**Deciphering categories & context of handwritten/typed text in images from Illicit & Counterfeit Medicines (ICM) marketplaces (Project 4)**

**Grad Student Mentor: Timothy Burt (taburt@uh.edu)**

**COSC 4397 Computer Vision, Fa21**

Data README

This document explains the folder contents which Prof. Kakadiaris provided you. It contains all data needed to start the project covered in the proposal document**. If you have any questions, please carefully review the project proposal document to see if the answer is there before contacting the grad student mentor.** I’m here to give feedback and guide your project, but you will implement your unique solution for this project.

# Summary

* 1-images: Contains images of medicinal products with/without PII, scraped from grey web marketplaces.
* 2-ppocr-inference-results: Contains preliminary results output from the PP-OCR software using the PGNet algorithm. It was run on all images. Unfortunately, the original parameters used when calling the code are not available, but they should give all of you a head start on understanding what we’re trying to accomplish.
* 3-annotations.xlsx: Contains ground truth annotations for the images with PII in the dataset, reviewed for accuracy. They also provide a complete listing of the images and which of the categories mentioned later the images fall in.
* 4-annotations\_data\_dictionary.docx: Describes the columns in the annotations document.

Other data may be available regarding the cases the images belong to. If all of you get to a point where you need to know specifics about each case to improve your method (e.g., reducing the type of drugs shown in the images to three), you may discuss your needs with Tim. He will consider them & provide additional data when needed.

## 1-images

This folder contains 2,897 HQ images, totaling 215 MB.

All images have been converted to RGB color space with JPEG compression. Images with both dimensions below 100 pixels were removed automatically, as well as duplicate images. Irrelevant images were also removed manually.

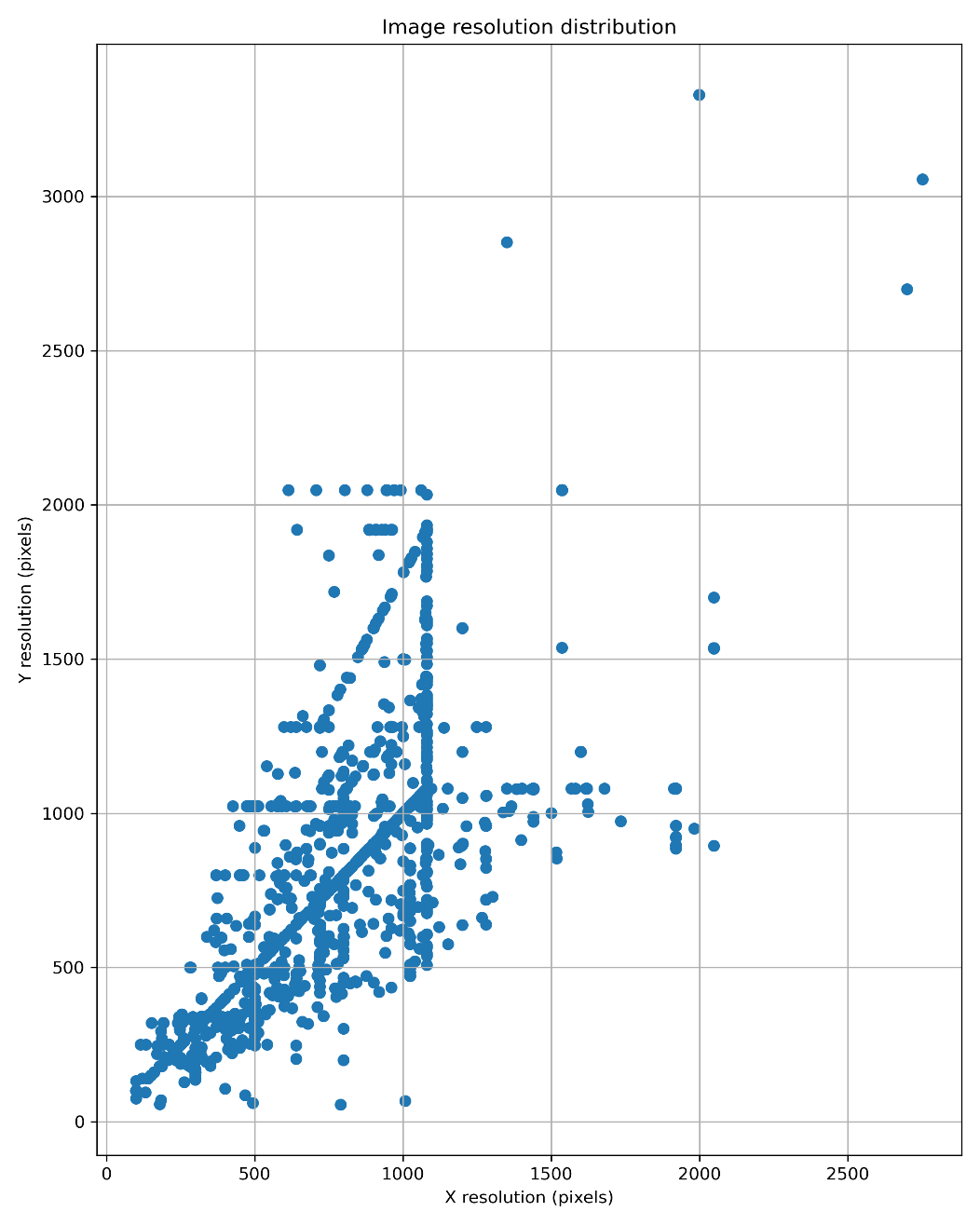
The images fall roughly into three categories, described next **(NOTE: they may have changed slightly, so don’t depend on any statistics described here. As researchers, the first task when starting a new project is to reproduce the SOTA algorithm/paper results. This is never an easy task.)**

* PII-likely-English (personally identifiable information)
  + 103 images, 12 MB, all contain English print or handwritten names, phone numbers, or emails.
  + Hand-sorted.
* PII-unknown (personally identifiable information)
  + 257 images, 34 MB, containing Chinese, Russian, Arabic, English printed text overlayed on an image, or handwritten English characters/numerals.
  + Many of these images are in Chinese print and may not contain names, phone numbers, or addresses. Chinese/Arabic numerals are not easy to spot since they each use their non-Latin numerals.
  + Most of the text output from these will be false alarms and have no PII. They will probably contain the drug, price, and quantity for sale. This method is prevalent on Instagram, Twitter, and Facebook.
  + If the image had a print that was possibly PII but couldn’t be understood, it belongs here.
* Medicines
  + 2,557 images, 176 MB. These contain the medicinal product, but NO possible PII or text is overlayed or present in the image.
  + These could have OCR run to determine the product with certainty instead of relying on correctly parsing the metadata available (if any).

## 2. ppocr-inference-results

These are .txt files that match the corresponding image from the dataset. They contain the output as discussed in the project proposal. They have not been reviewed, so we aren’t sure how good the quality is. They are divided into two folders:

* all\_images: results from all images in the dataset (2,908 files)
* pii\_images: results from only images with some PII in/on them (101 files)





There are also ground-truth annotations, and an accompanying README file is provided.