Passenger Screening Algorithm Challenge - Team 4

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

Description

While long lines and frantically shuffling luggage into plastic bins isn't a fun experience, airport security is a critical and necessary requirement for safe travel.

No one understands the need for both thorough security screenings and short wait times more than U.S. Transportation Security Administration (TSA). They're responsible for all U.S. airport security, screening more than two million passengers daily.

As part of their Apex Screening at Speed Program, DHS has identified high false alarm rates as creating significant bottlenecks at the airport checkpoints. Whenever TSA's sensors and algorithms predict a potential threat, TSA staff needs to engage in a secondary, manual screening process that slows everything down. And as the number of travelers increase every year and new threats develop, their prediction algorithms need to continually improve to meet the increased demand.

Currently, TSA purchases updated algorithms exclusively from the manufacturers of the scanning equipment used. These algorithms are proprietary, expensive, and often released in long cycles. In this competition, TSA is stepping outside their established procurement process and is challenging the broader data science community to help improve the accuracy of their threat prediction algorithms. Using a dataset of images collected on the latest generation of scanners, participants are challenged to identify the presence of simulated threats under a variety of object types, clothing types, and body types. Even a modest decrease in false alarms will help TSA significantly improve the passenger experience while maintaining high levels of security.

This is a two-stage competition. Please read our two-stage FAQs to understand more about what this means.

All persons contained in the dataset are volunteers who have agreed to have their images used for this competition. The images may contain sensitive content. We kindly request that you conduct yourself with professionalism, respect, and maturity when working with this data.

Data Description

This dataset contains a large number of body scans acquired by a new generation of millimeter wave scanner called the High Definition-Advanced Imaging Technology (HD-AIT) system. The competition task is to predict the probability that a given body zone (out of 17 total body zones) has a threat present.

The images in the dataset are designed to capture real scanning conditions. They are comprised of volunteers wearing different clothing types (from light summer clothes to heavy winter clothes), different body mass indices, different genders, different numbers of threats, and different types of threats. Due to restrictions on revealing the types of threats for which the TSA screens, the threats in the competition images are "inert" objects with varying material properties. These materials were carefully chosen to simulate real threats.

The volunteers used in the first and second stage of the competition will be different (i.e. your algorithm should generalize to unseen people). In addition, you should not make assumptions about the number, distribution, or location of threats in the second stage.

All volunteers have agreed to have their images used for this competition. The images may contain sensitive content. We kindly request that you conduct yourself with professionalism, respect, and maturity when working with this data.

Data size and access

The data for stage one is more than three terabytes in size. Stage two will have a similar size. Since most internet connections can not reliably download this much data, we are making the full dataset available on multi-regional Google Cloud Storage.

Two of the four file formats are downloadable from the competition page directly:

- Available to download from the competition files section: a full set of images in the .aps and .a3daps formats
- Available to download from or use with a virtual machine on Google Cloud: a full set of images in all four formats (.aps, .a3daps, .a3d, .ahi)

All four file formats represent the same underlying scan. They are simply different representations of a 3D image. Kaggle will provide example python code (as a Kernel) that shows how to read the images.

Access to the Google Cloud Storage bucket is controlled by membership in a Google group. Instructions to access are as follows:

1. Go to https://groups.google.com/forum/#!forum/kaggle-tsa-screening-challenge and request to join (make sure to say "Yes" to the question about the rules, or you will be rejected!)
2. Wait for approval. This may take up to a few business days. You will get an email once approved.
3. Access the bucket at https://storage.cloud.google.com/kaggle-tsa-stage1/

You can read more about accessing Google Cloud Storage buckets here. The large size and proprietary format of this dataset may seem daunting at first. We suggest you start with the smallest version of the images. Over time, the file formats will become familiar. Note that it is possible to work on--and potentially even win--this competition without the raw data files that make up the bulk of the dataset.

Image File Descriptions

The data for each scan performed by the HD-AIT system is referred to as an HD-AIT Frame. A frame consists of the following four binary files:

- _.ahi = calibrated object raw data file (2.26GB per file)
- _.aps = projected image angle sequence file (10.3MB per file)
- _.a3d = combined image 3D file (330MB per file)
- _.a3daps = combined image angle sequence file (41.2MB per file)

Each file is named with a scan Id. These file types are described in more detail below. Due to the dataset size, we have grouped the files into directories based on the file type. This will allow you to start with a small-but-complete set of image files (the .aps files) and work your way up to larger image files, as necessary.

The four files generated by the HD-AIT program set have a common file structure. All four files are binary and include a 512 byte header followed by the file's data. The header mostly contains technical scan parameters and is largely identical across all images. With the exception of the field 'data_scale_factor', we do not expect information from the header to be necessary or useful for the competition task. We have preserved the file formats used by the TSA in order to make the result of the competition more readily compatible with the images generated by their scanners.

After the binary header, the data is stored sequentially. When referring to the data storage order, the notation indicates, left to right, most significant (largest stride) to least significant (shortest stride) axes. For example, a data order defined as AYX means that the angular axis is the largest stride. Such a file would consist of a series of YX planes incremented in angle.

Projected Image Angle Sequence File (.aps)

The 'Projected Image' algorithm computes 3D images for 90-degree segments of data that are equally spaced around the region scanned. A maximum value projection of the result of each of these computations is written sequentially into a single file. The result of this is an image file that, when played back plane-by-plane, appears like the object is spinning on the screen.

```
Data file order: AYX (angle, vertical axis, horizontal axis)
Axis Name, Stride, Number of samples, Axis Length
XAxis, 1, Nx=512, Lx=1.0 meters
YAxis, 512, Ny=660, Ly=2.0955 meters
Angular, 337920, Na=16, La=360-degrees
```

The data type of this file is 16bit unsigned integer. Data scaling is achieved by multiplying each pixel value by the 'data_scale_factor' field in the header. In summary, the 'Projected Image Angle Sequence' file is a series of 2D mmWave snapshots equally spaced in angle around the object.

Combined Image 3D File (.a3d)

The 'Combined Image' algorithm computes eight 3D images that are equally spaced around the region scanned, and then combines these images into one composite 3D volumetric image. This computation is written into the .a3d file. When played back plane-by-plane, this image displays cross-section slices through the object at sequential heights.

```
Data file order: YZX (height, depth axis, horizontal axis)
Axis Name, Stride, Number of samples, Axis Length
XAxis, 1, Nx=512, Lx=1.0 meters
ZAxis, 512, Nz=512, Lz=1.0 meters
YAxis, 262144, Ny=660, Ly=2.0955 meters
```

The data type of this file is 16bit unsigned integer. Data scaling is achieved by multiplying each pixel value by the 'data_scale_factor' field in the header. In summary, the ‘Combined Image 3D’ file is the full composite 3D image volume stored in sequential height slices order.

Combined Image Angle Sequence File (.a3daps)

For visualization purposes, an 'Angle Sequence File' is rendered from the previously described 'Combined Image 3D' data. A projection through the 3D data at sequential angular increments is written contiguously into a single file. Similar to the 'Projected Image Angle Sequence' file, the result of this rendering is an image file that, when played back plane-by-plane, appears like the object is spinning on the screen.

```
Data file order: AYX (angle, vertical axis, horizontal axis)
Axis Name, Stride, Number of samples, Axis Length
XAxis, 1, Nx=512, Lx=1.0 meters
YAxis, 512, Ny=660, Ly=2.0955 meters
```

```
Angular,337920,Na=64, La=360-degrees
```

The data type of this file is 16bit unsigned integer. Data scaling is achieved by multiplying each pixel value by the 'data_scale_factor' field in the header. In summary, the 'Combined Image Angle Sequence' file is a series of 2D mmWave images equally spaced in angle around the object.

Calibrated Object Raw Data File (.ahi)

The three other types of images (.aps, .a3d, .a3daps) are processed projections based on the raw data file. These projection types were created by engineers with expertise in signal proccessing and are a more readily useful representation of the images. For this reason, and due to its vast size, using the raw data is optional and potentially redudundant with the other data formats.

```
Data file order: FXY (frequency, horizontal angular, vertical)
Axis Name, Stride, Number of samples, Axis Length
YAxis, 2, Ny=660, Ly=2.0955 meters
XAxis, 1320, Nx=900, Lx=451.8 degrees
Frequency, 1080000, Nf=512, Freq=10-40 GHz
```

The raw data is captured as follows. A synthetic focusing algorithm performs a waveform calibration operation as the first step. The algorithm combines a set of pre-scan calibrations files to reduce the effect of common mode direct coupling signals and to align the phase component of each waveform captured from each of the array elements. The result of this operation is a calibrated version of the raw data in units of volts. The data is complex floating point words ordered in frequency planes.

Competition File Descriptions

Every scan in the dataset has a unique (hashed) Id. Every Id will have four associated image files, grouped into folders by their type.

- stage1_sample_submission.csv - shows the submission format for stage 1. You will use this file to determine which Ids to submit to the leaderboard during stage 1.
- stage1_labels.csv - gives the labels for the training images in stage 1.
- sample.tar.gz - a small file containing samples of the image formats.
- stage1_aps.tar.gz - the complete set of stage 1 .aps images.
- stage1_a3daps.tar.gz - the complete set of stage 1 .a3daps images.