Large Scale OCR Extraction, Image Captioning and Object Recognition and Enrichment of UFO Sightings Data

Soravis Taekasem
taekasem@usc.edu

Surasit Prakunhungsit
prakunhu@usc.edu

Theerapat Chawannakul
tchawann@usc.edu

Yifan Wu
ywu352@usc.edu

Abstract

In this assignment, we explore and analyze some different form of UFO sightings data which are portable document format (.pdf) and image file format that is a rather more complex data type to handle. Specifically, extract, clean, analyze and visualize using several tools and framework including Image Magick, Poppler, Tesseract, Tika Parser, Tika Dockers, TikaAndVision, Tika ImageCaption, Selenium, TikaNER, OpenNLP, CoreNLP, NLTK, MITIE and Grobid Quantities. Finally, convert and aggregate it with a pre-process UFO sightings data to answer the question previously unanswered.

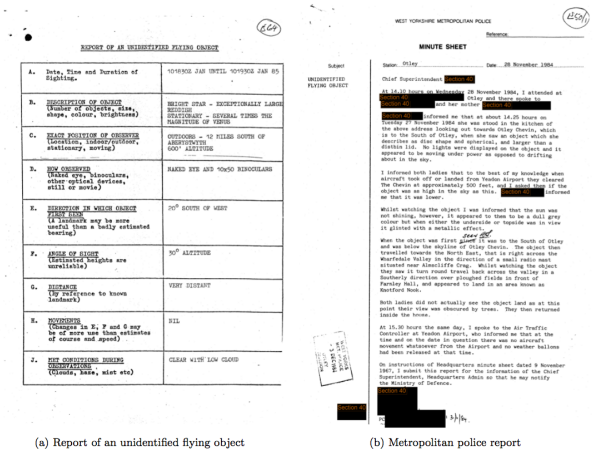
1 Introduction

Our ultimate goal is to discovering useful information, suggesting conclusions and supporting decision-making from UFO sightings dataset. Hence, we consider 2 new dataset which are British Ministry of Defence?s UFO Sightings data and UFO sightings images from the UFO stalker dataset to figure out how the UFO sightings are influenced by other factors.

Moreover, to analyze portable document format and image file we apply several advanced extractions technique such as Optical Character Recognition (OCR) with Apache Tika and Tesseract.

2 Datasets Observation

In this experiment we mainly focus on 2 datasets including The National Archives and UFO stalker.



REPORT OF AN UNIDENTIFIED FLYING OBJECT

A.	Date, Time and Duration of Sighting.	1058GMT JAN UNTIL 0759GMT JAN 85
B.	DESCRIPTION OF OBJECT (Number of objects, size, shape, colour, brightness)	BRIGHT STAR - EXCEPTIONALLY LARGE REDDISH STATIONARY - SEVERAL TIMES THE MAGNITUDE OF VENUS
C.	BEARING POSITION OF OBSERVER (Location, indoor/outdoor, stationary, moving)	OUTDOORS - 42 MILES SOUTH OF ABERDEEN/ON 600' ALTITUDE
D.	HOW OBSERVED (Naked eye, binoculars, other optical devices, still or movie)	NAKED EYE AND 50x50 BINOCULARS
E.	DIRECTION IN WHICH OBJECT FIRST SEEN (A landmark may be more useful than a hazily estimated bearing)	30° SOUTH OF WEST
F.	ANGLE OF SIGHT (Estimated heights are unreliable)	30° ALTITUDE
G.	DISTANCE (By reference to known landmark)	VERY DISTANT
H.	MOVEMENT (Changes in E, F and G may be of more use than estimates of course and speed)	NIL
J.	SKY CONDITIONS DURING OBSERVATION (Cloud, haze, mist etc)	CLEAR WITH LOW CLOUD

WEST YORKSHIRE METROPOLITAN POLICE

MINUTE SHEET

Subject: UNIDENTIFIED FLYING OBJECT

Status: Otley Date: 28 November 1984

Chief Superintendent

At 15.30 hours on Wednesday 28 November 1984, I attended at Section 40 Otley and there spoke to Section 40 and her mother Section 40

informed me that at about 14.25 hours on Tuesday 27 November 1984 she was stood in the kitchen of the above address looking out towards Otley Chevin, which is to the South of Otley, when she saw an object which she described as being circular in shape, dark in colour, and larger than a dustbin lid. No lights were displayed on the object and it appeared to be moving under power as opposed to drifting about in the sky.

I informed both ladies that to the best of my knowledge when aircraft took off or landed from Yeadon Airport they cleared The Chevin at approximately 500 feet, and I asked them if the object was as high in the sky as this. Section 40 informed me that it was lower.

Whilst watching the object I was informed that the sun was not shining, however, it appeared to them to be a dull grey colour but when either the underside or topside was in view it glinted with a metallic effect.

When the object was first seen it was to the South of Otley and was below the skyline of Otley Chevin. The object then travelled towards the North East, that is right across Wharfedale Valley in the direction of a small radio mast situated near Almscliffe Crag. Whilst watching the object they saw it turn round travel back across the valley in a Southerly direction over ploughed fields in front of Farnley Hall, and appeared to land in an area known as Hartford Nook.

Both ladies did not actually see the object land as at this point their view was obscured by trees. They then returned inside the house.

At 15.30 hours the same day, I spoke to the Air Traffic Controller at Yeadon Airport, who informed me that at the time and on the date in question there was no aircraft movement whatsoever from the Airport and no weather balloons had been released at that time.

On instructions of Headquarters minute sheet dated 9 November 1967, I submit this report for the information of the Chief Superintendent, Headquarters Admin so that he may notify the Ministry of Defence.

Section 40

(a) Report of an unidentified flying object

(b) Metropolitan police report

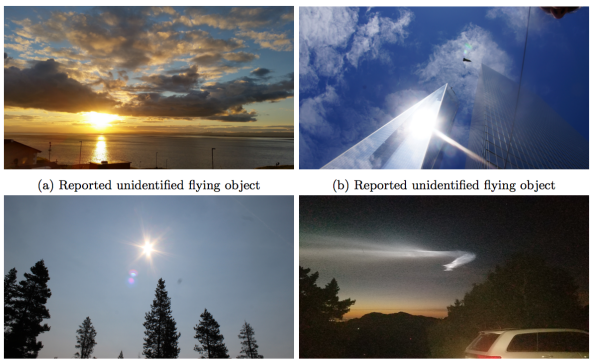
Figure 1: UFO sighting report

2.1 British UFO Sightings Dataset

British UFO sightings dataset from The National Archives is a portable document format dataset contain a wide range of UFO-related documents, drawings, letters, photos and parliamentary questions covering the final two years of the Ministry of Defence's UFO Desk (from late 2007 until November 2009). This data are in a form of 8 PDF files and a total of 1,968 pages. A sample of UFO sighting data from the British UFO files dataset is shown in Figure 1.

2.2 UFO Sightings Images Dataset

UFO sightings images dataset from UFO stalker is a images dataset which includes a collection of 4,065 sightings that have one or more images associated with them. A sample of UFO sighting data from UFO stalker dataset is shown in Figure 2.



(a) Reported unidentified flying object

(b) Reported unidentified flying object

(c) Reported unidentified flying object

(d) Reported unidentified flying object

Figure 2: UFO images submitted

3 Implementation

3.1 British UFO Sightings Dataset

For analyzing British UFO sighting datasets we implement 3 major component which are data retrieval, data cleaning and content extraction.

3.1.1 Data Retrieval

In this module we use PDF rendering library called Poppler to handle British UFO sightings datasets and extract 8 PDF files into 1,968 single page PDF file with `pdfseperate` command. After that we using ImageMagick which is a image processing and manipulation library to process each PDF

file and convert them into Tagged Image File Format (TIF) files by `convert` command.

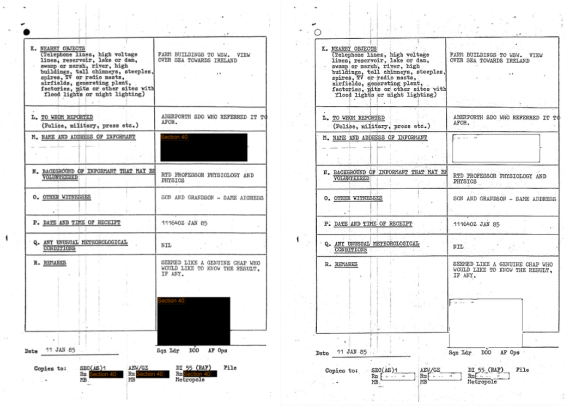
3.1.2 Data Cleaning

With the given pipeline, tesseract library return several errors for some input images, because the image is not suitable to apply optical character recognition (OCR). We use `white` parameter to convert the image into grayscale image and `-alpha off` parameter to turn off the alpha/matte channel of the image. This solved the problem.

In order to process image files more accurately, we try to clean the data using several image manipulation function from ImagesMagick such as `noise` and `median` command to reduce some noise from the image.

Also `contrast` and `threshold` command to increase the difference in luminance between bright and dark pixel. Next, render and force the pixel above the threshold into white and pixel below the threshold into black.

We also add ImageMagick plugin called Fred's ImageMagick script and use `textcleaner` function for further image cleaning. And by using `"-g -e none -f 15 -o 10"` parameter, we received the better image with less noise.



(a) Before processing (b) After processing

Figure 3: Image manipulation

3.1.3 Content Extraction

To extract the data we develop a Tika parser script that read the PDF files then extract some text content. Also by using Tesseract OCR from TikaOCR library, we extract text content from the given UFO sighting PDF files.

Moreover, we modify some JAVA script to clean both files from the Tika parser and the Tesseract OCR. The script will cleans bogus text and then add good ones to the TSV file.

3.2 UFO Sightings Images Dataset

For analyzing UFO sightings images dataset we implement 2 major components which are data retrieval, image captioning and objects recognition.

3.2.1 Data Retrieval

In this module we scraping images and retrieving sighting reports with `image_grabber.py` script. In order to get all images from ufostalker.com, our team created a python script to retrieve a JSON data contains UFO sighting information. Luckily, the JSON data contains image URLs for photo submitted by the reporter, so all images can be downloaded directly without any trouble. While it may sound easy, there was some obstacles during the process.

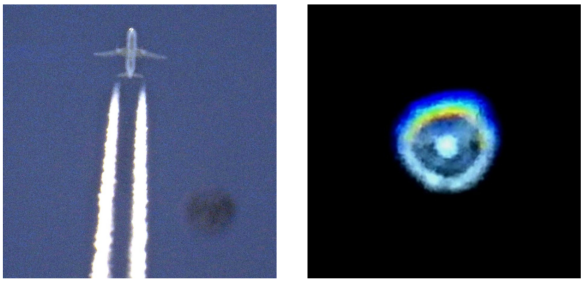
- There is a pagination in JSON data, so we have to use a for-loop to iterate through every page.
- ufostalker.com though that we are a DDoS attacker and blocked us from the site from accession because we forgot to add a delay in each loop.
- Some images are no longer there, for all sighting reports with id below than '69929', the uploaded images are not exist anymore. So those images we downloaded from the python script were unable to open and we had to remove them manually.

During the scraping process on images, we also stored sighting reports into `stalker_reports.tsv`

3.2.2 Image Captioning & Objects Recognition

After retrieved images from scraping, we perform image captioning process and object recognition process to determine best sentences as well as identify objects that describe an image. We used 2 scripts for these processes.

- `img2captions.py` to extract captions from image as `sentences.tsv`
- `img2objects.py` to identify objects in image as `objects.tsv`



(a) an airplane is flying in the sky with a space shuttle on it (b) a man flying through the air while riding a snowboard

Figure 4: Image captioning result

Figure 4 shows two sample image ran image captioning and object recognition. Figure 4a is a sample image with an apparent object, the result we got after running image captioning is "an airplane is flying in the sky with a space shuttle on it". In Figure 4b is a sample image with ambiguous object, the result we got after running image captioning is "a man flying through the air while riding a snowboard"

4 Extra Credit Implementation

4.1 TikaNER

We used several named entity recognizers to parse a description of each sighting from UFO V2 dataset and union all extracted entities together group by its categories (Location, Organization, Person, Date, Time, Measurements, Misc, Units).

First, we decomposed each sighting's description into separate text files. For each file, we pass its content to NER process. As for extracted categories of each recognizer will be described below

- OpenNLP (`NER/run_openNLP.sh`) for Location, Organization, Person, Date, Time
- CoreNLP (`NER/run_coreNLP.sh`) for Location, Organization, Person
- NLTK (`NER/run_NLTK.sh`) for Units
- MITIE (`NER/run_MITIE.sh`) for Location, Organization, Person, Misc
- Grobid Quantities (`NER/run_Grobid.sh`) for Measurements

Finally, all extracted entities are grouped by categories and added to final TSV dataset (`v2_dataset_final.tsv`). We also compressed all generated text files into the directory.

4.2 Image Recognition Deep Learning

To improve the ImageCaptioning and Image Recognition Deep Learning model we re-training `inception_V4` for Image Recognition Deep Learning model and `inception_V3` for ImageCaptioning and have pull requested both models with instruction and python test files to `img2text` github.

For both models, we classified images using 5 labels which are

- Flare - for lense flare, scattered light or sun
- Aircraft - for air plane, helicopter, rocket or drone
- Bird - for flying animal or insect
- Dust - for sensor dust or lens dust
- UFO - for unidentified flying object

`Inceptionv3` and `inceptionv4` revealed a very interesting result for our validate image. Figure 4b shows a sample validate image. The result for the retrain `inceptionv3` is 52% UFO, 33% Flare, and 11% Dust. And the result for the retrain `inceptionv4` is 91% Flare, and 8% UFO.

As you can see from the result above, `inceptionv3` thought that the validated image is the UFO and the `inceptionv4` thought that the image is the flare. As a human being, I believe that this image can be interpreted as both UFO and flare, thus the result from different model might show different result.

5 Question & Answer

Could changing things with ImageMagick, convert, etc. improve the OCR? Image Orientation? Handwriting?

We have done some experiment with ImageMagick (convert) parameter to clean the PDF file image (TIF). Without any updated parameter, tesseract failed on some of the images because those images were not clear enough to be OCR. We solved this problem by using the `-white` parameter to convert the image into black white image and use `-alpha off` parameter to turn off the alpha/matte channel. After using these parameter, tesseract worked for all images.

Next we try to clean the noise of the image using the ImageMagick plugin called Fred's ImageMagick script textcleaner. We added the parameter `"-g -e none -f 15 -o 10"` to the plugin. The result image were very good for human to interpret. However, tesseract returned an even worst OCR result. We then summarize that the quality of the image for the computer to understand sometimes doesn't similar to the quality that human wanted.

What you noticed about the dataset as you completed the tasks?

If we extract the PDF by ourselves manually, we could get a better result. The given PDF files contain a lot of noise and handwriting that tesseract can't detect. It is a very difficult task to develop a program to OCR such images even though we try to clean the noise as much as possible.

In the end, we obtained some data for the PDF OCR. And for the UFO stalking image captioning, the result is very promising for most images where the caption is fairly accurate for most component of the image.

What questions did your new joined datasets allow you to answer about the UFO sightings previously unanswered?

Yes, we can answer more questions, because the British UFO sightings document introduced some new features, such as type of observation, the angle of sight, and distance to UFO. With this features, we can answer some new questions. For example, by looking at the type of observation, we can conclude that most of the UFO records are discovered by naked eye. Also, by looking at the other witness, we can conclude that most of the UFO records are not highly-reliable because there is no other witness for the same record.

How well did the image captions accurately describe the UFO object types?

There is no UFO object generated from running image captioning process on UFO Stalker dataset. However, some of UFO-like objects, such as space shuttle, spotlight, kite, missile, balloon, and parachute, have been generated. The percentage of sighting reports contained objects are 52.48%

from all sighting reports that can generate objects. This number is not good enough in our opinion but, maybe, because the reports itself are possibly not accurate and the UFO-like objects in most images are very small.

What about the identified objects in the image?

Some objects are related to the image such as objects that mention in previous question. However, most labels are not related at all. Some examples are 'jack-o'-lantern', 'matchstick', 'fountain', etc. The result of low correlation between identified objects and images might caused from the using of generalized model. If we trained a specialized model to identify objects in images (As we did in extra credit 4.2). It probably results in a better accuracy.

How well did OCR work?

At first it doesn't work well, because it cannot read anything from some PDF pages. For example, for some pages in `DEFE-24-1922.pdf`, using the original pipeline script, OCR generate a blank text file, although there is important information in this page. We then found out that the problem is that TIF image has an alpha channel and therefore the underlying Leptonica package used by Tesseract doesn't support it. So we adjust some parameters, more detail in section 3.1.2. After that, OCR works better and can recognizes more pages.

What did you have to do to clean up the noise in the data?

To clean up the noise in the data, we apply several technique including image manipulation function from ImageMagick and also using additional TextCleaner plugin. More detail in section 3.1.2.

Of the incorporated British UFO sightings, how many of them could also similarly be explained akin to the sightings from the first assignment?

Only few rows can be explained akin to the first assignment TSV, because most of the time, Tesseract can't extract all features on every page especially sighted date and location which are key features for joining data together.

Were there any new object types introduced by the British UFO sightings?

Yes, there were new features. For example, type of observation (naked eyes, goggles), the angle of sight (ex. 45 degree), Distance to UFO, Distance to the landmark, Movement, Other witness. These features could be used for further analysis of UFO sightings.

How well were the British UFO sightings described?

The files are readable by human, but aren't good for Tesseract OCR. The files mostly describe size, shape, color and brightness of the UFO. The UFO information were very well described by the British UFO sightings file.

Was there a lot of missing data?

Yes, there are a lot of missing data for many of the British UFO sighting OCRs. However, for few rows, we are able to extract meaningful data including the date, location, description, shape, and duration.

Of the UFO images, how many of the images actually generated image captions and/or objects that described the UFO and not just the background scenery?

For our team, we didn't determine if each image will generate captions or sentences but we based on each sighting reports (per id). Following are the result.

- Out of 4065 sighting reports, 1824 reports have attached photos. (`ufostalker/reports.tsv`)
- 1206 reports can be used to generate image objects and sentences from its attached photos. (`ufostalker/objects.tsv`, `ufostalker/sentences.tsv`)

If we search through generated objects using a keyword 'UFO', we didn't come up with any result. So, we used other objects that look similar to 'UFO' as keywords instead (for our case we use 'space shuttle', 'spotlight', 'kite', 'missile', 'balloon', 'parachute'), we got 633 sighting reports that appear to have an UFO-like object in the image.

Additionally, when we looked at the list of objects generated by Object Recognition processing, we found that 'nematode worm' object has significant occurrence and by looking at images that have 'nematode worm' object, the object appears to be some weird light in the sky. So we added this as another keyword and get 718 sightings as a result.

Thoughts about OCR pipelining, and Image Captioning/Object identification. What was easy about using it? What wasn't?

OCR Pipelining

- Pros
  - By pipelining, we can easily enhance each step
  - Easy to maintain and debug
  - Better understandability
  - Separation of concerns
- Cons
  - Need more knowledge pipelining and shell script

Image Captioning & Object Identification

- Pros
  - Install these 2 libraries is easy with Docker
  - Web service API make them easy to use
  - Easy object tracking with automatic recognition
- Cons
  - Some learning curve for people who never use Docker
  - No user interface, only command line interface available
  - General model can cause incorrect results