USING FACIAL RECOGNITION TO GATHER SOCIAL MEDIA INTELLIGENCE

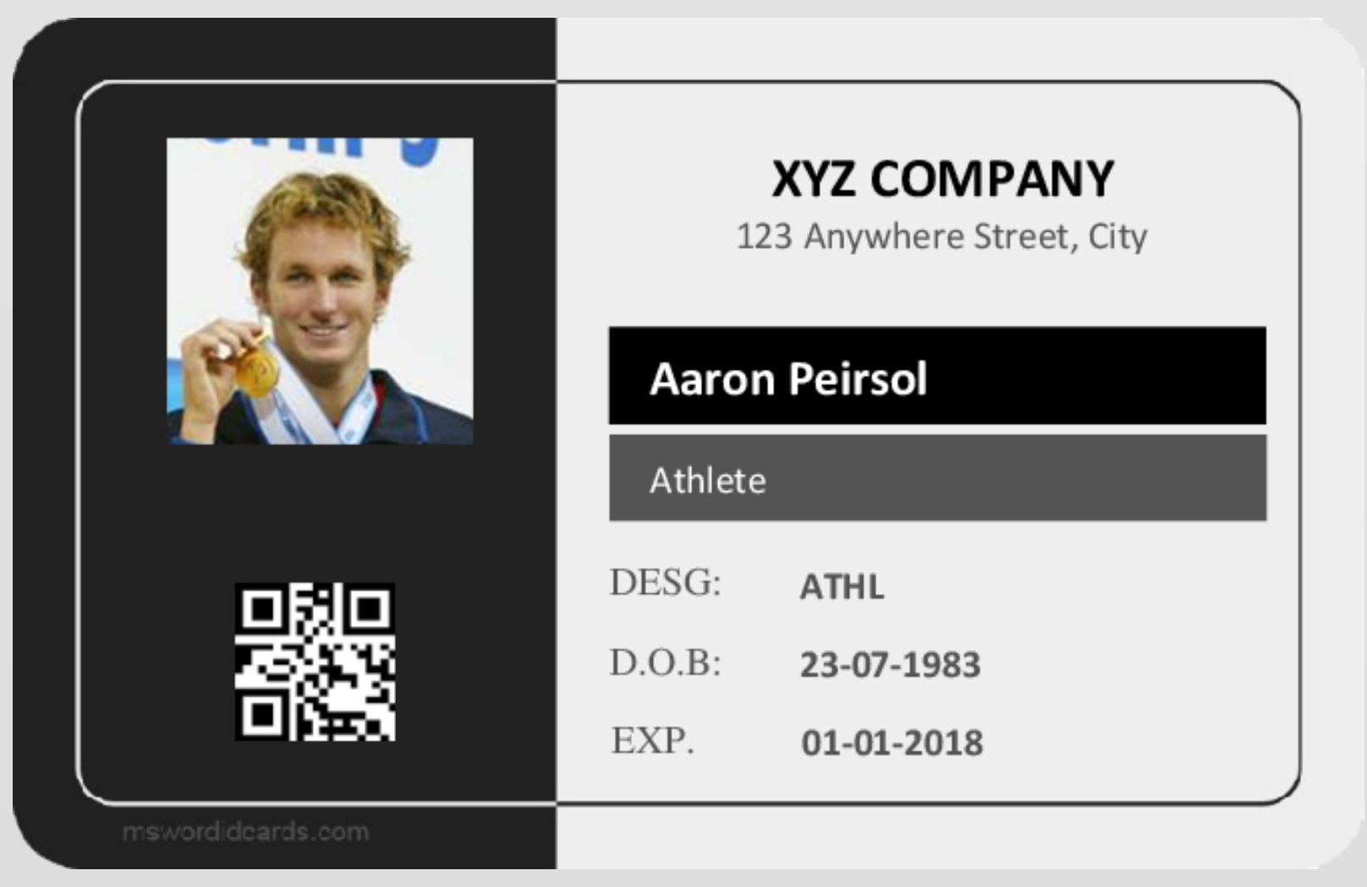
Abstract

Social Media Intelligence (SOCMINT) is a relatively unexplored source of intelligence, examining data contained within social media profiles. It is a subset of Open Source Information, where data on a subject is gathered using publicly available resources. Many see it as a vital aspect of modern information gathering – both the US FBI and the UK MOD have invested heavily in tools to utilize SOCMINT (Antonius and Rich, 2013)

There are very few examples of facial recognition being applied to SOCMINT to assist in gathering or analysis. It is therefore my aim to develop a tool will assist human operators in gathering social media intelligence by searching a database of faces to find potential matches to a face in a test image.

Introduction

Social media intelligence has been shown to have many uses, from detecting potential insider threads (Kandias and Stavrou, 2015) to increasing the effectiveness of prior knowledge attacks. The second point is particularly relevant, as it allows an adversary with no privileged information (OSINT only) to potentially gain access to networks or sensitive information by means of a spearphishing or social engineering attack.



XYZ COMPANY
123 Anywhere Street, City

Aaron Peirsol
Athlete

DESG: ATHL
D.O.B: 23-07-1983
EXP. 01-01-2018

Figure 1. An example of a fake ID card that could be created with a person of interest's name, date of birth and picture. Created by template from msword.dcards.com.

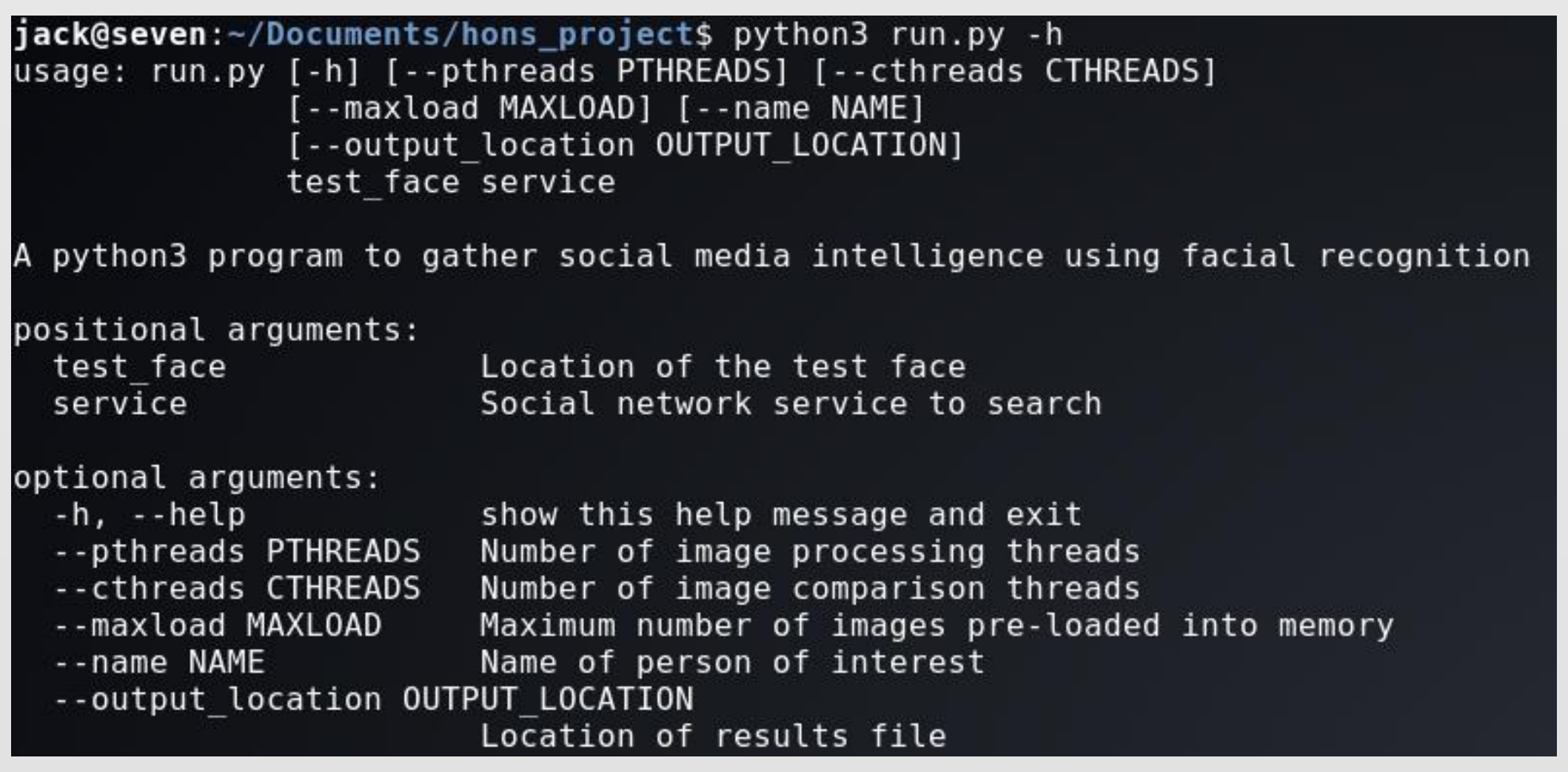
Although the usefulness of social media intelligence has been established, gathering and analysing it may not be as easy as it first appears. The problem for would-be analysts is the sheer amount of data contained within social media platforms – for example, 250 million photos are added to Facebook every day (Omand et. al., 2012).

To find useful information about a single person of interest in such a large data set is extremely challenging. Success rates increase dramatically if some starting point is known, such as a first name, a date of birth, or in this case the person's face. These factors may also be used in combination to increase accuracy and search speed.

Methods and Materials

Rather than search against actual social networks with real people, a database of test profiles was generated using the Labeled Faces in the Wild data set (Huang et. al., 2007), the Essex Face Recognition data set (Hond and Spacek, 1997), and the Security Camera Face data set (Grgic, Delac and Grgic, 2009). The tool used to search the data set was developed using Python version 3, using the face-recognition module with dlib as the underlying face recognition implementation. It includes a requirements.txt file to allow for easy installation of all pre-requisites using pip, and is used from the command line. A small demo file has been included that watches for results, and creates a HTML page that may be viewed for visualisation purposes.

The software was tested using data sets with different compression algorithms (JPG vs PNG), untrained data sets, data sets containing infra-red images, and data sets that were restricting by a piece of personally identifying information.



```
jack@seven:~/Documents/hons_project$ python3 run.py -h
usage: run.py [-h] [--pthreads PTHREADS] [--cthreads CTHREADS]
              [--maxload MAXLOAD] [--name NAME]
              [--output_location OUTPUT_LOCATION]
              test_face service

A python3 program to gather social media intelligence using facial recognition

positional arguments:
  test_face             Location of the test face
  service               Social network service to search

optional arguments:
  -h, --help            show this help message and exit
  --pthreads PTHREADS   Number of image processing threads
  --cthreads CTHREADS   Number of image comparison threads
  --maxload MAXLOAD     Maximum number of images pre-loaded into memory
  --name NAME           Name of person of interest
  --output_location OUTPUT_LOCATION
                        Location of results file
```

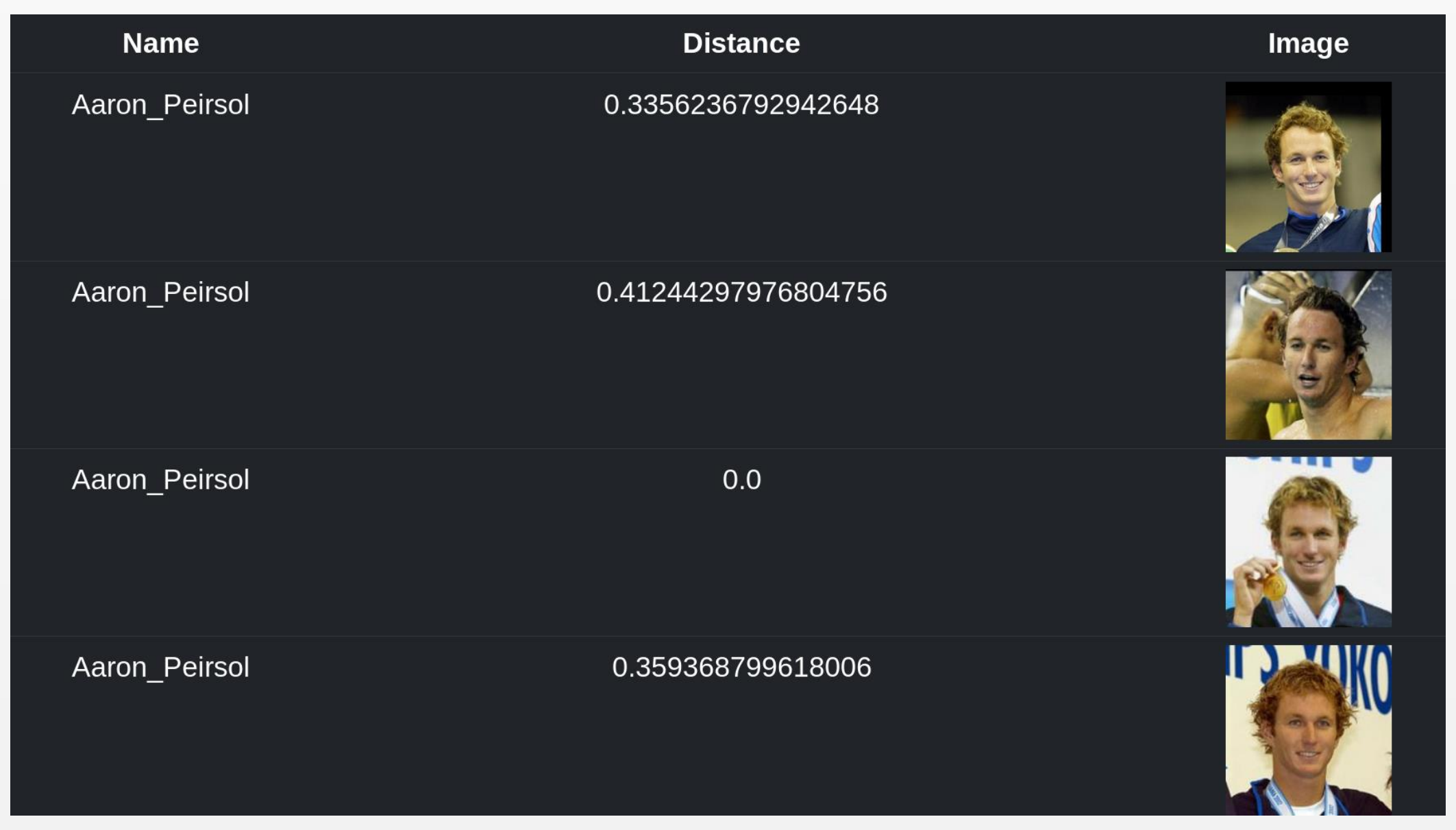
Figure 2. The "help" message that shows all positional and optional arguments that may be used with the program.

Results

The software performed excellently on both the untrained data set and the restricted data set, achieving 100% accuracy with 0 false positives over 5 tests in both cases. Of particular note is the restricted data set – by massively reducing the number of images to compare, the average length of each search was reduced to just 25 seconds.

When searching across a data set with a different compression algorithm, the same results were noted. However, the distance between two otherwise identical images was non-0 suggesting that the compression algorithm may produce some baseline error during comparison.

Searching across the SCFace data set with the infra-red images produced much worse results. While the original image was always compared with a distance of 0, the accuracy was negatively impacted by the presence of several false positives. It is theorised that the IR images lack facial detail, and are difficult to place descriptors on due to the lack of colour.



Name	Distance	Image
Aaron_Peirsol	0.3356236792942648	
Aaron_Peirsol	0.41244297976804756	
Aaron_Peirsol	0.0	
Aaron_Peirsol	0.359368799618006	

Table 1. Results when using searching with the test image Aaron_Peirsol_0001.jpg, restricted with the term "Aaron*" (i.e. search for all people named "Aaron").

Conclusions

The program implemented shows that, yes, it is possible to use facial recognition to assist in the gathering and analysis of social media intelligence. The automation of this task could be greatly beneficial to intelligence services, private firms wishing to gather intelligence, or private citizens wishing to see their SOCMINT footprint. There are however several caveats:

While testing showed promising results in terms of accuracy, the time taken for each search is excessive. The program can undoubtedly search large data sets faster than a human could, but having searches last over an hour over a small data set (in comparison to all profile photos on a social networking service) means that it is unlikely to be able to scale well.

As a corollary, since the time taken for each search is so long only limited testing could be performed. It is likely that the 100% accuracy rating achieved over the tests is close to the true accuracy that the program could achieve, the small sample size of tests makes it difficult to say this definitively.

Finally, the service this program provides may not be accessible to all. Using it against a real social network rather than a test data set would require setting up a proxy server to present data in the expected manner, as well as having the service in question allow access to the full (or at least, a sizeable portion) collection of profile images. This is feasible for security services and large corporations but is likely outside the scope of what a single person could acquire.

References

Antonius, N. and Rich, L. (2013), 'Discovering collection and analysis techniques for social media to improve public safety', 3, 42.

Kandias, M. and Stavrou, V. (2015), 'Personal traits analysis as a means to predict insiders'.

Omand, S. D., Bartlett, J. and Miller, C. (2012), 'Introducing social media intelligence (socmint)', *Intelligence and National Security* 27(6), 801-823.

Huang, G., Ramesh, M., Berg, T. and Learned-Miller, E. (2007), 'Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments'.

Hond, D. and Spacek, L. (1997), 'Distinctive Descriptions for Face Processing'.

Grgic, M., Delac, K. and Grgic, S. (2011), 'SCFace – surveillance cameras face database'.

ROBERT GORDON UNIVERSITY ABERDEEN

Jack Neilson; 1506801