Undergraduate Thesis Prospectus

Applying Computer Vision to Track Library Occupancy

(technical research project in Computer Science)

Resisting the Surveillance State:
How Americans Are Fighting to Stop Mass Data Collection

(sociotechnical research project)

by

Gabriel Silliman

October 27, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Gabriel Silliman

Technical advisor:	Briana Morrison, Department of Computer Science
STS advisor:	Peter Norton, Department of Engineering and Society

General Research Problem

How is data collected from individuals without their knowledge or consent?

As of 2022, there are an estimated 14.4 billion internet-connected devices around the world (IoT Analytics, 2022). The ubiquity of the internet allows for individuals' locations, opinions, and personal information to be tracked and monetized. Through location data from a weather app, a quick Google search, or a camera in a public area, organizations can capture a vast amount of data on much of the U.S. population without their knowledge. Regardless of how the data is used, the ubiquity of data surveillance has massive implications for the United States and the world.

Applying Computer Vision to Track Library Occupancy

How can UVA students see the occupancy of study spaces around campus in real time?

This is an independent technical project for the Computer Science department and my technical advisor is Professor Briana Morrison.

At the end of each semester, as deadlines approach and exams loom ahead, students pack the libraries around the grounds of the University of Virginia. Even in the late hours of the night, rows of students fill the tables on all floors of the libraries, making it difficult to find a quiet place to study. Nowadays, certain apps, such as Google Maps, will provide users with live trackers of how busy an area is at any moment, as well as predicted levels for the future. This is possible due to Google's mass data collection practices, giving them access to many users' precise location data (D'Zmura, 2020). Regardless of the privacy concerns that come with this level of location monitoring, it is not practical or feasible to monitor the number of students in a library using this method.

For this project, computer vision techniques can be used to identify how busy a specific area is by tracking the number of people in the camera's view. I will train a convolutional neural network (CNN) to identify individuals sitting or standing. The trained CNN will then be used to analyze frames from a video of the library to identify how many people are in the study space. There has been extensive research in object detection using neural networks, and CNNs are effective at identifying the existence of objects, including humans, in still images. Basalamah et. al. (2019) proposed a scale-driven convolutional neural network (SD-CNN) model to count the number of people in crowded areas. Their model assumes that heads are the dominant, visible feature in each static frame of analysis. The SD-CNN scales features by their 2-dimensional location in the frame to account for the perspective of the camera's view on the size of each head. Based on this scale map, object proposals are generated and classified into head or background classes which are used to generate a response map of each head in the image. This can be used to count or locate each person and analyze the human density in the image. This method is especially effective for the analysis of crowded areas where multiple people overlap within the frame and therefore applies to study spaces where students are sitting.

Once the CNN model can effectively determine the number of people in a space based on a video frame, I will test it on footage recorded in the libraries with varying numbers of students to generate accurate qualitative metrics, such as packed, busy, quiet, or empty. This information will be accessible via a mobile application. I will build the user interface using React Native, a mobile development software framework, and deploy the backend with Amazon Web Services. The app will allow students to check the activity levels of areas on campus to determine where they want to study. This will be a useful tool for students, especially around midterms and finals when the libraries are packed.

Resisting the Surveillance State: How Americans are Fighting to Stop Mass Data Collection

How are social groups resisting mass surveillance and data collection?

Mathematician Clive Humby coined the phrase "data is the new oil" in 2006, claiming that neither can be used unrefined, but both are extremely valuable when processed correctly (Humby, 2019). Today, digital media companies are worth trillions of dollars due to their monetization of data. In 2022, companies are predicted to generate and collect around 100 trillion gigabytes of data (IDC & Statista, 2021). Some data is used internally, but it is mostly sold to advertisers, data brokers, and government agencies. Meta alone generated $118 billion in revenue last year, 97% of which was from advertising (Meta, 2022).

Despite their claims, companies do not always handle their data correctly. In 2014, Cambridge Analytica gained access to up to 87 million Facebook users' data, including names, birthdays, locations, and likes (Schroepfer, 2018). Using these data, Cambridge Analytica matched over 30 million users with other public records to build psychographic voter profiles, which were in turn used by the political campaigns of Ted Cruz and Donald Trump in 2016 to micro-target political advertisements to individuals' views (Rosenberg et al., 2018). In response, Facebook was fined $5 billion for the mishandling of users' data, however, the damage had already been done and the extent to which it affected the 2016 election may never be known (Romm, 2019). Besides Meta, there have been countless other data breaches, and personal information tied to at least 11 billion compromised accounts has been released online (Hunt, 2022). In these scenarios, there is not much recourse outside of fines, which have little effect on massive corporations that can afford to pay them.

Beyond data leaks, users' data is frequently used in completely legal, but ethically ambiguous ways. Users give their location data to countless services, including seemingly harmless weather or navigation apps. Many developers sell access to their users' locations to data brokers who aggregate it and trade this data between themselves, making it difficult to determine the source. Some data brokers, such as Venntel and Babel Street, then sell the location data to US government agencies (Cyphers, 2022). According to public contracts available from the Federal Procurement Data System, these two companies appear to be selling hundreds of thousands of dollars of location data per year to government agencies including the Federal Bureau of Investigation, Drug Enforcement Administration, Secret Service, Customs and Border Protection, and Immigration and Customs Enforcement (GSA, 2022). According to an anonymous source who works with Venntel, the location data is device-specific, meaning that it could be associated with an individual if their location were known at a specific time (Cox, 2020). By using data brokers, government agencies can circumvent restrictions on what data they can collect without warrants and how they can use that data.

In the past decade, the public has become more aware of government geosurveillance tactics due to highly public data leaks. The most prominent leak was released by Edward Snowden, a former CIA employee and contractor for the National Security Agency. In 2013, Snowden leaked documents showing how the NSA was extracting personal "audio and video chats, photographs, e-mails, documents, and connection logs" from nine major US internet companies (Gellman & Poitras, 2013). The documents leaked by Snowden spurred nationwide awareness and discussion on government surveillance practices. According to Swanlund & Schuurman (2019), these leaks have also allowed individuals and organizations to evade existing techniques using "data minimization, obfuscation, and manipulation" and prompted

organizations to file lawsuits to attempt to stop their use by government agencies. To prevent the continued misuse of personal data, stronger data privacy laws must be passed, focused on providing transparency, disclosing data practices, and giving users direct control over their data (Isaak & Hanna, 2018). Europe made the largest single change to data protection practices when the EU passed the General Data Protection Regulation (GDPR) in 2016, which requires companies to inform users what data they collect and how it will be used, get consent before collecting data, and access or erase all user data on request (Wolford, 2018). California followed Europe's lead and became the first US state to implement a data privacy law when the California Consumer Privacy Act (CCPA) was passed in 2019. The CCPA gives residents the right to know what data is collected and how it is used or shared, the right to delete their data, and the right to opt out of the sale of their data (CDOJ, 2019).

While new laws are forcing companies to inform users about their data collection practices, many people do not have the time to read a complicated privacy policy every time they visit a new website and will blindly agree without fully understanding what data is being collected and how it is being used. Companies know that most people will not change the default data privacy settings, and users should not have to be data privacy experts or be technically savvy to limit how their data is used. The Institute of Electrical and Electronics Engineers provides frameworks for how this problem can be mitigated by allowing individuals to "create, curate, and control" their data and identity online (IEEE, 2019). The IEEE suggests that individuals should be able to create personalized terms and conditions for how their data is collected and used, terms should be presented in an easily understandable way, and data collection consent should be revokable at any point. They also recommend providing users with algorithmic agents that can act on their behalf to automatically reflect their data collection terms

to each website or service and inform the user if an organization is found to be using an individual's data in violation of their terms.

As users do not currently have control over most of their data, they should be aware of how their data is being collected and what can be done to stop it. Digital privacy groups such as the Electronic Frontier Foundation are working to educate consumers on the data collection practices of companies and help individuals avoid surveillance (EFF, 2017). The EFF also offers various tools such as Privacy Badger and Cover Your Tracks, which uncover and stop advertisers and trackers from secretly monitoring users' data across websites.

Technology companies and data brokers have profited from their roles in the global data market for years. However, more recently, due to wide-scale data leaks, new data privacy laws, and international advocacy, the public is starting to fight back against the aggressive collection and monetization of their data. If this momentum is to continue, various approaches must be taken to restrict pervasive data surveillance by companies and government agencies.

References

Basalamah, S., Khan, S. D., & Ullah, H. (2019). Scale driven convolutional neural network model for people counting and localization in crowd scenes. *IEEE Access*, *7*, 71576–71584. https://doi.org/10.1109/ACCESS.2019.2918650

Cox, J. (2020, August 25). Customs and Border Protection Paid $476,000 to a Location Data Firm in New Deal. Vice. https://www.vice.com/en/article/k7qyv3/customs-border-protection-venntel-location-data-dhs

Cyphers, B. (2022, June 13). How the Federal Government Buys Our Cell Phone Location Data. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2022/06/how-federal-government-buys-our-cell-phone-location-data

D'Zmura, M. (2020, October 15). Behind the scenes: popular times and live busyness information. Google. https://blog.google/products/maps/maps101-popular-times-and-live-busyness-information/

EFF. (2017, May 9). Tools from EFF's Tech Team. Electronic Frontier Foundation. https://www.eff.org/pages/tools

Gellman, B., & Poitras, L. (2013, June 7). U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program. *Washington Post*. https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html

Humby, C. (2019, July 4). Big Data and AI is not a panacea. LinkedIn. https://www.linkedin.com/pulse/big-data-ai-panacea-clive-humby-obe/?articleId=6552497074594734080

Hunt, T. (2022). Have I Been Pwned: Check if your email has been compromised in a data breach. Haveibeenpwned.com. https://haveibeenpwned.com/

IDC, & Statista. (June 7, 2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes) [Graph]. *Statista*. https://www.statista.com/statistics/871513/worldwide-data-created/

IEEE (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (1st Edition). IEEE. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf

IoT Analytics. (2022, September 1). IoT 2022: Connected Devices Growing 18% to 14.4 Billion Globally. IoT for All. https://www.iotforall.com/state-of-iot-2022

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, *51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268

Keegan, J., & Ng, A. (2021, September 30). There's a Multibillion-Dollar Market for Your Phone's Location Data – The Markup. The Markup. https://themarkup.org/privacy/2021/09/30/theres-a-multibillion-dollar-market-for-your-phones-location-data

Meta. (2022, February 2). Meta Reports Fourth Quarter and Full Year 2021 Results [Press Release]. https://s21.q4cdn.com/399680738/files/doc_financials/2021/q4/FB-12.31.2021-Exhibit-99.1-Final.pdf

Romm, T. (2019, July 12). FTC votes to approve $5 billion settlement with Facebook in privacy probe. *Washington Post*. https://www.washingtonpost.com/technology/2019/07/12/ftc-votes-approve-billion-settlement-with-facebook-privacy-probe/

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018, March 17). How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

Schroepfer, M. (2018, April 4). An Update on Our Plans to Restrict Data Access on Facebook. Facebook. https://about.fb.com/news/2018/04/restricting-data-access/

CDOJ (2019, March 29). Department of Justice, State of California. California Consumer Privacy Act (CCPA). Office of the Attorney General. https://www.oag.ca.gov/privacy/ccpa

Swanlund, D., & Schuurman, N. (2019). Resisting geosurveillance: A survey of tactics and strategies for spatial privacy. *Progress in Human Geography*, *43*(4), 596–610. https://doi.org/10.1177/0309132518772661

GSA (2022). United States General Services Administration. FPDS-NG ezSearch. Federal Procurement Data System. https://www.fpds.gov/ezsearch/fpdsportal?q=venntel&s=FPDS.GOV&templateName=1.5.2&indexName=awardfull&x=0&y=0

Wolford, B. (2018, November 7). What Is GDPR, the EU's New Data Protection law? GDPR.eu; European Union. https://gdpr.eu/what-is-gdpr/