IMPROVING THE COLLECTION OF DIGITAL EVIDENCE

BY **MARTIN NOVAK**

Two NIJ-funded projects introduce new methods and tools for collecting and processing digital evidence in cases involving child sexual abuse materials and large-scale computer networks.

Digital evidence can play a critical role in solving crimes and preparing court cases. But often the complexity and sheer volume of evidence found on computers, mobile phones, and other devices can overwhelm investigators from law enforcement agencies.

During an investigation of suspected child sexual abuse materials, for instance, a computer forensic analyst will typically spend hours reviewing hundreds of videos from seized media. The analyst looks at whether a human is present in a particular image. Next, the analyst needs to determine whether the human in the image is an adult or a child. This process is time-consuming, stressful, and prone to error.

This is just one example of the challenges facing law enforcement agencies when it comes to digital evidence. Departments around the country find themselves unable to keep up with rapidly evolving technologies and the quantity of digital evidence they produce. Many departments have limited budgets and lack proper equipment and training opportunities for officers. The result is often large backlogs in analyzing digital evidence.[1]

To help address these challenges and improve the collection and processing of digital evidence, the National Institute of Justice (NIJ) provided funding to Purdue University and the University of Rhode Island. Purdue University created the File Toolkit for Selective Analysis Reconstruction (FileTSAR) for large-scale computer

Both projects help move the field forward with new methods and tools for collecting and processing digital evidence, but these new methods and tools will need to be independently tested and validated.

networks, which enables on-the-scene acquisition of probative data. FileTSAR then allows detailed forensic investigation to occur either on site or in a digital forensic laboratory environment, with the goal of ensuring admissible digital evidence.[2] The University of Rhode Island developed DeepPatrol, a software tool using machine intelligence and deep learning algorithms to assist law enforcement agencies in investigating child sexual abuse materials.

Both of these projects are advancing the field of digital forensics. DeepPatrol may change the way law enforcement conducts forensic examinations by accelerating and streamlining efforts to identify children in videos of sexual exploitation. FileTSAR provides law enforcement with a portable, scalable, cost-efficient tool for examining complex networks.

Automating Image Detection

Automating the process for detecting sexually exploitative images of children would drastically reduce the amount of time that investigators have to spend looking at suspected files and would allow them to concentrate on other aspects of the case. However, poor image quality, image size, and the orientation of the individual in the image present significant challenges to automation. Also, determining whether an unidentified individual in an explicit video is an adult or a child often requires expertise in anthropomorphic indicators of age, knowledge that would be difficult to automate.

One current solution for detecting child sexual abuse in a video involves capturing representative key frame images that the analyst must review manually. Although this is an improvement over having to view an entire video, this method is still time-consuming and may not reduce the analyst's workload.

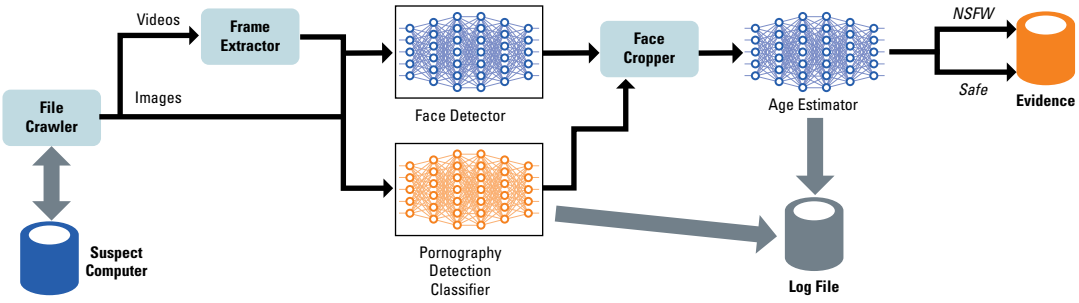
To help address this capability gap, NIJ sought proposals for the development of innovative tools that would automatically detect prepubescent individuals in videos of varying quality. Ideally, the tools would also be able to detect postpubescent individuals who have not yet assumed the full physical characteristics of an adult.

In developing DeepPatrol, researchers from the University of Rhode Island leveraged research in machine intelligence/vision and the implementation of deep learning algorithms and Graphic Processing Unit technology.[3] Instead of relying on expert-designed features, deep learning techniques learn useful feature hierarchies directly from the data, outperforming previous state-of-the-art methods on traditional and complex vision tasks. Media can be processed in real time, including live video, to detect for the presence of child sexual abuse imagery.

As shown in exhibit 1, the file crawler first identifies every image and video file in a given directory, including subdirectories. The frame extractor then separates each video into a sequence of unique frames, which will be processed as images. The extracted images from each video are saved in their own subdirectory. The current video sampling rate for DeepPatrol is one frame per second. In North America, the standard frame rate for video is 30 frames per second. For a two-minute video, then, 3,600 separate images could be extracted.[4]

Next are two steps that use deep learning:[5] the face detector and a pornography detection classifier. The face detector uses the Single Shot Scale-Invariant Face Detector (S^3FD), a publicly available, real-time face detector that uses a single deep neural network with a variety of scales of faces. Neural networks are

Exhibit 1. How DeepPatrol Works



Source: Marco Alvarez Vega, "DeepPatrol: Finding Illicit Videos for Law Enforcement," Final summary report to the National Institute of Justice, award number 2016-MU-CX-K015, April 2020, NCJ 254636, https://www.ojp.gov/pdffiles1/nij/grants/254636.pdf.

computational learning systems that use a network of functions to understand and translate a data input of one form into a desired output, usually in another form. The concept of the artificial neural network was inspired by human biology and the way neurons of the human brain function together to understand inputs from human senses.[6] S^{3}FD is particularly suited for detecting small faces.

Meanwhile, the pornography detection classifier uses OpenNSFW, a publicly available convolutional neural network developed by Yahoo, to detect pornography. A convolutional neural network is a specific type of neural network model designed for working with two-dimensional image data.[7] The pornography detection classifier inputs an image and provides a probability score between zero and one. Scores greater than 0.8 indicate a high probability that the image is pornographic. Scores less than 0.2 indicate that the image is safe. The pornography detection classifier then converts the image to an RGB color format and, using the face cropper, resizes it to 256 by 256 pixels.

The age estimator then takes the output from the face detector, pornography detection classifier, and face cropper to estimate the age of the person in the image. Pornographic images that contain minors are flagged as potentially being child sexual abuse materials. The estimated age and the pornography detection classifier score are sent to a log file.[8]

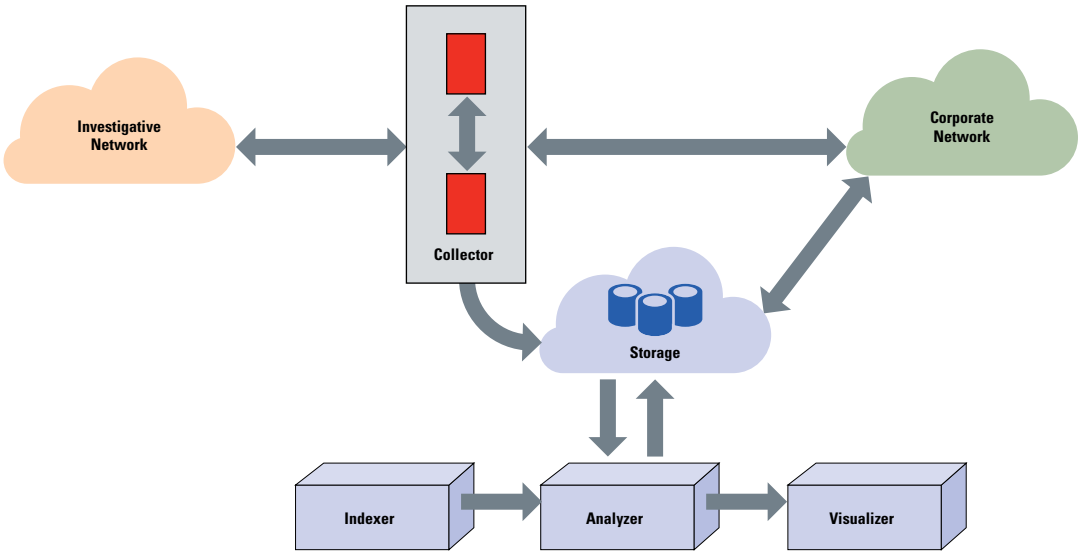
In order for this framework to become a commercially viable computer forensics tool that criminal justice practitioners can use on active cases, additional research will be necessary. For example, the current run time for a case with approximately one million files — including frame extraction, face detection, age estimation, and nudity detection — is 39 hours. This is an intensive resource demand on any agency for any computer forensics process. Reducing the duration of the DeepPatrol process is an essential step in making this platform commercially viable.

The Defense Cybercrime Institute is currently evaluating DeepPatrol using case studies that involve closed investigations with known outcomes to determine whether the algorithms and processes used by DeepPatrol would meet the *Daubert* standard for repeatability, reliability, and acceptance by the scientific community.

Processing Large-Scale Computer Networks

In 2014, an NIJ-funded report by the RAND Corporation listed the lack of tools for examining computer networks as a continuing area of concern for state and local law enforcement in processing digital evidence.[9] Large-scale computer networks — those that contain at least 5,000 devices, including computers, printers, and routers — are often

Exhibit 2. How FileTSAR Works



Source: Based on Kathryn Seigfried-Spellar, "FileTSAR Final Summary Overview," Final summary report to the National Institute of Justice, award number 2016-MU-MU-K091, April 2020, NCJ 254635, https://www.ojp.gov/pdffiles1/nij/grants/254635.pdf, 5.

identified as a potential source of digital evidence in investigations ranging from terrorism to economic crimes.

Digital forensic processing of large-scale computer networks entails some significant challenges when compared to traditional computer forensics. Large-scale computer networks involve diverse configurations, operating systems, applications, connectivity, hardware, and components. In a distributed computing system, data are more volatile and unpredictable than on standalone devices. Applications, resources attached to the network, differing configurations, data storage, or the network topology may obscure information that may be of evidentiary interest. Because networks may be distributed across multiple jurisdictions, only portions or segments of the data may be readily accessible to investigators in the jurisdiction(s) where the crime occurred.

NIJ sought proposals to develop innovative new tools that would allow agencies to conduct digital forensics processing of large-scale computer networks in a forensically sound manner. This included tools capable of reassembling transferred files, searching for keywords, and parsing human communication such as emails or chat sessions from captured network traffic.

With funding from NIJ, Purdue University developed FileTSAR for large-scale computer networks. FileTSAR follows the Computer Forensics Field Triage Process Model,[10] developed by Marcus Rogers and his colleagues, for on-the-scene acquisition of probative data. It then allows detailed forensic investigation to occur either on site or in a digital forensic laboratory environment without affecting the admissibility of evidence gathered via the toolkit.[11]

As shown in exhibit 2, FileTSAR is connected to the large-scale computer network via the collector, which implements two distinct operational components:

a trigger engine and a capture engine. The trigger engine monitors all available network traffic flowing into and out of the network and indicates when specific criteria occur in those network flows.

Based on the criteria for the specific digital forensic investigation, multiple options exist. Those criteria can spawn an event that will initiate the capture engine to record the network data. The capture engine can capture all network traffic (referred to as "catch it as you can") or operate in a variety of selective modes (referred to as "stop, look, listen"). Both the trigger engine and capture engine will output data in an industry-accepted format that is compatible with existing incident response systems, and provide a standardized interface into the storage system and indexer module.

The indexer takes input from the collector and processes it for file contents. The data are archived into the active case directories within the storage subsystem and can be explored, searched, and visualized later. The analyzer identifies the interrelatedness of files, flows, packets, users, and timelines. The analyzer also reconstructs documents, images, email, and Voice over Internet Protocol.

The visualizer identifies trends, patterns, or repetitions. It contains a web-based dashboard, accessible only by authenticated users. This authentication provides system accountability, logs all activities, and maintains the chain of custody for any evidence gathered.

The key to using FileTSAR is its logging capability. This allows an investigator to maintain chain of custody and explain to a jury where the evidence was located and how it was obtained. Another investigator can also replicate FileTSAR's processes on the same evidence.

Purdue University achieved the original goal of FileTSAR — to capture network traffic and restore digital evidence, in its original file format, in large enterprise network settings. This capability, however, requires high-performance storage units and assumes that high-performance servers or workstations will be located on premise within the law enforcement agency.

Licensed versions of FileTSAR are distributed free of charge to law enforcement agencies via a dedicated website.[12] Currently, FileTSAR is licensed to 120 agencies around the world. At least 30 of these agencies have implemented FileTSAR, including the 308th Military Intelligence Battalion, the Nigerian Police, Portugal's Cyber Crime Unit, the Grant County (WI) Sheriff's Office, and the United Kingdom's Royal Navy. Of the remaining 90 licensed agencies, it is uncertain how many have implemented FileTSAR. Component parts for the virtual machines necessary to run the system are made in China and are currently unavailable due to the COVID-19 pandemic.

Although the current version of FileTSAR is ideal for large law enforcement agencies, a more easily deployable, compact version would have greater utility for the 73% of U.S. law enforcement agencies with 25 or fewer sworn officers. With this goal in mind, NIJ recently funded Purdue University's proposal to develop *FileTSAR+ An Elastic Network Forensic Toolkit for Law Enforcement*.[13]

More Testing Is Needed

Both of these projects help move the field forward with new methods and tools for collecting and processing digital evidence. DeepPatrol provides a framework for automating the detection of child sexual abuse in videos. FileTSAR provides law enforcement agencies with the capability to conduct digital forensics processing of large-scale computer networks in a forensically sound manner.

The acceptability of these approaches to the criminal justice community will depend on the admissibility of the evidence each produces. These new methods will need to be independently tested and validated, and subjected to peer review. Any error rates will need to be determined, and standards and protocols will

need to be established. And the relevant scientific community will need to accept the two approaches.

To this end, NIJ plans to have FileTSAR and DeepPatrol independently evaluated by NIJ's Criminal Justice Testing and Evaluation Consortium. This will help ensure that the tools operate in the manner described by the grantees, they can be used for their intended purposes, and — if applicable — they are forensically sound. NIJ expects both of these evaluations to produce reports that will be publicly available once the evaluations are completed.

About the Author

Martin Novak, M.P.A., is a senior computer scientist in NIJ's Office of Research, Evaluation, and Technology.

For More Information

Learn more about NIJ's work in digital evidence and forensics at https://nij.ojp.gov/digital-evidence-and-forensics.

This article discusses the following awards:

- "DeepPatrol: Finding Illicit Videos for Law Enforcement," award number 2016-MU-CX-K015
- "File Toolkit for Selective Analysis & Reconstruction (File TSAR) for Large Scale Computer Networks," award number 2016-MU-MU-K091

Notes

1. All data in this paragraph are from Sean E. Goodison, Robert C. Davis, and Brian A. Jackson, "Digital Evidence and the U.S. Criminal Justice System: Identifying Technology and Other Needs To More Effectively Acquire and Utilize Digital Evidence," RAND Corporation, 2015, https://www.ojp.gov/pdffiles1/nij/grants/248770.pdf.
2. Marcus K. Rogers et al., "Computer Forensics Field Triage Process Model," *Journal of Digital Forensics, Security and Law* 1 no. 2 (2006): 19-37, https://commons.erau.edu/cgi/viewcontent.cgi?article=1004&context=jdfsl.
3. Machine learning is the study of computer algorithms that improve automatically through experience. Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised, or unsupervised (see "What Is Deep Learning?: 3 Things You Need to Know," MathWorks, https://www.mathworks.com/discovery/deep-learning.html). Graphics processing units are specialized electronic circuits designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device (see "What Is a GPU?," intel, https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html).
4. The formula is: 30 frames per second × 60 seconds × 2 minutes = 3,600.
5. Deep learning is a machine learning technique that constructs artificial neural networks to mimic the structure and function of the human brain (see "A Beginner's Guide to Neural Networks and Deep Learning," Pathmind, https://wiki.pathmind.com/neural-network).
6. "What Is a Neural Network?," DeepAI, https://deepai.org/machine-learning-glossary-and-terms/neural-network.
7. Jason Brownlee, "How Do Convolutional Layers Work in Deep Learning Neural Networks?," *Machine Learning Mastery* (blog), April 17, 2020, https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/.
8. For results of the DeepPatrol algorithm training using the APPA-Real dataset, see Jared Rondeau and Marco Alvarez, "Deep Modeling of Human Age Guesses for Apparent Age Estimation," paper presented at the International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 2018.
9. Goodison, Davis, and Jackson, "Digital Evidence and the U.S. Criminal Justice System."
10. The Computer Forensics Field Triage Process Model proposes an onsite or field approach for providing the identification, analysis, and interpretation of digital evidence in a short time frame, without the requirement of having to take the systems/media back to the lab for an in-depth examination or acquiring complete forensic images (see Rogers et al., "Computer Forensics Field Triage Process Model").
11. Rogers et al., "Computer Forensics Field Triage Process Model," 20.

12. "Tools: FileTSAR," Purdue University, https://polytechnic.purdue.edu/facilities/cybersecurity-forensics-lab/tools.

13. National Institute of Justice funding award description, "FileTSAR+ An Elastic Network Forensic Toolkit for Law Enforcement," at Purdue University, award number 2020-DQ-BX-0008, https://nij.ojp.gov/funding/awards/2020-dq-bx-0008.

Image source: digicomphoto/iStock.

NCJ 300985

Cite this article as: Martin Novak, "Improving the Collection of Digital Evidence," *NIJ Journal* 284, December 2022, https://nij.ojp.gov/topics/articles/improving-collection-digital-evidence.