JPG Extraction Tool

Last Updated: 27/06/22 Owner: Dave Gee

Dev: Sam Bodza

The jpg extraction tool is used to extract images from the processed DR image Dataset for a data request. The tool automates the sample selection for which is then manually validated to determine the amount of identifiable images. The tool pulls a random sample of 1400 images to be manually validated by both an internal and external examiner. The tool also reads and parses the EXIF data, metadata attached to the images, from the entire sample highlighting any images with EXIF in a format that would not be expected for manual validation.

Instructions on how to use this tool are available in the documentation folder on DHP-NAS5.

Sampling

The original DR image dataset contains several types of images, some of which are considered identifiable. These images can be broken down into three main types; retinal, anterior and ’junk’. Retinal images are images taken of the back of the eye and are considered to be non-identifiable and are the images that are involved in the data requests. Anterior images are images taken of the front of the eye, and currently are considered to be identifiable by UHB IG and as such should not be passed on during a data request. Junk images are images that are not either retinal or anterior, and generally consist of images that were accidentally taken; they may contain a wall or an image of a patient’s full face and as such could be identifiable and cannot be released to the requester.

Due to the volume of images (~7,000,000) a deep learning AI algorithm was developed to be able to classify the images. A sample of images must be taken for every data request to validate the ongoing performance of the algorithm, and this sample must be large enough to be able to accurately describe the performance of the algorithm while also being small enough to not be too time consuming when being manually validated. The images will be validated by both an internal and external party, with any images that have been misclassified, being reported back to be updated in the database.

The sample size of 1400 was derived statistically using the IEEE definition for signal detection, where a signal must be detected at more than 3 standard deviations from zero. By using both Poisson and Normal distributions, we are able to derive the value needed for the above definition of signal detection.

Given *R* as the number of identifying images, *R* has an approximate Poisson distribution with mean *R=Np* where *p* is the probability of identifying an image and *N* is the size of the sample. Given the supposed accuracy of the ML in validation as 99.5% our probability p can be defined as:

From this, we can derive that given our value of p:

Using the Poisson distribution in the form of:

And for our threshold of 3 standard deviations:

A more accurate figure using a normal distribution can also be solved for *N* where:

Solving by brute force we find that

As such a random sample with a minimum size of 1379 will be taken from the extract to be manually validated.

EXIF data processing

EXIF is short for Exchangeable Image File, a standardised format for storing interchange information in image files that use JPEG compression. Most new digital cameras store EXIF data which has the possibility to include identifiable information. Some of this EXIF data is useful as it contains information around the cameras used which has been requested multiple times.

During the processing of these DR images, a CLI tool known as ‘exiftool’ is used to first scrape the useful information form the images EXIF data, and once the useful data has been collected, remove all the EXIF data from the image. To validate that this has been completed successfully during pre-processing multiple checks are implemented during the extraction process.

Initially all EXIF data from all the images is scraped and each image is checked to confirm that the EXIF is in the desired format. Any images that fail this check are removed from the main extract into a separate bucket to initially be stripped of all their EXIF data again. The images are checked over once more and any images that have ‘stubborn’ EXIF data that will not be removed, are again scraped and a set of the tags are pushed into a text document to be manually checked. This document is examined by both an internal and an external examiner and any tags that could potentially contain identifiable information are scraped form the images and the value of the tags are passed forward to both examiners.