Mapping of Facial Recognition on Social Media using Machine Learning

Priyanka Tyagi[1]
“Department of Computer Science & Engineering”
School of Engineering & Technology Sharda University
Greater Noida, India
priyanka.tyagi@sharda.ac.in

Neelam Shrivastava[2]
Dept. of Computer Science Accurate Institute of Managemnt & Technology
Greater Noida, India
neelam.engr@gmail.com

Rakshita Mall[3]
“Department of Computer Science & Engineering”Sharda University
Greater Noida, India
rakshita.mall@sharda.ac.in

Kajol Mittal[4]
“Department of Computer Science & Engineering”
School of Engineering & Technology Sharda University
Greater Noida, India
kajolmittal28@gmail.com

ABSTRACT: **More and more people are sharing on social media knowing they are being watched, but not knowing how much or how their data is being processed. Performing monitoring is resource intensive, both in terms of effort and calculation. This research looks at the scope of data collection and processing by demonstrating that data from Twitter and Reddit can only be acquired, processed, and stored in real-time utilizing a personal computer. The primary goal was to discover individual persons in the data stream using face recognition, but the obtained data may also be utilized for other purposes. We also considered ethical issues related to the collecting and processing of such data.**

Keywords: Mapping, Machine Learning, social media, surveillance

I. INTRODUCTION

We're using social media more and more to share our experiences with the rest of the world through text, photographs, and videos. Nations and organizations utilize systems to collect information on its inhabitants and users at the same time [20]. Monitoring, on the other hand, is not a novel notion. Surveillance technology has long been utilized by police and security services to undertake investigations. State-owned firms like China, as well as private corporations like Facebook, now have unprecedented access to data [9, 4]. They also have access to computer technology that allow them to seek and analyze information at fast speeds [19].This project's goal is to increase awareness by demonstrating that this sort of processing and collecting is not limited to huge companies or countries, but is equally feasible for smaller groups.

According to a Pew Research Center analysis, social media use more than tenfold rose between 2005 and 2015 [30]. More specifically, in 2005, In 2010, only 7% of all American adults used social media; by 2015, that figure had risen to 65% [30]. As more individuals utilize social media, it has the potential to become a useful source of monitoring data. Twitter, Facebook, Snapchat, Reddit, and Instagram, for example, make user data available on the internet via APIs and interfaces [22, 8]. Although consumers of these services are concerned about their digital footprint, the privacy regulations that define the data gathered sometimes mislead them [31].The goal of this research is to show that data from social media platforms can only be gathered, processed, and stored in real time utilizing a personal computer. It uses facial recognition to analyze data in order to identify a specific individual inside it. It is crucial to assess if collecting personal data without consent is ethical or lawful. Even if the project only shows what is possible, the end result may not always justify the methods. As a result, our system has been halved. The system continues to download data but does not save it anyplace. Instead of using social media data, use a Reddit-like data source. This method does not gather or store "live" data. This research has resulted in the capacity to gather and interpret data streams in real-time, similar to Reddit. We have not examined the facial recognition component's accuracy.

II. LITERATURE REVIEW

This section begins with a quick overview of monitoring. It then goes into existing social media platforms, biometrics that can be followed using social media facial recognition, and lastly the General Data Protection Regulation.It is critical to specify the monitor. Surveillance is defined as the monitoring of personal data for reasons such as influencing, controlling, protecting, or guiding [23]. It's critical to stress that this is a general data collecting issue, not merely a physical surveillance one involving cameras. This definition is used throughout the paper.Long before computers and cameras, people could be watched. Physical surveillance was used at first. Because everything has to be done manually by the individuals tracking the people being monitored, this sort of surveillance is quite resource demanding [23].

The earliest documented usage of video surveillance was in 1942. This technology has enabled more individuals to be monitored [22]. The initial surveillance system could not capture video and had to be manually presented. Years later, video recording became possible, and surveillance systems in 1990 enabled him to display video footage from several cameras on a single screen.China has lately demonstrated how contemporary technology may be used to monitor enormous populations. An NPR [34] piece describes how face recognition technology may be used with surveillance cameras to watch individuals.We can collect information about our residents using this monitoring technology. The data is then compiled into a ‘score,' often known as a social credit score. Scores determine everything from internet speed to dog ownership [24].The popularity of social media has skyrocketed in recent years. In the previous year, the number of active social media users increased by 9% to 3.5 billion. This equates to 45% of the global population [12]. Personal information such as location information, postings, and photographs are tracked on social media [23].The data presently given by the five main social media networks is summarized here. This information can be collected via application programming interfaces (APIs), which provide programmatic access to the data, or by web scraping, in which your computer acts as the user and acquires the information.Facial recognition software allows devices to locate and identify people based on their facial images. By calculating a unique ID based on facial features, you can verify that two photos of faces contain matching faces. [25].A person can be identified in a variety of ways. B. Iris and fingerprint scanning, the latter of which is often used in applications such as telephones [25].

Face recognition has an advantage over other biometric identification systems in that the data subject is often unaware of it, and the source data is publically available. The acquisition of fingerprints is difficult to hide. B. When requesting a biometric ID card or passport. People are required to lay their hands on the scanner. Modern technology, like as in-display fingerprint scanners, can help to obscure data collecting, but still remains visible. Recognizing individuals on social media with face recognition is a growing area. His Social Mapper [24], an open-source utility, is one of his applications that performs this. This application allows you to locate a user's numerous social media profiles without having to search for them manually. All you need is the name and photo of the individual whose social media profiles you wish to collect. Simply enter this information into the Social Mapper tool, which will then compare the image you supply to photographs from the most prominent social networking applications that employ face recognition. Following this comparison, the individual's social media profiles are displayed.Social Mapper and our software are connected since both utilize face recognition to discover individuals on social networking. Aside from that, our method is rather unique. B. Use photographs from the live broadcast for facial recognition, but use previously released photos. Our system's application is also distinct. Collect data about people while social mappers supply various accounts.

III. EXPERIMENTAL ANALYSIS

Our system is split into two sections (see Figure 1). The first component is the data gathering component, which gathers information from social media networks like Twitter and Reddit. The second stage is to analyze false data using facial recognition.

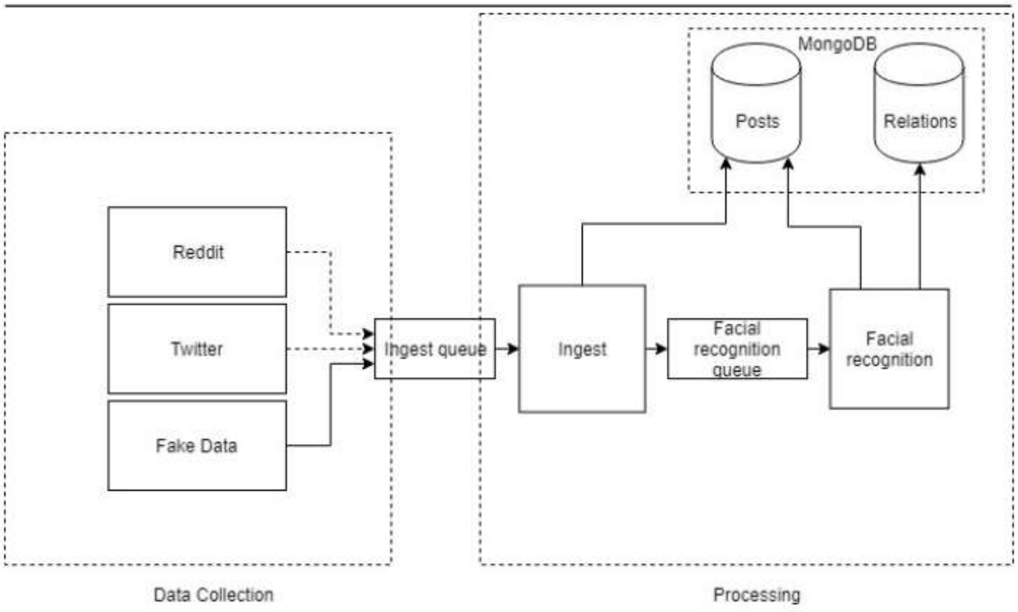

Figure 1: System flow of data collection and processing

A. COLLECTION OF DATA

The first component of the system is data collection, which is divided into two modules: one for Reddit data and one for Twitter data. As soon as these applications are available, data is gathered and live broadcast. The system mines the data for picture URLs, users, geolocations, and other information. Instead of processing personal data, incorrect data sources are utilized. This data source is patterned after Twitter and Reddit data collecting. The frequency with which bogus data is sent is intended to match the frequency with which it is posted on selected social networking sites. The data is subsequently submitted as a standardized JSON object to the ingestion queue.

B. PROCESSING OF DATA

The system's processing section deals with bogus data supplied from data collecting.

1. INGEST

An ingest queue in this component of the system gets JSON objects containing data from the faked stream. Insert the supplied information into the post database. After that, the post ID and picture URL are routed to the face recognition queue.

2. RECOGNISING FACE

A JSON object is received from the facial recognition queue. Checks if the object contains an image URL and if so downloads the image. When the system successfully downloads an image, it checks if it contains faces. If no face is found, the image is discarded. When the system recognizes a face, it compares it to a stored data set of faces to find similar faces. If a match is detected, the individual in the photo's name is added to the relationship database.

C. MONITORING OF DATA

Queue monitoring is required to comprehend performance concerns, particularly those linked to queue length. RabbitMQ, the system's message broker, has an administrative webpage that displays basic information such as messages entering and exiting each queue. We used Prometheus, a time-series database, to store our metrics because RabbitMQ allows us to export them and Prometheus has a query API. Grafana was also used to display the status of the pipeline based on Prometheus data. Grafana presents the current backlog along with a graph of messages entering and exiting the ingestion and face recognition queues. Figure 2 depicts a picture of the arrangement.

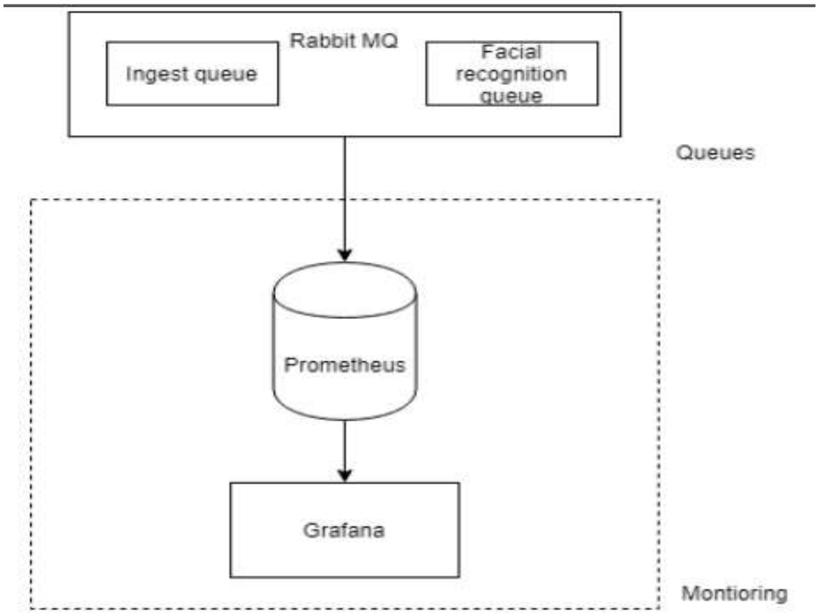

Figure 2: System overview of the queue setup

IV. APPROACH

This section describes how to design your pipeline. First, let's talk about data collection. Here we explain how we collect data, the social media from which we collect it, and how we treat it. After this section, continue processing the data. The processing part discusses two topics: selecting a facial recognition system and selecting a database. Let's move on to system design and scalability now that we've covered processing. We describe why we scaled, how we scaled, and design alternatives for scalability here. Finally, there is a section outlining why the system logs are significant and the techniques we employed.

A. COLLECTION OF DATA OR DATA GATHERING

This section discusses the system's data collecting component.

A.1. COLLECTING DATA FOR THE PIPELINE

Web scraping and APIs are the two major methods for collecting data from social media.Online scraping is the technique of extracting information from web pages using a computer application. B. Social networking sites [11]. According to the authors, this is useful when social media platforms do not allow official access to their data, or when looking for data that has not been publicly disclosed but is subject to the platform's terms of service [11]. There is no web scraping there. We didn't employ his web scraping since we wanted to stick with his data supplier.Another alternative is to directly acquire data from social networking sites via API [11]. The entire data collection may be purchased via services such as Twitter [11]. This, however, is not an option for individuals with little finances. Some businesses offer free public APIs with data samples [11]. We used public APIs for this project since social media firms' public APIs are available.

B. FAKING DATA

Because I didn't want to deal with authentic social media data for ethical and legal reasons, I opted to create bogus data. The false data was created using images from the Labeled Faces in the Wild collection and faceless stock shots [13]. We used a variety of photographs to represent the social media material. Fake data, including data from Reddit and Twitter, is provided to our system.

C. PROCESSING

This section describes design options for facial recognition systems and databases.

C.1. RECOGNISING FACE

Facial recognition technologies capture photographs and evaluate faces to determine identities. This requires a data bank of known people's photos to compare with the received photographs, as well as a system to perform the actual comparison.It is time-consuming to manually create massive databases of tagged faces. Images with annotated faces contain faces whose owners are known. Manual marking, on the other hand, is not necessary. Online, you may find photos with facial tags. Labeled Faces in the Wild [13] is one of his. The photos in the data set are in their raw form, with varying postures, lighting, features, accessories, occlusions, and backdrops. We believed this was a good fit because the social media photographs were shot under a range of settings.There are several methods for comparing faces and records. The purpose of this project is to demonstrate that it is doable and very simple, thus I choose to utilize a Python-based face recognition package maintained by Adam Geitgey [10].

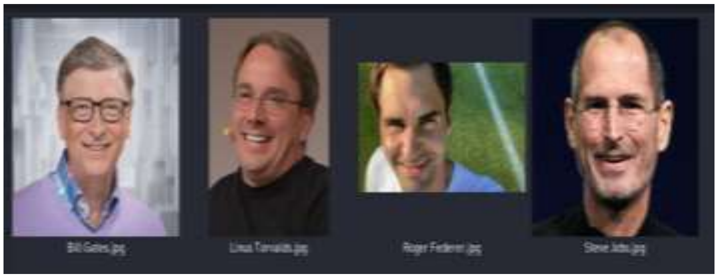

Figure 3: Input for face recognition

The DLIB's facial recognition API [18] is implemented in this library. DLIB is a machine learning algorithms library written in C++. The face recognition library lets you transform a face-filled picture into numerous vectors, one for each face. When two vectors are near together, they are most likely from the same person's face.

C.2. DATABASE

It required a database to hold the collected posts, as well as the face vectors and names of the people in the photographs. We decided to select the database based on the work of Parker et al. [28]. The author compares his MongoDB and SQL databases for NoSQL databases. This whitepaper explains why MongoDB is a great choice for schema-free databases. It is also mentioned that MongoDB outperforms SQL databases when it comes to inserts. We chose MongoDB, which doesn't require a strict schema, as we may in the future, the database schema will need to be changed. Another factor that contributed to our choice is that Mongo DB is faster during inserts. We want to collect data in real time, therefore quicker inserts help.

D. SCALABLE DATA

One of the project's objectives is to make the system scalable. This section examines system modularization and how it aids scalability.

D.1. MODULARIZATION

One of the requirements is that the system must be scalable. H. There may be multiple instances of each module on multiple servers. Modules also have varied hardware requirements. B. The advantages of facial recognition with GPU access As a result, I required a method to configure and deploy the system. These issues can be addressed by using virtual machines or containerization. Figures 3 and 4 show the difference between the two solutions. Essentially, a virtual machine (VM) runs a guest operating system (OS) running on top of a host OS, reducing resource requirements, as opposed to containers running directly on the host [27].Tools for packaging and delivering programmers are provided by containerization systems such as the Docker platform [27]. I chose Docker to containerize it. This occurred because our approach does not necessitate complete insulation. Because of the guest operating system, VMs demand extra resources. This contradicts the purpose of utilizing as little energy as possible. Furthermore, because Docker is extensively utilized as a containerization technology, It features a well-established user base and support platform [27].

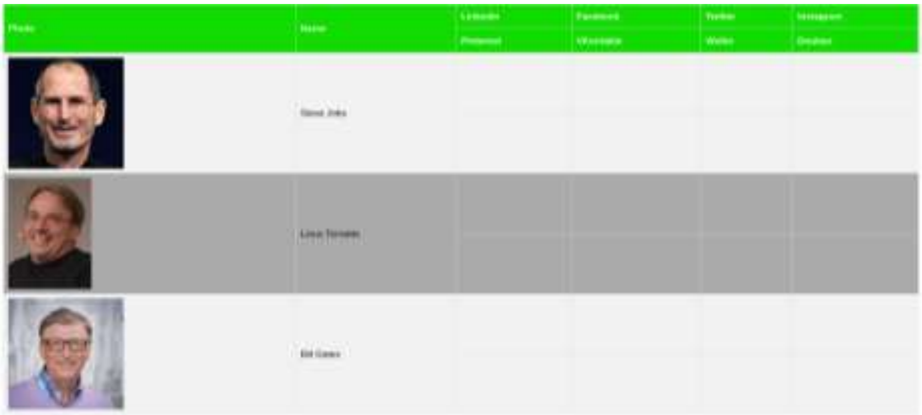

Photo	Name	LinkedIn	Facebook	Twitter	Instagram
		Pinterest	VKontakte	Weibo	Douban
	Steve Jobs				
	Linus Torvalds				
	Bill Gates				

Figure 4: Result of the input data

E. QUEUE

To control the flow of data between containers, use queues. Queues let each container to consume at their own pace. Data collectors can also push data as quickly as they can. This is excellent. Because data rates may not be consistent, you might accumulate a backlog during peak data collecting and use it up when it drops. Because we needed real-time processing, we chose a first-in-first-out (FIFO) queue for our project. RabbitMQ was used to implement queues.

V. CONCLUSION

Our study employs real-time social media data as a source for a facial recognition system. To gather real-time data from social media, two data collecting modules are built. Face recognition can identify a specific individual using real-time data and create a database of facial data that may be utilized later. The goal was to show that picture data from social media platforms may be acquired and processed for surveillance purposes. We demonstrated how to harvest face data from two major social media networks, Reddit and Twitter. It also demonstrated how to make use of this information. We intend to raise awareness of the possibility of monitoring individuals through social media by demonstrating how simple our system is to adopt.

VI. FUTURE SCOPE

This experiment demonstrates how personal information may be gathered via social media. We feel you can tell folks even more with this knowledge. B. People's geographical location and political leaning. Text from tweets and Reddit is also collected by the system posts that may contain information about places, friends, interests, etc. You can extract this data using natural language processing to create a richer profile.

The only data sources implemented in the system are Reddit and Twitter, Data may also be gathered from other social media networks. Snapchat and Instagram both show potential. Snap Map is a feature offered by Snapchat. Users may upload photographs and movies to public maps, which include strippable images as well as location data. Instagram doesn't have a public API, but it's accessible without authentication, so it's possible to scrape data from it.

Currently the system can only use databases that contain annotated images. B. Label wild faces. The integration of self-training is another future enhancement. This will add fresh faces from the streaming data to the database. This enables the system to constantly increase facial recognition accuracy.

References

[1] T. Bray, "The JavaScript Object Notation (JSON) Data Interchange Format," Internet Requests for Comments, RFC Editor, RFC 8259, December 2017, retrieved 2019-04-28. [Online]. Available: https://tools.ietf.org/html/rfc8259

[2] G. W. Brown, The concise Oxford dictionary of politics and international relations: edited by Garrett Wallace Brown, Iain McLean, and Alistair McMillan, 4th ed. Oxford, United Kingdom: Oxford University Press, 2018.

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[4] C. Cadwalladr and E. Graham-Harrison. (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. Retrieved 2019-05-06. [Online]. Available: https://www.theguardian.com/news/2018/mar/ 17/cambridge-analytica-facebook-influence-us-election

[5] S. Chinoy. (2019, April) We built an 'unbelievable' (but legal) facial recognition machine. Retrieved 2019-04-17. [Online]. Available: https://www.nytimes.com/ interactive/2019/04/16/opinion/facial-recognition-new-york-city.html

[6] EU. EU Charter of Fundamental Rights. European Union. Retrieved 2019-04-21. [Online]. Available: https://fra.europa.eu/en/charterpedia/title/ii-freedoms

[7] EU. (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. European Union. Retrieved 2019-04-08. [Online]. Available: https://eurlex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[8] (2019) Facebook for developers. Facebook. Retrieved 2019-05-06. [Online]. Available: https://developers.facebook.com/docs/graph-api

[9] R. Gallagher. (2016) Documents reveal secretive U.K. surveillance policies. Retrieved 2019-05-06. [Online]. Available: https://theintercept.com/2016/04/20/ uk-surveillance-bulk-datasets-gchq/

[10] A. Geitgey. (2017) Face recognition. Retrieved 2019-04-17. [Online]. Available: https://github.com/ageitgey/face recognition

[11] S. Halford, M. Weal, R. Tinati, L. Carr, and C. Pope, "Understanding the production and circulation of social media data: Towards methodological principles and praxis," New Media & Society, vol. 20, no. 9, pp. 3341–3358, 2018. [Online]. Available: https://doi.org/10.1177/1461444817748953

[12] (2019) Digital 2019. Hootsuite Inc. Retrieved 2019-05-06. [Online]. Available: https://p.widencdn.net/kqy7ii/Digital2019-Report-en

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[14] (2019) UN: China Responds to Rights Review with Threats. Human Rights Watch. Retrieved 2019-05-20. [Online]. Available: https://www.hrw.org/news/ 2019/04/01/un-china-responds-rights-review-threats

[15] (2019) API endpoints. Instagram. Retrieved 2019-05-06. [Online]. Available: https://www.instagram.com/developer/endpoints/

[16] P. Karp. (2018) Australia's war on encryption: the sweeping new powers rushed into law. Retrieved 2019-05-20. [Online]. Available: https://www.theguardian.com/technology/2018/dec/08/australias-war-onencryption-the-sweeping-new-powers-rushed-into-law

[17] O. S. Kerr, "Internet surveillance law after the USA Patriot Act: The big brother that isn't," Nw. UL Rev., vol. 97, p. 607, 2002.

[18] D. E. King, "Dlib-ml: A machine learning toolkit," Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[19] A. Kofman. (2018) Interpol rolls out international voice identification database using samples from 192 law enforcement agencies. Retrieved 2019-05- 06. [Online]. Available: https://theintercept.com/2018/06/25/interpol-voiceidentification-database/

[20] L. Kuo. (2019) Chinese surveillance company tracking 2.5m xinjiang residents. Retrieved 2019-04-07. [Online]. Available: https://www.theguardian.com/world/ 2019/feb/18/chinese-surveillance-company-tracking-25m-xinjiang-residents

[21] J. Lechtenborger. (2018) Figures. Retrieved 2019-05-20. [Online]. Available: ¨ https://gitlab.com/oer/figures

[22] A. Longdin. (2014) The history of CCTV – from 1942 to present. Retrieved 2019- 05-09. [Online]. Available: https://www.pcr-online.biz/2014/09/02/the-historyof-cctv-from-1942-to-present/

[23] D. Lyon, Surveillance Studies: An Overview. Cambridge CB2 1UR, UK: Polity Press, 2007.

[24] A. Ma. (2018) China has started ranking citizens with a creepy 'social credit' system — here's what you can do wrong, and the embarrassing, demeaning ways they can punish you. Retrieved 2019-04-08. [Online]. Available: https://nordic.businessinsider.com/china-social-credit-system-punishmentsand-rewards-explained-2018-4

[25] J. McLaughlin. (2016) Private intellligence firm proposes "google" for tracking terrorists' faces. Retrieved 2019-04-07. [Online]. Available: https://theintercept.com/2016/11/04/private-intellligence-firm-proposesgoogle-for-tracking-terrorists-faces/