Simplified Digital Forensics with AI and Cloud

Akin Alkan

Abstract— **Cyber security crimes are increasing and it tends to increase in the following years. Almost every organization effected at least by one successful attack. There is a cyber security gap in all over the world including USA. One way to close this gap is to reduce burden on digital forensics experts. This can be achieved by AI and cloud systems with a simplified GUI. Our model depends on automated data collection, automated data analysis, and automated data report features. A simple GUI can collect the data from different resources, AWS Amazon Rekognition is a good example of image and video recognition cloud tool that helps to analyse the collected data, and to report the results.**

I. INTRODUCTION

A. Why Digital Forensics?

There are different definitions of digital forensics which is also known as computer and network forensics. In general, digital forensics is defined as " application of science to the identification, collection, examination, and analysis of data while preserving the integrity of the information and maintaining a strict chain of custody for the data" [1] The main component of the digital forensics is data and there are different types of data from different sources. Images, text files, videos, location information, chats, network information can be collected from computers, mobile devices, tablets, Internet of Things (IoT) devices, network cards and from different media and storage equipment like cash, memory, USB and disks. This data can be used for different purposes like understanding the root cause of operating systems errors, recovery of the failed systems, data recovery, data acquisition, investigating incidents, and crime investigation.

Every large organization has a complex IT system and they have sensitive data to protect. It is hard to understand what event has occurred in their computer systems and network without a capability of forensics. Depending on the situation, widely investigators, IT professionals, or incident handlers are staffed in organizations to accomplish the above mentioned missions. Sometimes organizations hand over this responsibility to third parties.

B. Necessity of Digital Forensics Teams

There are four main part of the digital forensics process; collection, examination, analysis and reporting. Collection is the beginning phase and includes making decisions about what to collect, labeling them and recording what you have done. Examination is the retrieving and extracting the collected data. Analysis is inferring useful information from examined data. Reporting is the results of all process from collection to analysis and providing recommendations. All these phases take time and have a significant cost. Costs depend on the tools used to collect and examine data like software, hardware, equipment and personnel expenditures like staffing and training. Organizations which has forensics team have an ability to respond the incidents rapidly than the organizations outsource forensics service. Frequency of the incidents, required respond time, sensitivity of the data and the costs are the important factors to make the decision to have a forensics team on board or gain the capability with outsourcing. Whatever the decision is, whole digital forensics process take a significant amount of time and cost, and this is a challenging problem for public and private sectors.

II. ARTIFICIAL INTELLIGENCE AND DIGITAL FORENSICS

Evolution in the internet era and rapid increasing of different IT technologies has changed the world dramatically. Cyber crimes has also increased related to this development. Thousands of different computers, smartphones, tablets make investigation of these equipment extremely difficult and time consuming. There is a huge difference between reported cyber crimes and cases convicted. Lack of digital forensics experts is one of the reason of this huge gap. It takes time for a digital forensics expert to be well trained to deal cases with speed. Besides, the tools used for digital forensic are not fully automated and they need human interaction.

There are lots of work done in improving digital forensics tools and frameworks. Martini and Choo studied on integrated conceptual digital forensic framework for cloud computing [2], Beebe and Clark worked on hierarchical objectives-based framework for the digital investigations process [3], Marziale, Richard and Roussev focused on GPU's to increase the performance of digital forensic tools[4]. These studies heavily focused on to improve the performance and capability of digital forensics tools, but all these tools are still user dependent [5].

AI is an increasing topic for the last two decades. Since the computing power has increased and cloud systems are available everywhere, AI is reachable by everybody and it is also used in digital forensics. The largest cloud provider is AWS, and it has lots of services that can be used by customers.

This paper is about using AWS Rekognition services to reduce burden on digital forensics experts, increase the speed and lower the cost while making the process easy.

A. Amazon Rekognition [6]

Amazon Rekognition automates image and video analysis with Machine Learning and Deep Learning. It offers pre-trained and customizable computer vision (CV) capabilities to extract information and insights from images and videos.

B. Features

Amazon Rekognition has different features to be used for different purposes.

- Content moderation
- Face comparison and search
- Face detection and analysis
- Label detection
- Text detection
- Celebrity recognition
- Video segment detection
- Streaming video events detection

Content Moderation:

Content moderation is used for detecting the inappropriate, unwanted, or offensive content in videos or images. Amazon Rekognition uses a two-level hierarchical taxonomy to label categories of inappropriate or offensive content. Each top-level category has a number of second-level categories.

Top-Level-Category	Second-Level Category
Explicit Nudity	Nudity Graphic Male Nudity Graphic Female Nudity Sexual Activity Illustrated Explicit Nudity Adult Toys
Suggestive	Female Swimwear or Underwear Male Swimwear or Underwear Partial Nudity Barechested Male Revealing Clothes Sexual Situations
Violence	Graphic Violence or Gore Physical Violence Weapon Violence Weapons Self Injury
Visually Disturbing	Emaciated Bodies Corpses Hanging Air Crash Explosions and Blasts
Rude Gestures	Middle Finger
Drugs	Drug Products Drug Use Pills Drug Paraphernalia
Tobacco	Tobacco Products Smoking
Alcohol	Drinking Alcoholic Beverages
Gambling	Gambling
Hate Symbols	Nazi Party White Supremacy Extremist

These labels can be used to detect inappropriate content by determining the minimum confidence level parameter. If any confidence is not specified, Amazon Rekognition takes 50% confidence as standard.

On Amazon Rekognition website it is noted that "Amazon Rekognition isn't an authority on, and doesn't in any way claim to be an exhaustive filter of, inappropriate or offensive content. Additionally, the image and video moderation APIs don't detect whether an image includes illegal content, such as child pornography."

Face Comparison and Search:

Faces in images and videos can be detected by Amazon Rekognition. Service can determine the similarity of a face against another picture or from a private image repository. It can also analyzes the facial attributes of the face, and return a label. Service does not persist actual image bytes. Instead, the underlying detection algorithm first detects the faces in the input image, extracts facial features into a feature vector for each face. Amazon Rekognition uses these feature vectors with the determined similarity threshold when performing face matches. Similarity threshold and similarity parameters can be used to correlate the false positives, and increase the accuracy of the detection algorithm.

With Amazon Rekognition face comparison and search ability, it is easy to index hundreds of thousands mugshots in a couple of days. It also shortens the time to identify the person from days and weeks in traditional method to seconds.

Detecting Labels:

This service detects objects, scenes, activities, landmarks, dominant colors, and image quality in images and videos. It uses a hierarchical taxonomy of ancestor labels to categorize labels. This feature may help use parent labels to build groups of related labels and to allow querying of similar labels in one or more images. For example, it can be easily detected hundreds of common objects, such as guns, masks, yellow cars, and running man.

Custom Labels:

Custom labels detect custom objects using automated machine learning to train the models with as few as 10 images. This service can be used for a more detailed image like a brands, logos, or specific weapons, or unique parts. It extends Amazon Rekognition's existing capabilities, which are already trained on tens of millions of images across many categories. Instead of thousands of images, it can be upload a small set of training images that are specific to the use case. If the images are already labeled, Amazon Rekognition Custom Labels can begin training a model in a short time. If not, images need to be labeled directly within the labeling interface, or Amazon SageMaker Ground Truth can be used to label them. After Amazon Rekognition Custom Labels begins training from the image set, it can produce a custom image analysis model in just a few hours. Behind the scenes, Amazon Rekognition Custom Labels automatically loads and inspects the training data, selects the right machine learning algorithms, trains a model, and provides model performance metrics.

Text Detection:

Amazon Rekognition can detect text in images and videos. It can then convert the detected text into machine-readable text. Both image and video text detection operations support most fonts, including highly stylized ones. After detecting text, Amazon Rekognition creates a representation of detected words and lines of text, shows the relationship between them, and tells you where the text is on an image or video frame. Amazon Rekognition is designed to detect words in English, Arabic, Russian, German, French, Italian, Portuguese and Spanish.

Streaming Video Events Detection: Amazon Rekognition Video detects labels in streaming video. It uses another AWS service Amazon Kinesis Video Streams to receive and process a video stream. People, pets, and packages can be selected as labels. The results will be sent to Amaon S3 bucket and Amazon SNS notifications.

III. BIG PICTURE OF CYBER CRIMES

Cyber crimes are increasing day by day effecting almost everybody lives. It is predicted the damages will cost up to $10.5 trillion annually by 2025. It is not only large companies effected by cyber crimes, smart businesses account for 43% of cyber attacks. Figure 1 shows the distribution of cyber attacks according to the industries.

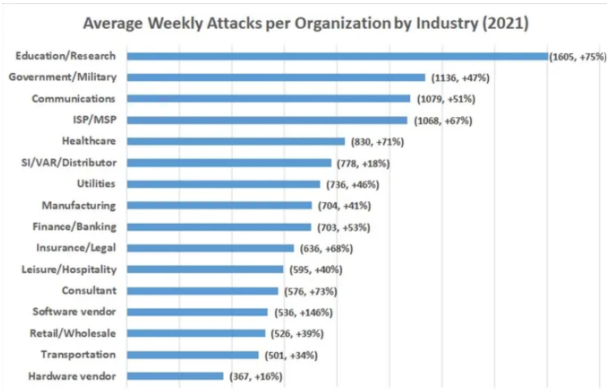

Figure 1 [7]

Number of organizations compromised by an attack is also increasing. Moreover, organizations tend to not to report the incidents due to effects of negative reputation. Figure 2 shows percentage of organizations compromised by at least one successful attack year by year.

Percentage of organizations compromised by at least one successful attack

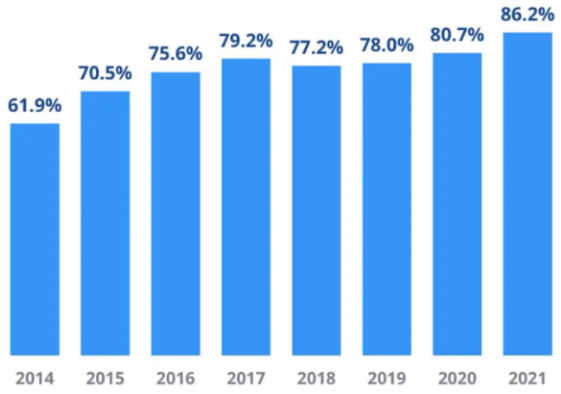

Figure 2 [8]

Despite the increasing threat of cyber security, there is a gap in cyber security demand and the need for workforce will likely to increase. In 2022, the number of active cybersecurity jobs in the United States amounted to approximately 150 thousand. By 2032, the cybersecurity workforce is forecast to reach almost 187 thousand jobs according to Statista 2022 report.

Total number of cybersecurity jobs in the United States from 2020 to 2032

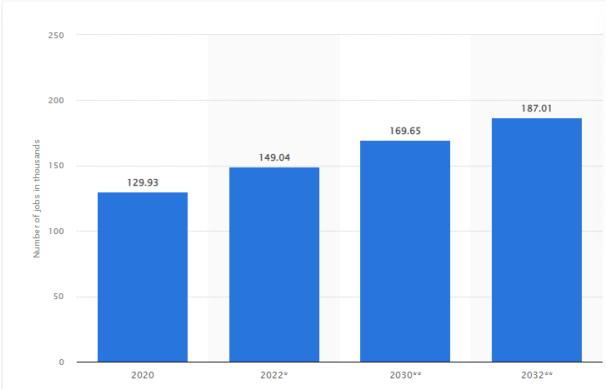

Figure 3 [9]

IV. PROPOSED FRAMEWORK

We propose a conceptual framework in this paper which can reduce the investigation time and cost using AI and cloud infrastructure due to the gap of investigation experts. It is not easy to implement all of the technical details at once, but starting a point will diminish the time and cost of digital investigations.

Main purpose of this framework is to make the whole investigation process, from collection to report, available for anybody who has a few knowledge about computer systems. So, simplicity is the main principal in all framework. Framework consists of 3 main parts: Data acquisition, data analysis, and data reporting.

Data acquisition: Data acquisition is the one of the most important phase, and it is composed of gathering and recovery of data. This phase can be done with a simple GUI for computer forensics, but it may be challenging for mobile

forensics due to various types of mobile phones. Our demo python script scans the folder which the user selected and discovers the image files. This feature can be extended to videos, texts, chat and any kind of file types.

Demo program scans the selected folder

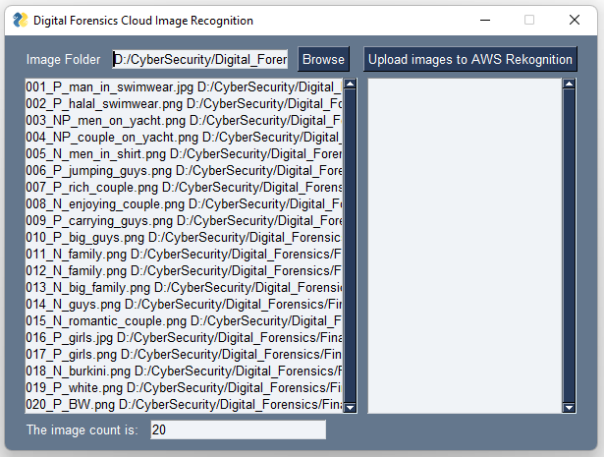

Figure 4

Data analysis: Data analysis is the hard one in terms of handling the collected data. High performance computer systems, and auto-ML models in clouds make it simple than ever. To use these cloud based auto-ML models, nobody needs to know machine learning but how to use them. AWS Rekognition has lots of features from image and video recognition, to labeling and content analysis. To use AWS Rekognition features, one does not need to know machine learning or deep learning, but some programming skills. With our demo program we also bypassed the need of any knowledge about programming. It is simple to send the discovered pictures to AWS S3 storage and analyse their content. Test has been done with free tier AWS account and it took almost 30 seconds to find the 20 images in a folder, load them to S3 storage, analyze the pictures with deep learning and report them back to the GUI. The 20 pictures has been selected from different concepts, and AWS Rekognition reported them with 95% accuracy.

A centralized data center for all these labelled images, videos and texts may contribute to the increase of the accuracy of the detecting and clustering algorithms.

Demo program scans the selected folder

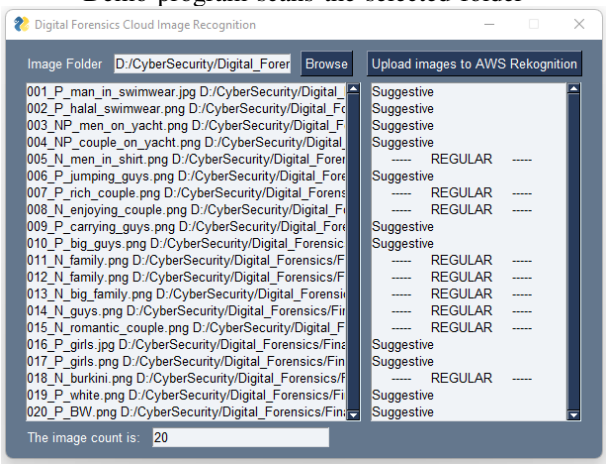

Figure 4

V. CONCLUSIONS

Cyber security crimes are increasing and it tends to increase in the following years. Almost every organization effected at least by one successful attack. There is a cyber security gap in all over the world including USA. One way to close this gap is to reduce burden on digital forensics experts. This can be achieved by AI and cloud systems. AWS Rekognition is a good application for this simplified method. Our simplified GUI shows how a person who has a basic knowledge of computer can use it for basic data acquisition, analysis and report. A centralized data center for all labelled data may increase the accuracy of the detecting and clustering systems, but privacy will be a concern.

REFERENCES

[1] K. Kent, S. Chevalier, T. Grance, and H. Dang, “Guide to integrating forensic techniques into incident response,” NIST Special Publication, pp. 800–86, 2006.

[2] Martini, B., Choo, K. K. R. (2012). An integrated conceptual digital forensic framework for cloud computing. Digital Investigation, 9(2), 71-80.

[3] Beebe, N. L., Clark, J. G. (2005). A hierarchical, objectives-based framework for the digital investigations process. Digital Investigation, 2(2), 147-167.

[4] Marziale, L., Richard, G. G., Roussev, V. (2007). Massive threading: Using GPUs to increase the performance of digital forensics tools. digital investigation, 4, 73-81.

[5] Rughani, H.R., (2017). Artificial Intelligence Based Digital Forensics Framework, 8(8), http://dx.doi.org/10.26483/ijarcs.v8i8.4571.

[6] This section information is gathered from AWS Rekognition webpage, (2022), https://aws.amazon.com/rekognition/

[7] Elliot, M., (2022). Top Security Statistics for 2022, https://www.cobalt.io/blog/top-cybersecurity-statistics-for-2022

[8] Zaharia, A., (2022). 300+ Terrifying Cybercrime and Cybersecurity Statistics, https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends/

[9] Sava, J.A., (2022). Cybersecurity workforce in the United States 2020-2032, https://www.statista.com/statistics/674151/united-states-computer-systems-design-jobs/