scalars for single energy case, two-tuples for dual energy case, etc.). Suppose the ORS accounts for m MOIs $\{M_k\}_{k=1}^{m}$. For each MOI, there is an associated likelihood function $p(\underline{\mathbf{v}} \mid M_k)$ (either a sample PDF or an estimated ROR-based PDF, as illustrated in Fig.1 for the single energy case). The mean likelihood of MOI k within the neighborhood of $\underline{\mathbf{v}}(x,y,z)$ is

$$\bar{p}(\underline{\mathbf{v}}(x,y,z) \mid M_k) = \underset{(x',y',z') \in N(x,y,z)}{\text{mean}} p(\underline{\mathbf{v}}(x',y',z') \mid M_k) \tag{3.1}$$

where $N(x,y,z)$ is the neighborhood of (x,y,z) – a rectangular window with center at (x,y,z) that extends $\pm w_x$ in x, $\pm w_y$ in y and $\pm w_z$ in z. For the test data described in this paper, we use $w_x = w_y = w_z = w$, where w is the consensus relaxation parameter. $\bar{p}(\underline{\mathbf{v}}(x,y,z) \mid M_k)$ amounts to a moving average of $p(\underline{\mathbf{v}}(x,y,z) \mid M_k)$ in 3D, which can be computed efficiently using a fast moving average algorithm whose time complexity does not depend on window extent. The consensus likelihood for voxel $\underline{\mathbf{v}}(x,y,z)$ is

$$p^*(\underline{\mathbf{v}}(x,y,z)) = \max_{k=1 \ldots m} \bar{p}(\underline{\mathbf{v}}(x,y,z) \mid M_k) \tag{3.2}$$

and the ID of the consensus MOI for voxel $\underline{\mathbf{v}}(x,y,z)$ is

$$k^*(\underline{\boldsymbol{v}}(x,y,z)) = \begin{cases} \arg\max_{k=1\ \ldots\ m} \bar{\mathrm{p}}\,(\underline{\boldsymbol{v}}(x,y,z) \mid M_k) & \mathrm{p}^*(\underline{\boldsymbol{v}}(x,y,z)) \geq \mathrm{p}^*_{\mathrm{crit}}\,(\underline{\boldsymbol{v}}(x,y,z)) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

$k^*(\underline{\boldsymbol{v}})$ in (3.3) varies from 0 to m. $k^*(\underline{\boldsymbol{v}}) = 0$ is reserved for background voxels (voxels believed to contain air or material that is of no interest). $k^*(\underline{\boldsymbol{v}}) > 0$ is reserved for foreground voxels (voxels believed to contain one of the materials of interest). In (3.3),

$$\mathrm{p}^*_{\mathrm{crit}}\,(\underline{\boldsymbol{v}}(x,y,z)) = \mathrm{p}_{\mathrm{crit}} \cdot \max_{\underline{\boldsymbol{v}}}\ \mathrm{p}(\underline{\boldsymbol{v}} \mid M_{k^*(\underline{\boldsymbol{v}}(x,y,z))}) \tag{3.4}$$

is a lower bound on the admissible consensus likelihood below which the ID of the consensus MOI is set to zero. We use $\mathrm{p}_{\mathrm{crit}} = 0.2$ as the value at the ROR boundary of ROR-based PDFs (see Fig.1b) normalized to a peak value of one. $\mathrm{p}_{\mathrm{crit}}$ separates voxels that potentially belong to OOIs in the bin or bag from background voxels that do not potentially belong to OOIs. The likelihood and MOI ID formulas in (3.2 - 3.3) are neighborhood operations (as opposed to point operations). They express consensus within a local neighborhood as to what the material composition of the voxel at the center is, and the degree of belief in that consensus. The consensus relaxation transformation from a CT image to MOI likelihood images, a consensus likelihood image and a material map is summarized in Fig.3. Consensus relaxation classifies each voxel by material type using a maximum likelihood classifier that requires MOI RORs or PDFs, but no training.

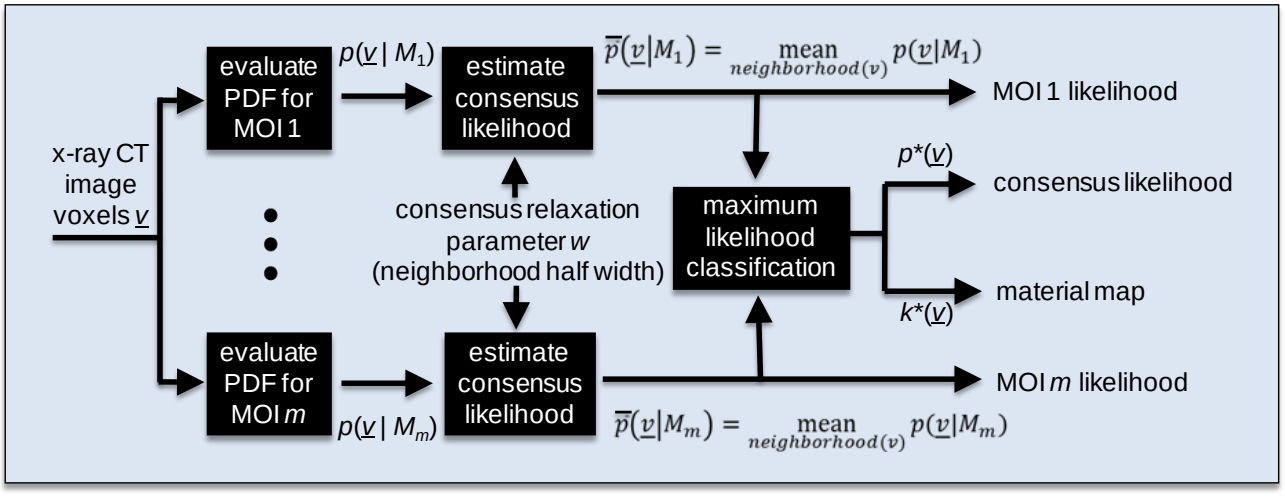


Fig.3 Consensus relaxation transformation from an x-ray CT image to MOI likelihood images, a consensus likelihood image and a material map.

Fig.4a shows the CT image of a plastic bin rendered in 3D using LLNL's AATR application. Fig.4b shows one slice (slice 160) of the CT image. Fig.4c-e show the consensus likelihood images of that slice for consensus relaxation parameters of $w = 1$, 2 and 3. Fig.4f-h show the corresponding material maps of that slice. The consensus image tends to become less fragmented and less busy as the degree of consensus relaxation increases. However, if the degree of consensus relaxation is too large, the consensus becomes more ambiguous (less localized), and the extracted objects begin to spatially distort.

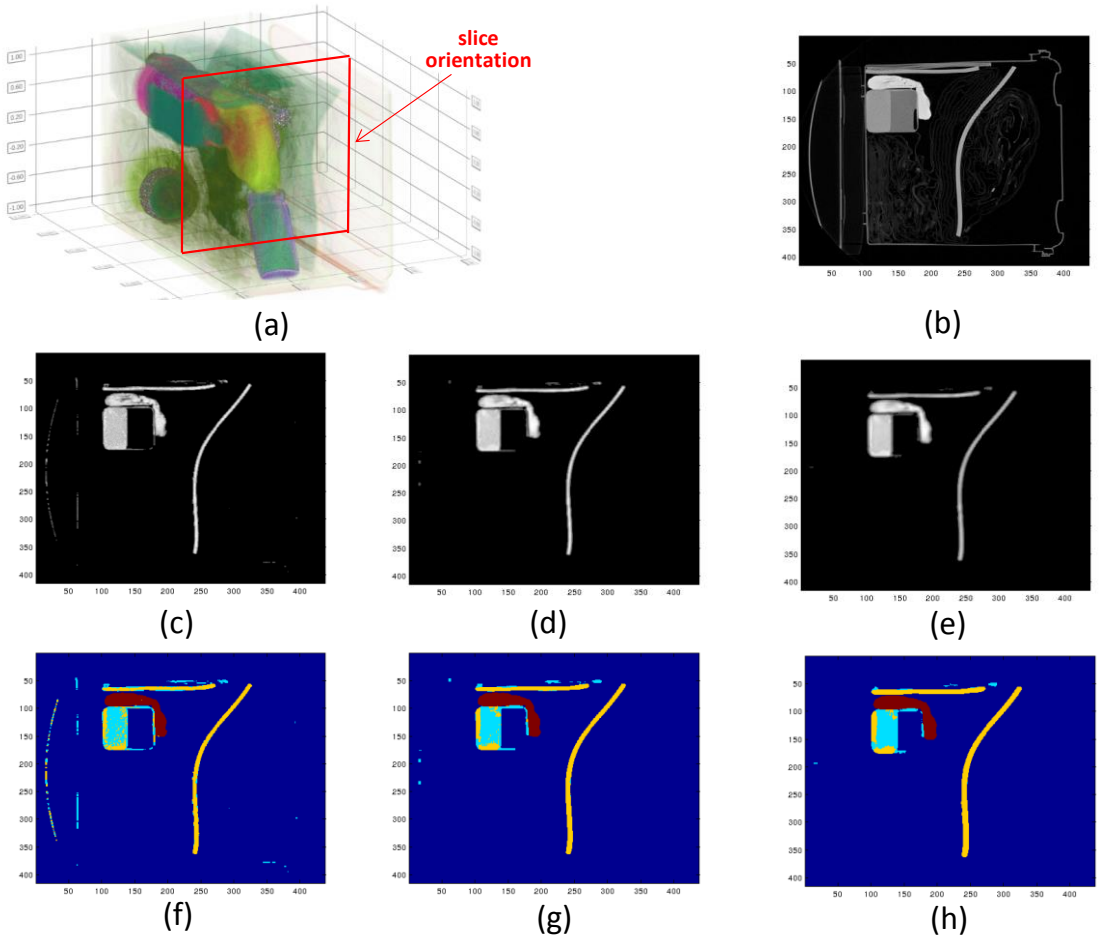


Fig.4 (a) CT image of a plastic bin with OOIs and clutter rendered in 3D, (b) CT image slice, (c-e) consensus likelihood images for w = 1, 2 and 3, respectively, (f-h) consensus MOI ID images for w = 1, 2 and 3, respectively (saline in light blue, rubber in orange, clay in red).

4. RELEVANCE SCORES FOR CT IMAGE VOXELS AND SEGMENTS

The idea of using segmentation to extract image volumes (segments) from CT images is consistent with the requirement to identify locations of potential OOIs within baggage. Image segmentation has been heavily researched over many decades. One general method is to spatially cluster voxels with similar properties. To determine the number of voxel categories present in the image, one might use variants of standard clustering methods such as K-means [13-14], or analyze the sample PDF of voxel values [15].

Our AATR extracts volumes (segments) from CT images by applying connected component segmentation to the material map in 3D. Each extracted image volume contains a complete set of spatially connected voxels for which the consensus MOI is the same, leading to extracted volumes of homogeneous material composition. No attempt is made to split or merge the extracted segments. Note that if particular mixed or heterogeneous material compositions are important, the mixtures themselves could be defined as MOIs.

The relevance score (classification statistic) that we compute for an extracted image volume (segment S) is the product of a segment material composition factor (which depends on the MOIs) and a segment physical features factor (which depends on the FOIs), both of which vary from zero to one:

$$c(S) = P_{\text{material}}(S)\, P_{\text{feature}}(S) \tag{4.1}$$

4.1 Segment Material Composition Factor

The material composition factor for extracted image volume or segment S is

$$P_{\text{material}}(S) = \frac{1}{n(S)} \sum_{(x,y,z) \in S} \tilde{p}^*(\underline{\boldsymbol{v}}(x,y,z)) \tag{4.2}$$

where $n(S)$ is the number of voxels in segment S. In (4.2),

$$\tilde{p}^*(\underline{\boldsymbol{v}}) \triangleq p^*(\underline{\boldsymbol{v}}) / \max_{\underline{\boldsymbol{v}}} p^*(\underline{\boldsymbol{v}}) \in [0,1] \tag{4.3}$$

where $p^*(\underline{\boldsymbol{v}})$ in (3.2) is the consensus likelihood for CT image voxel $\underline{\boldsymbol{v}}$. $\tilde{p}^*(\underline{\boldsymbol{v}})$ in (4.3) is the normalized consensus likelihood (some fraction of the peak consensus likelihood function value for CT image voxel $\underline{\boldsymbol{v}}$).

4.2 Segment Physical Features Factor

Regardless its material composition, the segment physical feature component for an extracted image volume restricts that volume to a low relevance score when any of its physical features is inconsistent with the ORS. For an ORS that includes n FOIs $\{f_i\}_{i=1}^{n}$, $P_{\text{feature}}(S)$ is thus expressed as the product of n factors, one for each FOI:

$$P_{\text{feature}}(S) = \prod_{i=1}^{n} P_{\text{feature},i}\,(f_i(S)\,;\,\underline{\boldsymbol{\theta}}_i(k(S))) \tag{4.4}$$

In (4.4), $P_{\text{feature},i}\,(f\,;\,\underline{\boldsymbol{\theta}})$ is the constraint function for physical feature i, $f_i(S)$ is the value of feature i for segment S, and $k(S)$ is the ID of the most likely MOI for segment S. Also, $\underline{\boldsymbol{\theta}}_i(k)$ is the set of parameters for the feature i constraint function when the most likely MOI for the segment has an ID of k. These parameters can vary not only from feature to feature, but also from MOI to MOI.

"Soft rectangular" constraint functions are sufficient for our purposes. As shown in Fig.5, these constraint functions are unit isosceles trapezoids, with three parameters, namely the start θ_0, end θ_1, and tail width θ_2 of the rectangular pulse (or alternatively, the center $(\theta_0+\theta_1)/2$, half width $(\theta_1-\theta_0)/2$, and tail width θ_2 of the rectangular pulse). If the tail width parameter is unspecified, it is set to the half width. "Soft rectangular" functions become rectangular pulses when $\theta_2 = 0$, triangular pulses when $\theta_0 = \theta_1$, and Kronecker deltas when $\theta_0 = \theta_1$ and $\theta_2 = 0$.

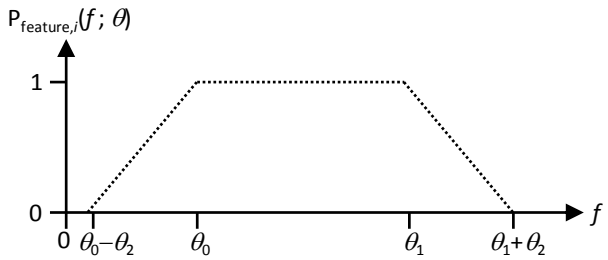


Fig.5 "Soft rectangular" constraint function for an object physical feature of interest.

4.3 Segment Physical Features

We define segment physical features as object macroscopic characteristics other than material composition. Because OOIs for explosives, drugs and other contraband can have vastly different shapes and sizes, the OOI physical features of interest are kept fairly general.

The physical features of interest used in LLNL's AATR are object mass and thickness, both of which relate to object shape or geometry. Two other features, namely texture (which is unrelated to object shape) and containment, have also been considered, but are not currently being used. Object mass is estimated by multiplying object volume by the mean of segment voxel attenuations and a mass density factor of 1/1024 g/cm^3 [2]. Object thickness is estimated in separate passes along the *x*, *y* and *z* axes over the 3D image of segment IDs. The mean of segment run lengths is computed along one direction in each pass. The minimum of over all three passes of the mean segment run length is used as the estimate of segment thickness. While this computationally efficient object thickness algorithm applies to 3D objects of any shape, the thickness estimate tends to be less accurate for thin objects (where uncertainty in the thickness estimate is on order of the actual thickness of the object).

4.4 Relevance Scores for CT Image Voxels

The normalized consensus likelihood

$$c(\underline{\boldsymbol{v}}) = \tilde{p}^*(\underline{\boldsymbol{v}}) \in [0,1] \tag{4.5}$$

in (4.3) could serve as a consensus-based relevance score or classification statistic $c(\underline{\boldsymbol{v}})$ for CT image voxel $\underline{\boldsymbol{v}}$. While this classification statistic does account for the material composition of CT image voxel $\underline{\boldsymbol{v}}$, it does not account for physical features of the object that CT image voxel $\underline{\boldsymbol{v}}$ belongs to (segment $S(\underline{\boldsymbol{v}})$). However, following the classification statistic for segment S in (4.1), the classification statistic

$$c(\underline{\boldsymbol{v}}) = \tilde{p}^*(\underline{\boldsymbol{v}})\, P_{feature}(S(\underline{\boldsymbol{v}})) \in [0,1] \tag{4.6}$$

for CT image voxel $\underline{\boldsymbol{v}}$ accounts for both.

5. AATR PERFORMANCE

The AATR performance metric used by ALERT [2] evaluates segmentation performance. It focuses on how accurately objects extracted from CT images by the segmenter match the emplaced OOIs (the "ground truth").

The AATR computes a relevance score $c(S)$ (see (4.1)) for each segment S that it extracts from the CT image. The number of positive objects N^+ is the number of OOIs packed into the test set of plastic bins. The number of negative objects N^- is the number of non-OOIs packed into the test set of plastic bins. At decision threshold c^* on segment

relevance score $c(S)$, the number of true positives $N_{TP}(c^*)$ is the number of segments extracted by the AATR for which $c(S) \geq c^*$ and for which precision and recall exceed some prescribed value (0.2 for sheet objects, and 0.5 for bulk objects [2] [2-3]). In the present context, precision is defined as the number of voxels in the extracted image segment that belong to the OOI (the number of voxels in the overlap) divided by the number of voxels in the extracted segment. Likewise, recall is defined as the number of voxels in the overlap divided by the number of voxels in the OOI. The number of false positives $N_{FP}(c^*)$ is the number segments extracted by the AATR for which $c(S) \geq c^*$ and for which the overlap criteria are not met. At a decision threshold c^*, the detection false alarm probabilities are then estimated as

$$P_D(c^*) = N_{TP}(c^*) / N^+, \qquad P_{FA}(c^*) = N_{FP}(c^*) / N^- \tag{5.1}$$

$P_D(c^*)$ in (5.1) will vary from zero to one. However, $P_{FA}(c^*)$ will be greater than one when the number of extracted image segments that are negatives exceeds the number of non-OOIs packed into the plastic bins. If $P_{FA}(c^*)$ exceeds one, it is clipped to one.

The AATR performance metric used by ALERT has the following properties:

(a) The P_{FA} ratio can produce values greater than one.

(b) The computed numbers of true and false positives can vary depending on the "special" threshold value used for precision and recall (a heuristic).

(c) Nearly identical image segments extracted by different AATRs could potentially contribute differently to P_D and P_{FA}. For example, for an image volume extracted by one AATR, a contribution will be made to P_D when the precision and recall relative to a specific OOI are slightly above the threshold. However, for a nearly identical volume extracted by another AATR, a contribution will instead be made to P_{FA} when either precision or recall relative to that same OOI are slightly below the threshold.

(d) An OOI will not be detected even when most of its voxels are covered by multiple extracted volumes that are each too small by themselves to be called detections. Each of these extracted volumes will instead be considered a false positive.

(e) An OOI will not be detected even when it is completely covered by an extracted volume that is too large to be called a detection. The extracted volume will instead be considered a false positive.

In order to classify volumes extracted by the AATR as positive vs. negative, the AATR must use a specific value for the decision threshold c^*. This value is currently specified manually and supplied as an input to LLNL's AATR. The performance of LLNL's AATR is highly dependent on this manually specified value. LLNL's algorithm for automatic decision threshold estimation is still under development.

LLNL's AATR performance is summarized in Fig. 6. LLNL's AATR code was supplied to the ALERT team. The team then independently conducted a series of performance studies on images of baggage from TO4 (to which the performers had access) and TO7 (which was sequestered).

For TO4 adaptability metrics (AMs) 1-3, the AATR P_D was often lower than the P_D goal, and the AATR P_{FA} was always higher than the P_{FA} goal. By manually specifying a lower decision threshold, LLNL's AATR could be made to

[2] Sheet objects are thin, and the ratio of thickness to surface area is very small. Bulk objects are either thicker, or the ratio of thickness to surface area is not small.

always meet the P_D goal, but the P_{FA} goal would never be met. This implication here is that the simple segmentation algorithm used in LLNL's AATR was often unable to extract objects from the TO4 images that matched closely enough to the ground truth objects, especially when the ground truth objects were small. When the segments extracted by LLNL's AATR did not match the ground truth segments well enough, P_D was lowered and P_{FA} was raised. This happened even when the segments extracted by LLNL's AATR collectively covered most of the ground truth segments.

For TO4 AM 4 (varying mass), the P_{FA} goal was always met or nearly met, but the P_D values were lower than for AM 1-3. This suggests that the simple segmentation algorithm used in LLNL's AATR was often unable to extract small objects from the CT images that match smaller ground truth objects well enough.

Performer Training / TO4 Data

AM 1: AROC

OOI	Required PD [%]	Required PFA [%]	AATR PD [%]	AATR PFA [%]
S	70	2	66	20
S	80	5	66	20
S	85	8	66	20
S	90	10	66	20
S	95	20	66	20

AROC	NOT DEFINED

AM 2: PD/PFA for Varying OOIs

OOI	Required PD [%]	Required PFA [%]	AATR PD [%]	AATR PFA [%]
C,S,R	90	10	82	24
C	90	10	95	24
S	90	10	72	24
R	90	10	82	24

AM 3: Varing PD Weight

OOI	Req PD [%]	Req PFA [%]	AATR PD [%]	AATR PFA [%]
C,S	C:90, S:90	10	C: 95, S: 66	26
C,S	C:20, S:90	10	C: 95, S: 66	26
C,S	C:90, S:20	10	C: 95, S: 66	26

AM 4: PD/PFA for Varying Mass

OOI	Min Mass [g]	Required PD [%]	Required PFA [%]	AATR PD [%]	AATR PFA [%]	Incremental Mass Range [g]	AATR Incremental PD [%]
S	400	90	10	53	11	N/A	N/A
S	300	90	10	22	6	300 - 400	33
S	100	90	10	50	16	100 - 300	53

AM 5: PD/PFA for Varying Thickness

OOI	Min Thickness [mm]	Required PD [%]	Required PFA [%]	AATR PD [%]	AATR PFA [%]	Incremental Thickness Range [mm]	AATR Incremental PD [%]
R	10	90	10	68	16	N/A	N/A
R	6.5	90	10	18	2	6.5 - 10	30
R	0	90	10	5	1	0 – 6.5	12

ALERT Testing / TO7 Data

AM 2: PD/PFA for Varying OOIs

OOI(s)	Required PD [%]	Required PFA [%]	AATR PD [%]	AATR PFA [%]
m1	90	10	94	11
m2	90	10	86	4
m3	90	10	85	2
m4	90	10	80	1

Fig.6 LLNL AATR performance as measured by the ALERT team. The tables surrounded by blue background reflect performance on TO4 data (to which performers had access). The table surround by pink background reflects performance on TO7 data of baggage (which was sequestered). The OOIs were S for saline, R for rubber and C for clay.

For TO4 AM 5 (varying thickness), the P_{FA} goal was always met or nearly met, but the P_D goal was never met. For objects of thickness from 0-6.5mm (i.e. of less than 6.5mm) or 6.5-10mm, one would get better results by manually specifying a smaller decision threshold as input to LLNL's AATR. However, since the CT image voxels have

dimensions that vary from roughly 0.9 to 1.5mm, AM 5 is attempting to detect objects with thickness of less than roughly 5 voxels and from roughly 5-7 voxels respectively. The uncertainty in our estimate of object thickness may not allow LLNL's AATR to reliably detect objects which are that thin.

Since LLNL's AATR does not train on TO4 data, one would expect a different AATR that was trained on TO4 data to perform better on TO4 data. However, for a sequestered test set of baggage (such as TO7) and a different ORS, LLNL's system cannot be over-trained and may adapt well. As shown in the ALERT testing table on previously unseen baggage (TO7), LLNL's AATR nearly met the (P_D, P_{FA}) goals for each of four types of unknown material. In one case, P_D exceeded the goal and in three cases, P_{FA} exceeded the goal. It should be noted that the decision threshold used was manually chosen, was not adaptive (did not vary from image to image or MOI to MOI), and was suboptimal. The results recorded in Fig. 6 were based on a second manual choice of decision threshold, which improved results somewhat relative to the first choice (but the same AATR code was used in each case). The need for manual decision threshold estimation will be eliminated once the adaptive automatic decision threshold estimation algorithm in [1] is fully integrated with LLNL's AATR. We hope automatic adaptive decision threshold estimation will improve performance beyond what is recorded in Fig. 6.

6. DISCUSSION AND FUTURE DIRECTIONS

6.1 Performance Metrics

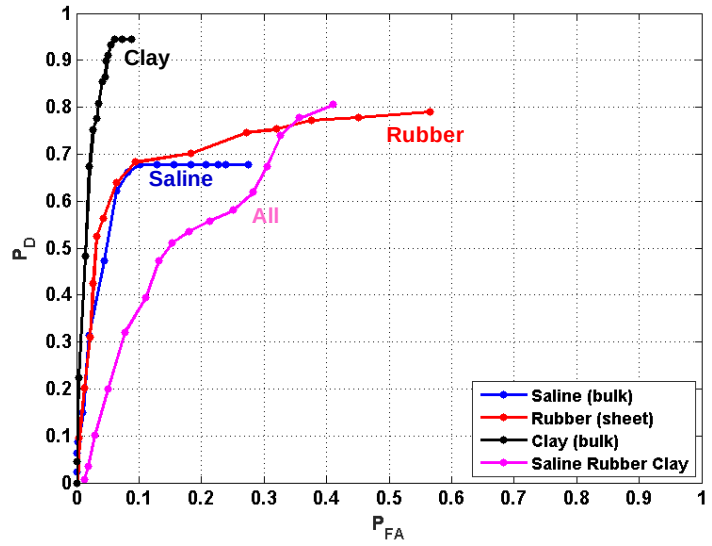
Some might view certain properties (a)-(e) of the ALERT performance metric in Section 5 as drawbacks. The issues suggested by those properties can be addressed by adopting a performance metric based on similarity between consensus relaxation images and the ground truth image as a whole (rather than on similarity between specific ground truth objects and extracted objects).

One such performance metric was proposed in [1]. $N_{TP}(c^*)$ is the number of voxels across all CT images for which $c(\underline{\boldsymbol{v}}) \geq c^*$ that belong to an OOI (where $c(\underline{\boldsymbol{v}})$ is given by (4.5)). Similarly, $N_{FP}(c^*)$ is the number of voxels across all CT images for which $c(\underline{\boldsymbol{v}}) \geq c^*$ that do not belong to an OOI. N^+ is the number of voxels across all CT images that belong to an OOI. N^- is the number of voxels across all CT images that do not belong to an OOI for which the mean of voxel attenuations across all energies is $\geq \mu_{\min}$. $\mu_{\min}$ represents a lower bound on attenuations for any potential material of interest. We use $\mu_{\min} = 300$ (for comparison, $\mu = 0$ for air; $\mu = 1000$ for water). $\mu_{\min} = 300$ excludes CT image voxels that occupy air and lightweight objects (such as clothing) from consideration when calculating N^-. $P_D(c^*)$ and $P_{FA}(c^*)$ are computed using (4.1).

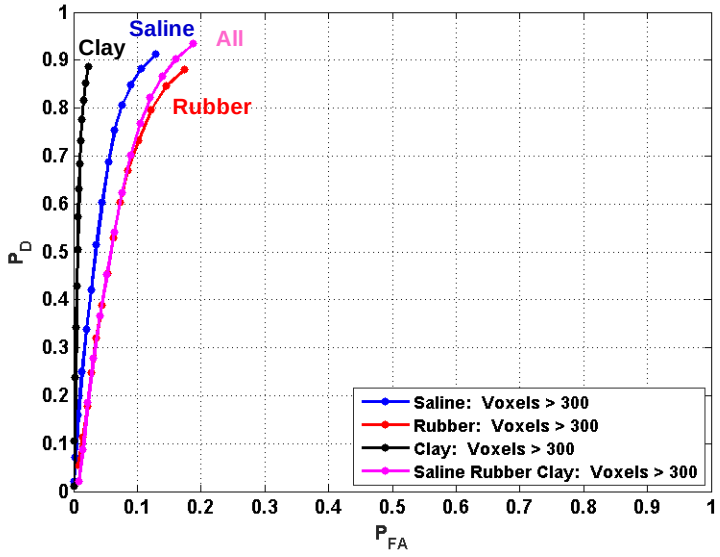
Using this performance metric, P_D and P_{FA} are easy to interpret. Specifically, P_D reflects the fraction of OOI volume alarmed on. P_{FA} reflects the fraction of non-OOI volume alarmed on (excluding volume composed of material for which the attenuation is less than $\mu_{\min}$). While $\mu_{\min}$ does impact the computed number of false positives, it is not a heuristic. In particular, $\mu_{\min}$ is based on RORs, theoretical values, or CT scanner measurements of MOI samples in the laboratory.

Fig.7 shows TO4 ROC curves (taken from [1]) for a specific ORS based on the ALERT performance metric and the proposed performance metric described above. The ROC curves in Fig.7a will vary depending on the heuristic degree of overlap specified as allowable for object matches. The ROC curves in Fig.7b are not subject to heuristics. One can see that these sets of ROC curves look very different. The AATR judged to have the best performance could thus conceivably change depending on which performance metric is chosen. Choice of performance metric is clearly very important. Even though the ROC curve appears to improve when the proposed metric is used, one must still understand

what "good" false alarm performance means in the context of a given performance metric. The bar for "good" performance might be higher for one performance metric than for another.



(a)



(b)

Fig.7 TO4 ROC curves for a specific ORS based on (a) the ALERT performance metric and (b) the proposed performance metric.

6.2 Adaptive Decision Threshold Estimation

While the goals of performance characterization using ROC curves can be achieved by varying the decision threshold c^*, a specific decision threshold must be chosen in order for the AATR to return concrete OOI assessment results to the operator. A key weakness of LLNL's AATR is that this decision threshold must be supplied as an input. However, LLNL is developing a method for estimating c^* that can adapt to different MOIs and to the clutter in different bags. For

a given CT image, the general idea is to estimate ROC curves for each MOI based on the values of voxels in the consensus relaxation images and their associated PDFs. To that end, decision threshold estimation is facilitated by using PDFs derived from RORs (as in Fig.1b) that vary continuously from 0 to 1 (and which would extend somewhat beyond the ROR support interval). If a (P_D, P_{FA}) goal is specified for a given MOI, the point on the estimated ROC curve for that MOI which best meets or exceeds that goal is used as the basis for the decision threshold. If no (P_D, P_{FA}) goal is specified, the point on the ROC curve is used that maximizes an objective function whose value increases (i) as P_D increases (for fixed P_{FA}), and (ii) as P_{FA}decreases (for fixed P_D). By allowing AATR decision thresholds to adapt to different MOIs and to the clutter in different bags, a level of performance that exceeds what the traditional ROC curves predict may potentially be achieved.

6.3 Remarks on Additional Aspects of the Project

It is LLNL's opinion that ALERT should consider adopting an AATR performance metric that does not rely on a heuristic degree of overlap between pairs of objects. Two alternative performance metrics were proposed in [1], and some discussion was provided in Section 6.1.

The idea of using an ALERT team to test AATRs supplied by the various performers on sequestered data is good, as it limits the ability of the various performers to adapt to the test. The process used in generating the data sets seems reasonable. However, architecting our codes so as to deliver a virtual machine (VM) to the ALERT team for use in independent testing took longer than anticipated, and was done at the expense of other key project deliverables (notably, adaptive decision threshold estimation, lack of which could be limiting our P_D-P_{FA}performance). Also, since we did not have access to sequestered data, it was harder for us to troubleshoot runtime failures on said data (these were typically due to mishandled "corner" cases).

LLNL appreciates being invited to participate in this effort. We learned a great deal about the AATR problem. Some key lessons learned are listed below:

- ATR and AATR are very different problems. While ATR may benefit from classifier training (because the ORS does not change), classifier training can actually be harmful in the context of AATR (because over-training in the context of one ORS can actually lead to reduced performance in the context of a different ORS). Classifier training should occur only in an ORS-agnostic context.

- AATR is not all about CT image segmentation. Computer vision specialists instinctively think of AATR as an object identification problem that requires segmentation to extract objects of interest from CT images. In reality, one can conclude that a CT image contains OOIs and highlight their locations for an operator without extracting any segments.

- The choice of AATR performance metric is important. This is easy to understand when one considers that choice of performance metric can have a profound impact on the ROC curve and which AATRs are judged to be superior.

- Strong similarities between MOI and non-MOI PDFs (of attenuations) pose a major problem for any AATR. The community must continue to explore methodologies (e.g., (ρ,Z) characterizations derived from dual or multi-energy scanners) and feasible sensor modalities that lead to signatures useful for material discrimination. Linear attenuation coefficients from dual energy systems will often be insufficient.

- Sequestered data and training: Even though the TO7 data was sequestered and could not be used for training, enough information was provided to the performers (in the form of non-sequestered data and performance

feedback) to enable them to adjust the algorithms to improve their performance and "pass the test". It is not clear that this was the intent of the project.

6.4 Summary

The key summary points for this report are as follows:

1. LLNL's AATR requires no training. The choice to avoid training is supported by realizing that it may not always make sense to train an AATR for a new ORS on the available CT images. Without training, LLNL's AATR has demonstrated (P_D, P_{FA}) performance close to the goal on sets of sequestered CT images.

2. A key weakness of LLNL's AATR is that the decision thresholds on relevance scores must be supplied as input. In lieu of using a trained classifier, an adaptive decision threshold estimation method is being developed to address this weakness. Since it produces decision thresholds that vary from image to image and from MOI to MOI (i.e., potentially from one extracted volume to another), it has the potential to perform at a level above what traditional ROC curves would predict.

3. LLNL's AATR is not based on sophisticated methods for image segmentation. However, we have observed that when an AATR is evaluated based on segmentation performance, the computed level of performance suffers tremendously when the CT images are over segmented – even if the extracted segments, when considered collectively, cover the OOIs almost entirely. We have thus developed an alternative performance characterization measure based on matches, not between segments, but between binary ground truth images and images of relevance scores (from 0 to 1). The proposed performance measure can cope with over and under segmented ground truth objects, and it produces results that are qualitatively consistent with intuition.

REFERENCES

[1] D. Paglieroni, H. Chandrasekaran, C. Pechard and H. Martz Jr., "Consensus Relaxation on Materials of Interest for Adaptive ATR in CT Images of Baggage", Proc. SPIE Defense and Security Symposium, Anomaly Detection and Imaging with X-Rays (ADIX) III, Orlando, FL, USA, April 17-18, 2018.

[2] Crawford, C., T07 Top Level Spec, V10 (URL: http://www.csuptwo.com/T07/T07 Top Level Spec V10.pdf)

[3] Crawford, C., T04 Final Report, (URL: https://myfiles.neu.edu/groups/ALERT/stretegic_studies/T04_FinalReport.pdf)

[4] Landry, G., Seco, J., Gaudreault, M., and Verhaegen, F., "Deriving effective atomic numbers from DECT based on a parameterization of the ratio of high and low linear attenuation coefficients," Phys. Med. Biol., 58, 6851–6866 (2013).

[5] Bond, K. C., Smith, J. A., Treuer, J. N., Azevedo, S., Kallman, J. S. and Martz, H. E., ZeCalc Algorithm Details, Version 6, LLNL Tech. Rep., LLNL-TR-609327, (Jan. 2013). To request a copy of ZeCalc software, contact Mary Holden-Sanchez at holdensanchez2@llnl.gov.

[6] Azevedo, S. G., Martz, H. E., Aufderheide, M. B., Brown, W. D., Champley, K. M., Kallman, J. S., Roberson, G. P., Schneberk, D., Seetho, I. M. and Smith, J. A., "System-independent characterization of materials using dual-energy computed tomography," IEEE Trans. Nuc. Sci., 63(1), 341-350 (2016).

[7] Stierstorfer, K., et al., Weighted FBP - a Simple Approximate 3D FBP Algorithm for Multislice Spiral CT with good Dose usage for Arbitrary Pitch., *Physics in Medicine and Biology*, 49, pp. 2209–2218, 2004.

[8] Champley, K. M., and Martz, H. E., "Statistical-analytic regularized reconstruction for x-ray CT," 12th Int'l Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, pp. 173-176 (2012).

[9] Karimi, S., Metal Artifact Reduction in Computed Tomography, Ph.D. Dissertation, UC San Diego: Electrical Engineering (Signal and Image Processing), Local identifier # b8199984, 2014.

[10] Mansoor, M. and Rajashankari, R., "Detection of Concealed Weapons in X-Ray Images using Fuzzy K-NN", *Int. Journal of Comp. Sci., Eng., and Inf. Tech.*, 2(2), 2012.

[11] Bastan, M., Yousefi, M. R., and Breuel, T. M., "Visual Words on Baggage X-Ray Images", in *Computer Analysis of Images and Patterns*, Springer, 2011, pp.360-368.
[12] Akcay, S., Kundegorski, M. E., Devereux, M. and Breckon, T. P., "Transfer Learning using Convolutional Neural Networks for Object Classification within X-Ray Baggage Security Imagery", *Proc. 2016 IEEE Int. Conf. Image Proc. (ICIP)*, September 25-28, 2016.
[13] Pappas, T., "An Adaptive Clustering Algorithm for Image Segmentation", IEEE Trans. Sig. Proc., 3, March 1994, pp.162-177.
[14] Achanta, R. et al, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods", IEEE Trans. PAMI, 34, 11, November 2012, pp.2274-228
[15] Shapiro, L. and Stockman, G., "Computer Vision", New Jersey, Prentice-Hall, 2001, pp.279-325.

AUTHOR BIOGRAPHIES

David Paglieroni

David Paglieroni received B.S., M.S. and Ph.D. degrees in Electrical and Computer Engineering from UC Davis, with an emphasis on signal / image processing and computer vision. He joined the Lawrence Livermore National Laboratory (LLNL) in 1999 to become Co-PI on the Image Content Engine (ICE) strategic initiative for broad area search in overhead images. He is currently a Science and Engineering Lead in the Computational Engineering Division at LLNL. He serves as Principal Investigator (PI) for adaptive automatic threat recognition (AATR) in x-ray CT images of baggage (funded by DHS), analytics lead for LLNL's ground penetrating radar (GPR) program area, and Co-PI on a tunnel risk assessment project (funded by CBP). Prior to joining LLNL, Dr. Paglieroni was manager of the imagery software section at the San Jose office of Lockheed Martin.

Dr. Paglieroni served as lead author on the majority of more than 60 papers and patents that he co-authored. He is a member of Tau Beta Pi, a Senior Member of the IEEE and was an associate editor for the Journal of Pattern Recognition for nearly two decades. His current research interests include signal / image processing, computer vision, graph-based data fusion, inter-visibility analysis in networks of moving objects, tunnel risk assessment along borders, buried threat detection in GPR images and AATR in x-ray CT images of baggage.

Hema Chandrasekaran

Hema Chandrasekaran received her B.Tech. degree in electronics engineering from Madras Institute of Technology, India in 1985 and M.S. and Ph.D. degrees in electrical engineering from the University of Texas at Arlington in 1994 and 2000 respectively. She has been a member of technical staff in the computational engineering division at Lawrence Livermore National Laboratory since 2010. Prior to joining LLNL, she was a senior scientific programmer for the Kepler Mission at NASA Ames Research Center at Moffet Field, CA.

Christian Pechard

Christian Pechard received his M.Sc. degree in computer engineering from ESIEE-Paris, France. From 1996 to 2007, he worked in the field of networking, telecommunication, real-time embedded systems, network management at Cisco Systems, San Jose, California. From 2007 to 2010, he worked in the field of enterprise wireless, real-time embedded system for Trapeze Networks, Pleasanton, California. Since 2010, he joined Lawrence Livermore National Laboratory (LLNL) to research and develop applications in the field of telecommunication, wireless protocols, real-time embedded systems. He is currently responsible for software pipeline development on a large multi-years ground penetrating radar program.

Harry Martz, Jr.

Harry Martz is the Director for Non-destructive Characterization Institute and a distinguished member of the technical staff at Lawrence Livermore National Laboratory. He is also Principal Investigator (PI) on Department of Homeland Security, Science and Technology, Explosive Division Projects and Domestic Nuclear Detection Office, Nuclear and Radiological Imaging Platform and Passive And X-ray Imaging Scanning projects. Harry originally joined the Laboratory to research and develop X-ray imaging and proton energy loss computed tomography for the non-destructive inspection of materials, components, and assemblies. He received his M.S. and Ph.D. in Nuclear Physics/Inorganic Chemistry from Florida State University, and his B.S. in Chemistry from Siena College.

Harry is leading a team of scientists and engineers to research, develop and apply nonintrusive characterization methods to better understand material properties and inspection of components and assemblies. He has applied CT to inspect one-millimeter sized laser targets, automobile and aircraft components, reactor-fuel tubes, new production reactor target particles, high explosives, explosive shaped charges, dinosaur eggs, concrete, and non-destructive radioactive assay of waste drum contents. Recent R&D efforts include CT imaging for conventional and homemade explosives detection in luggage and radiographic imaging of cargo to detect special nuclear materials and radiological dispersal devices. Dr. Martz has authored or co-authored over 300 papers and is co-author of a chapter on Radiology in Non-destructive Evaluation: Theory, Techniques and Applications, Image Data Analysis in Non-destructive Testing Handbook, third edition: Volume 4, Radiographic Testing, and contributed a chapter entitled Industrial Computed Tomographic Imaging to the Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar and Medical Imaging Real-Time Systems. In 2016 Dr. Martz and co-authors' published a text book entitled X-ray Imaging: Fundamentals, Industrial Techniques and Applications. He has also served on several National Academy of Sciences Committees on Aviation Security and was the Chair of the Committee on Airport Passenger Screening: Backscatter X-Ray Machines. Harry has been co-chair of the Awareness and Localization of Explosives-Related Threats, Advanced Development for Security Applications Workshops. Dr. Martz has presented a short course on CT imaging at The Center for Non-destructive Evaluation, Johns Hopkins University and a course on X-ray Imaging for UCLA's Extension Program. He is currently working with colleagues to develop a week-long course and lab on x-ray imaging for first responders. Awards include 2000 R&D 100 WIT-NDA (Waste Inspection Tomography for Nondestructive Assay), 1998 Director's Performance Award Active and Passive Computed Tomography and Federal Laboratory Consortium for Technology Transfer 1990 Award of Merit. He is a member of the Physics Honor Society Sigma Pi Sigma.