COMPUTING edge

- Virtual and Augmented Reality
- Mobile Computing
- Machine Learning
- Cybersecurity Training

FEBRUARY 2022

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

COMPUTING edge

IEEE COMPUTER SOCIETY computer.org

IEEE COMPUTER SOCIETY
IEEE
Printed with inks containing soy and/or vegetable oils

STAFF

Editor
Cathy Martin

Publications Operations Project Specialist
Christine Anthony

Production & Design Artist
Carmen Flores-Garvey

Publications Portfolio Managers
Carrie Clark, Kimberly Sperka

Publisher
Robin Baldwin

Senior Advertising Coordinator
Debbie Sims

Circulation: *ComputingEdge* (ISSN 2469-7087) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036.

Postmaster: Send address changes to *ComputingEdge*-IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Printed in USA.

Editorial: Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *ComputingEdge* does not necessarily constitute endorsement by the IEEE or the Computer Society. All submissions are subject to editing for style, clarity, and space.

Unsubscribe: If you no longer wish to receive this *ComputingEdge* mailing, please email IEEE Computer Society Customer Service at help@computer.org and type "unsubscribe *ComputingEdge*" in your subject line.

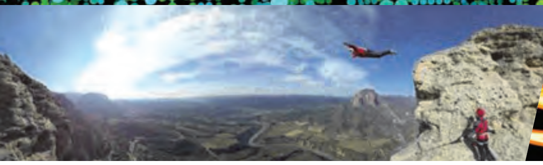
IEEE prohibits discrimination, harassment, and bullying. For more information, visit www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

IEEE Computer Society Magazine Editors in Chief

2469-7087/22 © 2022 IEEE

FEBRUARY 2022 · VOLUME 8 · NUMBER 2

COMPUTING edge



Virtual and Augmented Reality

Mobile Computing

Machine Learning

Cybersecurity Training

Departments

Magazine Roundup

The IEEE Computer Society's lineup of 12 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to visualization and microchip design. Here are highlights from recent issues.

Computer

A Comparative Study of Design Paradigms for PUF-Based Security Protocols for IoT Devices: Current Progress, Challenges, and Future Expectation

Device authentication is an essential security feature for the Internet of Things (IoT). Physical unclonable functions (PUFs) are a promising technology for developing lightweight authentication protocols in these environments. This article from the November 2021 issue of *Computer* presents the current progress and challenges in designing PUF-based authentication protocols for IoT devices.

Computing in Science & Engineering

Performant, Portable, and Productive Parallel Programming With Standard Languages

The perfect solution to the P3 (performance, portability, productivity) problem is a single version of an application that gives high performance across a wide range of target systems and is easy to develop and maintain. Actual solutions give up some level of performance, portability, or productivity, or all three. In this article from the September/October 2021 issue of *Computing in Science & Engineering*, the authors review three periods in the past 65 years when the P3 problem had good solutions. But it is harder today, with greater parallelism. The authors propose a machine model to help programmers design algorithms and data structures that will exhibit performance portability.

IEEE Annals of the History of Computing

Tymshare's Changes to the Project Genie SDS-940 Operating System: A Historical Review

The University of California's Project Genie was an ARPA-sponsored project to create an interactive programming environment. The SDS-930 was a relatively inexpensive machine with a maximum of 32 K of physical memory. To create a time-sharing machine from the base SDS-930 architecture, Pirtle and Lichtenberger extended the '930 architecture to include protection and memory mapping. Deutsch and Lampson implemented an operating system for the modified '930. Ultimately, SDS decided to market the Berkeley modifications to the '930 and denoted it the '940. They sold the machine to a new company called Tymshare. Tymshare heavily modified the Berkeley operating system over more than five years. Read more in this article from the July–September 2021 issue of *IEEE Annals of the History of Computing*.

IEEE Computer Graphics and Applications

Interactive Visualization of Hyperspectral Images Based on Neural Networks

It is challenging to interpret hyperspectral images in an intuitive and meaningful way, as they usually contain hundreds of dimensions. The authors of this article from the September/October 2021 issue of *IEEE Computer Graphics and Applications* develop a visualization tool for hyperspectral images based on neural networks,

which allows a user to specify the regions of interest, select bands of interest, and obtain hyperspectral classification results in a scatterplot generated from hyperspectral features. A cascade neural network is trained to generate a scatterplot that matches the cluster centers labeled by the user. The inferred scatterplot not only shows the clusters of points, but also reveals relationships of substances. The trained neural network can be reused for time-varying hyperspectral data analysis without retraining.

IEEE Intelligent Systems

FedRec: Federated Recommendation With Explicit Feedback

Recommendation models have been widely embedded in various online services, most of which are designed with the assumption that users' original behaviors are available in a central server. This may violate user privacy. This article from the September/October 2021 issue of *IEEE Intelligent Systems* follows a recent work called federated collaborative filtering (FCF) for item recommendation with implicit feedback. The authors propose a novel and generic federated recommendation (FedRec) framework for rating prediction with explicit feedback. Specifically, they federate some basic and advanced factorization-based recommendation models both in batch style and in stochastic style.

IEEE Internet Computing

Revisiting the Arguments for Edge Computing Research

This article from the September/October 2021 issue of *IEEE Internet Computing* argues that low latency, high bandwidth, device proliferation, sustainable digital infrastructure, and data privacy and sovereignty continue to motivate the need for edge computing research even though its initial concepts were formulated more than a decade ago.

IEEE micro

Universal Graph-Based Scheduling for Quantum Systems

High-fidelity operation of a quantum system requires precise tuning of control parameters. Calibration of a quantum system is often achieved by running complex series of dependent experiments, and a full system calibration can require tens of calibration experiments to complete. Optimal control parameters drift over time, and components of experimental quantum systems are susceptible to failure. Hence, continuous operation of a quantum system requires automated background processes such as frequent recalibration and monitoring. In this article from the September/October 2021 issue of *IEEE Micro*, the authors present a scheduling toolkit that schedules experiments based on a directed acyclic graph using a configurable traversal algorithm. The scheduler can be triggered from any process, enabling universal feedback between the scheduler and the quantum control system.

IEEE MultiMedia

From Semantic to Spatial Awareness: Vehicle Reidentification With Multiple Attention Mechanisms

The rapid development and popularization of video surveillance highlight the critical and challenging problem of vehicle reidentification, which suffers from the limited inter-instance discrepancy between different vehicle identities and large intra-instance differences of the same vehicle. In this article from the July–September 2021 issue of *IEEE MultiMedia*, the authors propose a novel multilevel attention network to hierarchically learn an efficient feature embedding for vehicle re-ID. Three kinds of attention are designed in the network: hard local-level attention to localize vehicle salient parts, soft pixel-level attention to refine attended pixels both globally and locally, and spatial attention to enhance the encoder's spatial awareness of salient regions within the windscreen area.

IEEE pervasive COMPUTING
MOBILE SYSTEMS | UBIQUITOUS COMPUTING | INTERNET OF THINGS

Sensing Social Behavior With Smart Trousers

Nonverbal signals play an important role in social interaction. Body orientation, posture, hand movements, and leg movements all contribute to successful communication, though research has typically focused on cues transmitted from the torso alone. The authors of this article from the July–September 2021 issue of *IEEE Pervasive Computing* explore lower-body movements and address two issues: the empirical question of what social signals they provide and the technical question of how these movements could be sensed unintrusively and in situations where traditional methods prove challenging. The authors propose a soft, wearable sensing system for clothing. Bespoke "smart" trousers with embedded textile pressure sensors are designed and deployed in seated, multiparty conversations.

IEEE SECURITY & PRIVACY

Parental Controls: Safer Internet Solutions or New Pitfalls?

Parental-control solutions often require dangerous privileges to function. This article from the November/December 2021 issue of *IEEE Security & Privacy* analyzes privacy and security risks of popular solutions and finds that many leak personal information and are vulnerable to attacks, betraying the trust of parents and children.

IEEE Software

Data-Driven Technical Debt Management: Software Engineering or Data Science Challenge?

In this article from the November/December 2021 issue of *IEEE Software*, the authors summarize experience with data-driven technical debt management that they gained through several industry research projects. They report challenges and their consequences, propose solutions, and sketch improvement directions.

IT Professional

Data Anonymization for Maintenance Knowledge Sharing

Formerly considered part of general enterprise costs, industrial maintenance has become critical for business continuity and a real source of data. Despite the heavy investments made by companies in smart manufacturing, traditional maintenance practices still dominate the industrial landscape. Maintenance knowledge sharing between industries can significantly optimize maintenance activity and improve process efficiency. Different international standards and initiatives are promoting such an approach. However, this trend failed to gain ground in the manufacturing industry. In this article from the September/October 2021 issue of *IT Professional*, the authors present the results of an investigation about the real roadblocks that obstruct the progress of the maintenance knowledge-sharing approach.

Editor's Note

High-Quality Virtual and Augmented Reality

Virtual and augmented reality (VR and AR) have many exciting applications, but the technology faces challenges when it comes to the quality of the user experience. No one wants to use a VR or AR system that lags or makes them queasy. Recently, progress has been made that greatly improves the usability of VR and AR platforms. This *ComputingEdge* issue describes promising VR and AR developments for higher-quality systems.

In "Multimedia in Virtual Reality and Augmented Reality," from *IEEE MultiMedia*, the author highlights the advances in video compression and human–computer interfaces that have led to better VR and AR experiences. In "Saliency Computation for Virtual Cinematography in 360° Videos," from *IEEE Computer Graphics and Applications*, the authors propose a spherical harmonics-based method for efficiently streaming and rendering 360° videos.

Usability is equally important in mobile computing. *IEEE Software*'s "Matt Lacey on Mobile App Usability" discusses how usability affects productivity and morale in both consumer-facing and internal business apps. *IEEE Pervasive Computing*'s "Empowering Communities With a Smartphone-Based Response Network for Opioid Overdoses" gives an example of an easy-to-use mobile app that is making a positive impact in people's lives.

Machine learning (ML) technology faces challenges related to big data and security. The authors of *IEEE Internet Computing*'s "Toward Distributed, Global, Deep Learning Using IoT Devices" present a scalable training system for large Internet of Things datasets. *Computer*'s "Attacking Machine Learning Systems" warns of the vulnerabilities in today's ML systems and urges further research in adversarial ML and ML security.

This *ComputingEdge* issue concludes with perspectives on cybersecurity training. The authors of "Preparing America's Cyber Intelligence Workforce," from *IEEE Security & Privacy*, propose a framework for standardizing cyber-intelligence education and training. In *Computer*'s "Security Awareness Training for the Workforce: Moving Beyond 'Check-the-Box' Compliance," the authors argue that organizations and employees would benefit from more in-depth cybersecurity training.

DEPARTMENT: EIC MESSAGE

This article originally appeared in IEEE MultiMedia vol. 28, no. 2, 2021

Multimedia in Virtual Reality and Augmented Reality

Shu-Ching Chen, *Florida International University, Miami, FL, 33199, USA*

Multimedia is one of the key drivers improving virtual reality and augmented reality (VR/AR), which are promising to reform human–computer interaction in the future with lower-cost and all-in-one headsets containing powerful hardware. Advances in multimedia research on video compression and human–computer interfaces have further enhanced the immersion and efficiency of experiences on the platform. However, many VR/AR experiences are still very difficult to build using traditional engineering methods and many available behavioral and biometric data have not been well explored. Further research in the multimedia community is needed to enhance the usefulness of these systems, with potential in affective learning, resource generation, and developer tools.

Virtual reality and augmented reality (VR/AR) could be considered as one of the key technologies for the next generation of a human–computer interaction. VR/AR headset technology has become powerful enough that many traditional mobile or PC applications can now run on such headsets. Multimedia has played an important role to enable VR/AR technologies. To enable VR/AR devices to display the high-resolution virtual environments to the user seamlessly, immersive content needs to be effectively and efficiently projected and displayed on a virtual three-dimensional spherical surface. Continuous efforts from the multimedia community have led to the development of video-coding techniques, including layered video coding,[1] entropy equilibrium optimization,[2] etc. While the technology is ready for personal use and single-user applications, effective compression, and efficient transmission to enable interconnected VR/AR remains a fundamental challenge and further research is required.

While the users are able to see high-quality videos using VR/AR devices, their behaviors and surrounding environment are captured by the devices using cameras and sensors as well. So, multimedia tools and techniques can be applied to improve user experiences. Specifically, the multimedia data collected by the device can be used to facilitate user's interactions with the VR/AR environment. Advanced deep neural networks have been utilized to analyze the multimedia data collected by VR/AR devices to recognize patterns, such as speech,[3] hand postures, and gestures[4] to interact with the applications, removing the need for conventional controllers. However, the usability and reliability of these techniques to process data remain the key problem to allow for smooth and natural interaction in the VR/AR environment.[5]

VR/AR also provides access to user data that were previously difficult to collect, such as hand pose, head pose, eye-tracking, image, and audio data. These data types can provide great insights into a user's status, which could benefit domains, such as affective learning. Many VR/AR experiences mainly consist of rule-based systems utilizing the immersive nature of the platform, particularly in domains such as education. Such experiences have continuously been found to help improve user engagement,[6] which can be quite challenging using traditional online platforms, especially with hands-on operational work[7] or instructions to younger children such as those with disabilities.[8] However, others have shown that significant numbers of participants in certain VR/AR environments may actually learn less due to the distraction on the platform, even when they are engaged.[9] With the additional data that VR/AR hardware can allow researchers to collect passively, it may be possible to integrate these data into multimodal affective learning systems like those described by Verma *et al.*[10], Tao *et al.*[11] to modify experiences dynamically based on a user's

1070-986X © 2021 IEEE
Digital Object Identifier 10.1109/MMUL.2021.3086275
Date of current version 30 June 2021.

predicted needs and emotional state, keeping user engagement while augmenting learning.

The immersive nature of the platform implies that many assets and details are necessary to construct a realistic environment and a good sense of immersion, leading to enormous time and efforts developing assets in three-dimensional (3-D) space to accompany VR/AR experiences. This problem hinders developing new applications and immigrating existing ones for the VR/AR environment. Research in multimedia could also prove quite useful to mitigate and facilitate content generation for the VR/AR environments. Continued research into converting images into 3-D assets,[12] natural language descriptions into environments,[13] and the generation of audio[14] can all help minimize the work needed to create such environments for any use-case. To ensure that these generated environments are not simply static scenes, more research needs to be done for natural language-based code generation since current methods[15,16] lack those datasets needed to create robust models that could be utilized to generate functionality within the VR/AR environments.

Another challenging problem faced by VR/AR would be integrating the aforementioned multimedia methods and many other tools to allow developers to take advantage of the sensors and data available to them. Platforms such as Unity are extremely useful to create environments using traditional methods of asset creation, as well as providing some infrastructure to integrate the artificial intelligence models directly into projects.[17] However, it can still take a very long time to create VR/AR experiences. By creating frameworks and tools, it could be possible to allow users to generate fully functional environments themselves without the need for extensive computer science knowledge and avoiding lengthy development timelines.

In conclusion, VR/AR is a high-impact platform for the research and development of novel multimedia techniques and shows incredible promise in improving user outcomes in various domains. Advances in compression and interaction technologies as well as the multimedia data that can be collected have greatly improved the usability of the platform for user applications. However, more research needs to be done to take full advantage of the platform. VR/AR has proven to be a modality that can drive improved engagement within users, and further research into directions, such as affective learning, content generation, and development platforms can help fully realize the potential of the platform.

REFERENCES

1. A. T. Nasrabadi, A. Mahzari, J. D. Beshay, and R. Prakash, "Adaptive 360-degree video streaming using layered video coding," in *Proc. IEEE Virtual Reality*, 2017, pp. 347–348, doi: 10.1109/VR.2017.7892319.
2. Y. Zhou, L. Tian, C. Zhu, X. Jin, and Y. Sun, "Video coding optimization for virtual reality 360-degree source," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 118–129, Jan. 2020, doi: 10.1109/JSTSP.2019.2957952.
3. D. Hepperle, Y. Wei, A. Siess, and M. Wölfel, "2D, 3D or speech? A case study on which user interface is preferable for what kind of object interaction in immersive virtual reality," *Comput. Graph.*, vol. 82, pp. 321–331, 2019, doi: 10.1016/j.cag.2019.06.003.
4. K. M. Sagayam and D. J. Hemanth, "Hand posture and gesture recognition techniques for virtual reality applications: A survey," *Virtual Reality*, vol. 21, no. 2, pp. 91–107, 2017, doi: 10.1007/s10055-016-0301-0.
5. H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "GestureGAN for hand gesture-to-gesture translation in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 774–782, doi: 10.1145/3240508.3240704.
6. B. I. Edwards, K. S. Bielawski, R. Prada, and A. D. Cheok, "Haptic virtual reality and immersive learning for enhanced organic chemistry instruction," *Virtual Reality*, vol. 23, no. 4, pp. 363–373, 2019, doi: 10.1007/s10055-018-0345-4.
7. P. Wang, P. Wu, J. Wang, H.-L. Chi, and X. Wang, "A critical review of the use of virtual reality in construction engineering education and training," *Int. J. Environ. Res. Public Health*, vol. 15, no. 6, 2018, Art. no. 1204, doi: 10.3390/ijerph15061204.
8. H. H. S. Ip *et al.*, "Enhance emotional and social adaptation skills for children with autism spectrum disorder: A virtual reality enabled approach," *Comput. Educ.*, vol. 117, pp. 1–15, 2018, doi: 10.1016/j.compedu.2017.09.010.
9. G. Makransky, T. S. Terkildsen, and R. E. Mayer, "Adding immersive virtual reality to a science lab simulation causes more presence but less learning," *Learn. Instruct.*, vol. 60, pp. 225–236, 2019, doi: 10.1016/j.learninstruc.2017.12.007.
10. M. Verma, S. K. Vipparthi, and G. Singh, "AffectiveNet: Affective-motion feature learning for micro expression recognition," *IEEE MultiMedia*, vol. 28, no. 1, pp. 17–27, Jan./Mar. 2020, doi: 10.1109/MMUL.2020.3021659.
11. Y. Tao *et al.*, "Confidence estimation using machine learning in immersive learning environments," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2020, pp. 247–252, doi: 10.1109/MIPR49039.2020.00058.

12. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421, doi: 10.1007/978-3-030-58452-8_24.
13. Q. i. Chen, Q. i. Wu, R. Tang, Y. Wang, S. Wang, and M. Tan, "Intelligent home 3D: Automatic 3D-house design from linguistic descriptions only," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12625–12634, doi: 10.1109/cvpr42600.2020.01264.
14. S.Ö. Arık *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 195–204.
15. B. Wei, G. Li, X. Xia, Z. Fu, and Z. Jin, "Code generation as a dual task of code summarization," in *Proc. Adv. Neural Inf. Process. Syst. Neural Inf. Process. Syst.*, 2019, pp. 1–11.
16. Z. Sun, Q. Zhu, Y. Xiong, Y. Sun, L. Mou, and Lu Zhang, "TreeGen: A tree-based transformer architecture for code generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8984–8991, doi: 10.1609/aaai.v34i05.6430.
17. A. Juliani *et al.*, "Unity: A general platform for intelligent agents," pp. 1–28, 2018, *arXiv:1809.02627*.

SHU-CHING CHEN is currently a Professor with Florida International University, Miami, FL, USA. Contact him at chens@cs.fiu.edu.

EDITOR: Mike Potel, potel@wildcrest.com

DEPARTMENT: APPLICATIONS

This article originally appeared in Computer Graphics and Applications vol. 41, no. 4, 2021

Saliency Computation for Virtual Cinematography in 360° Videos

Ruofei Du and Amitabh Varshney, *Department of Computer Science, University of Maryland, College Park, MD, 20742, USA*

Recent advances in virtual reality cameras have contributed to a phenomenal growth of 360° videos. Estimating regions likely to attract user attention is critical for efficiently streaming and rendering 360° videos. In this article, we present a simple, novel, GPU-driven pipeline for saliency computation and virtual cinematography in 360° videos using spherical harmonics (SH). We efficiently compute the 360° video saliency through the spectral residual of the SH coefficients between multiple bands at over 60FPS for 4K resolution videos. Further, our interactive computation of spherical saliency can be used for saliency-guided virtual cinematography in 360° videos.

With recent advances in consumer-level virtual reality (VR) head-mounted displays (HMD) and 360° cameras, omnidirectional videos are becoming ubiquitous. These 360° videos are becoming a crucial medium for news reports, live concerts, remote education, and social media. One of the most significant benefits of 360° videos is immersion: users have a sustained illusion of presence in such scenes. Nevertheless, despite the rich omnidirectional visual information, most of the content is out of field of view (FoV) of the HMD as well as human eyes. The binocular vision system of human eyes can only interpret 114° FoV horizontally, and 135° FoV vertically. As a result, over 75% of the 360° videos are not being perceived. Furthermore, as shown in Table 1, almost 90% of pixels are beyond the FoV of the current generation of consumer-level VR HMDs.

Therefore, predicting where humans will look, i.e., saliency detection, has great potential over a wide range of applications, such as

- efficiently streaming 360° videos under constrained network conditions;[1]
- salient object detection in panoramic images and videos;[2]
- information overlay in panoramic images, videos, and for augmented reality displays;[3]
- directing the user's viewpoint to salient objects or automatic navigation and synopsis of the 360° videos.[4]

Saliency of regular images and videos has been well studied thoroughly since Itti *et al.*[5] However, unlike classic images which are stored in rectilinear or gnomonic projections, most of the panoramic videos are stored in equirectangular projections. Consequently, classic saliency may not work for 360° videos due to the following challenges (as further shown in Figure 6):

- *Horizontal clipping* may slice a salient object into two parts on the left and right edges, which may cause a false negative result.
- *Spherical rotation* may distort the nonsalient objects near the north and south poles, which may cause a false positive result.

Our work addresses three interrelated questions: a) how should we formulate the saliency in $\mathbb{SO}(2)$ space[a] with spherical harmonics (SH), b) how should we speed up the computation by discarding the low-frequency information, and c) how should we automatically and smoothly navigate 360° videos with saliency maps?

0272-1716 © 2021 IEEE
Digital Object Identifier 10.1109/MCG.2021.3080320
Date of current version 12 July 2021.

[a] $\mathbb{SO}(2)$ space represents all 2-D rotations of the image sphere surrounding the observer.

TABLE 1. Approximate binocular FoV of human eyes, as well as the current generation of consumer-level HMD.

Visual Medium	Approximate Field of View (FoV)		Ratio Beyond FoV
	Horizontal	**Vertical**	
Human Eyes	114°	135°	76.25%
HTC Vive, Oculus Rift	85°	95°	87.53%
Samsung Gear VR	75°	85°	90.16%
Google Cardboard	65°	75°	92.48%

To investigate these questions, we present a novel GPU-driven pipeline for saliency computation and virtual cinematography based on SH, as shown in Figure 1. SH is the spherical analog of a 2-D Fourier transform for planar 2-D images and transforms the 360° images into frequency domain in spherical coordinates.

Our model reveals the multiscale saliency maps in the spherical spectral domain and reduces the computational cost by discarding low bands of SH coefficients. From the experimental results, it outperforms the Itti *et al.* model by over 5× to 13× in timing, and runs in real time at over 60 frames per second for 4K videos on present-day personal computer hardware.

COMPUTING THE SPHERICAL HARMONICS COEFFICIENTS

We begin by preprocessing the SH coefficient for representing the 360° videos. Our pipeline pre-computes a set of the Legendre polynomials and SH functions and stores them in the GPU memory. We adopt the highly parallel prefix sum algorithm to integrate feature maps of the downsampled 360° frames as 15 bands of SH coefficients on the GPU.

Evaluating SH Functions

First, we precompute the SH functions at each spherical coordinate (θ, ϕ) of the input panorama of $N \times M$ pixels. Since the values in the feature maps, which are used to define the intensity and color contrast are positive and real, we compute only the real-valued SH functions $Y(\theta, \phi)$ also known as the tesseral SH, as shown in Figure 2 and detailed by Green[10] and Du.[11]

Evaluating SH Coefficients

Next, we extract the feature maps such as the intensity I, red-green (RG) contrast, and blue-yellow (BY) contrast, inspired by Itti *et al.*'s model[5] and the MATLAB package *SaliencyToolbox* by Walther and Koch[12]

$$RG = \frac{r - g}{\max(r, g, b)}, \quad BY = b - \frac{\min(r, g)}{\max(r, g, b)}. \tag{1}$$

For each feature map, we extract its SH coefficients consisting of L^2 values in L bands. In the equirectangular representation of the 360° videos, we assume that each feature $f_{i,j}$ at the coordinate $(i, j), 0 \le i < N, 0 \le j < M$ represents the mean value $f(\theta_{i+0.5}, \phi_{j+0.5})$ at the solid angle $(\theta_{i+0.5}, \phi_{j+0.5})$, where θ_i and ϕ_j represent equally spaced spherical coordinates. Therefore, for the mth element of a specific band l, we evaluate the SH coefficients of the feature map f as

$$\begin{aligned} c_l^m(\theta, \phi) &= \int_{(\theta,\phi)\in S} f(\theta, \phi) \cdot Y_l^m(\theta, \phi) \sin\theta \, d\theta \, d\phi \\ &= \frac{2\pi}{M} \sum_{i=1}^{N} \sum_{j=1}^{M} f_{i,j} \cdot Y_l^m(\theta_{i+0.5}, \phi_{j+0.5}) \\ &\quad |\cos\theta_{i+1} - \cos\theta_i|. \end{aligned} \tag{2}$$

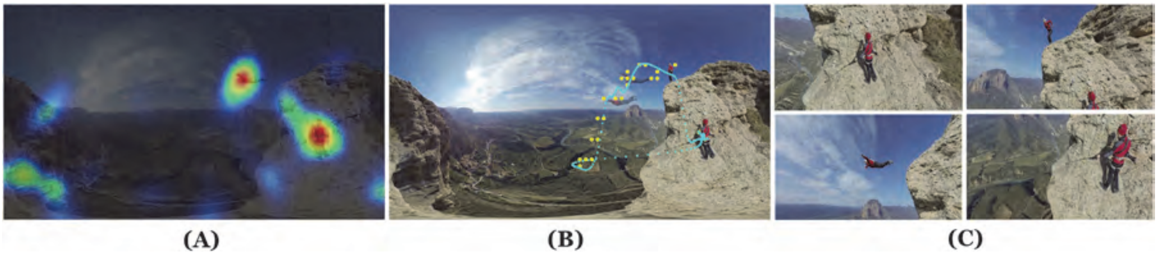


FIGURE 1. (A) shows the saliency map computed by our SSR model in 21.34 ms on a CPU and 10.81 ms on a GPU. (B) shows the optimized camera trajectory in blue. (C) shows four rendering results of virtual cinematography using our shader opensourced at https://shadertoy.com/view/ldBczm.

RELATED WORK

Our work builds upon a rich literature of prior art on saliency detection and spherical harmonics.

Visual Saliency

A region is considered salient if it has perceptual differences from the surrounding areas that are likely to draw visual attention. Prior research has designed bottom-up,[5] top-down,[6] and hybrid models for constructing a saliency map of images (see the review by Zhao *et al.*[7]). The bottom-up models combine low-level image features from multiscale Gaussian pyramids or Fourier spectrum. Top-down models usually use machine learning strategies and take advantage of higher level knowledge such as context or specific tasks for the saliency detection. Recently, hybrid models using convolutional neural networks[8] have emerged to improve the saliency prediction.

One of the most pivotal algorithms for saliency detection remains the Itti *et al.* model.[5] This model computes the center-surround differences of multilevel Gaussian pyramids of the feature maps, which include intensity, color contrast, and orientations, as conspicuity maps. It further combines the conspicuity maps with nonlinear combination methods and a winner-take-all network. Another influential algorithm is the spectral residual approach,[9] which computes the visual saliency by the difference of the original and smoothed log-Fourier spectrum of the image.

However, these and other image saliency approaches assume the input data as rectilinear images, which would not output consistent results for spherical images with horizontal clipping or spherical rotation. Inspired by the Itti *et al.* model[5] and the spectral residual approach,[9] we formulate a SSR model in $\mathbb{SO}(2)$ space. Our model achieves spherical consistency and can be applied to real-time virtual cinematography of 360° videos.

Spherical Harmonics

Spherical harmonics are a complete set of orthogonal functions on the sphere (see Figure 2), and can be used to represent functions defined on the surface of a sphere.[10] In visual computing, spherical harmonics have been widely applied to various domains and applications including, indirect lighting, volume rendering, spatial sound, and 3-D object retrieval. In this article, we use spherical harmonics to efficiently evaluate visual saliency by the difference of the high-frequency and low-frequency spectrum.

Let

$$H_{i,j} = \frac{2\pi}{M} Y_l^m(\theta_{i+0.5}, \phi_{j+0.5}) |\cos\theta_{i+1} - \cos\theta_i| \tag{3}$$

we have

$$c_l^m(\theta, \phi) = \sum_{i=1}^{N} \sum_{j=1}^{M} f_{i,j} \cdot H_{i,j}. \tag{4}$$

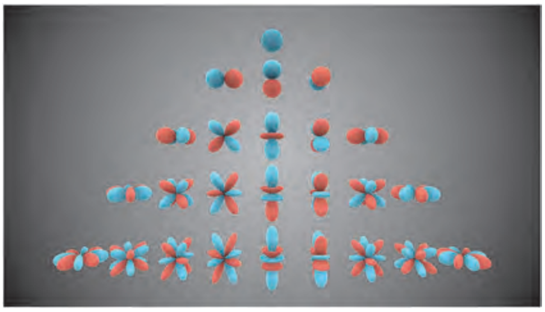


FIGURE 2. The first five bands of SH functions. Blue and red indicate positive and negative values respectively. See https://shadertoy.com/view/4dsyW8 for a live demo built upon Íñigo Quílez's work.

Hence, for a given dimension of the input frames, we can precompute the terms $H(i, j)$ and store them in a lookup table. The integration of the SH coefficients is then reduced to a conventional prefix sum problem.

Implementation Details

On the CPU-driven pipeline, we use *OpenMP* to accelerate the evaluation of SH coefficients with 12 threads. On the GPU-driven pipeline, we take advantage of the *Blelloch Scan* algorithm[13] with *CUDA* to efficiently aggregate the SH coefficients with 2048 kernels on an NVIDIA GTX 1080. The *Blelloch Scan* algorithm computes the cumulative sum in $O(\log N)$ for N numbers. Therefore, our algorithm runs at $O(L^2 \log MN)$ for L^2 coefficients.

Finally, we show the reconstructed image f' with 1–15 frequency bands of SH coefficients with regular RGB color maps in Figure 3.

Note that the low-band SH coefficients capture the background information, such as sky and mountains, while the high-band SH coefficients capture the details, such as parachuters.

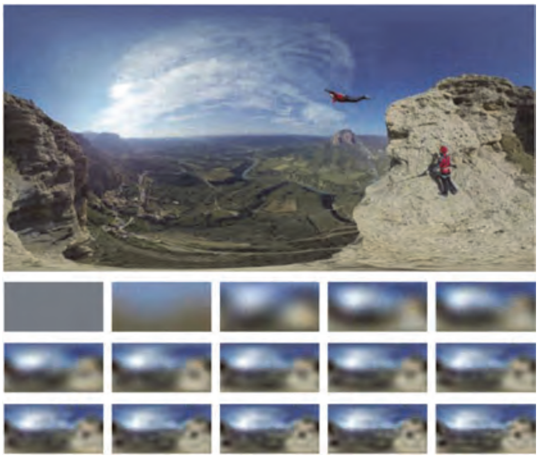


FIGURE 3. Reconstructed images using the first 15 frequency bands of SH coefficients extracted from the video frame.

SSR MODEL

Inspired by the spectral residual approach,[9] we define SSR as the accumulation of the SH coefficients between a low-frequency band and a high-frequency band. This model reveals the multiscale saliency maps in the spherical spectral domain and reduces the computational cost by discarding the low bands of SH coefficients.

SSR Approach

As shown in Figure 3, SH frequency bands can be used to compute the contrast directly across multiple scales in the frequency space. We define the SSR as the sum of the frequency bands between P and Q

$$\Re(\theta,\phi) \quad = \sum_{l=P+1}^{Q} \sum_{m=-l}^{l} c_l^m \cdot Y_l^m(\theta,\phi). \tag{5}$$

Here, $Y_l^m(\phi,\theta)$ are precomputed associated Legendre polynomials in the preprocessing stage. The SSR represents the salient part of the scene in the spectral domain and serves as a compressed representation using SH.

For better visual effects, we square the spectral residual to reduce estimation errors and smooth the spherical saliency maps using a Gaussian

$$\mathbf{S}(\theta,\phi) = \mathfrak{G}(\sigma) * [\Re(\theta,\phi)]^2 \tag{6}$$

where $\mathfrak{G}(\sigma)$ is a Gaussian filter with standard deviation σ (we empirically take $\sigma = 5$).

We show the SSR results of the intensity channel with all pairs of the lower band P and the higher band Q in Figure 4. As P increases, the low-frequency information such as the sky and mountains are filtered out. The SSR results within and close to the orange bounding box reveal the salient objects, such as the two people.

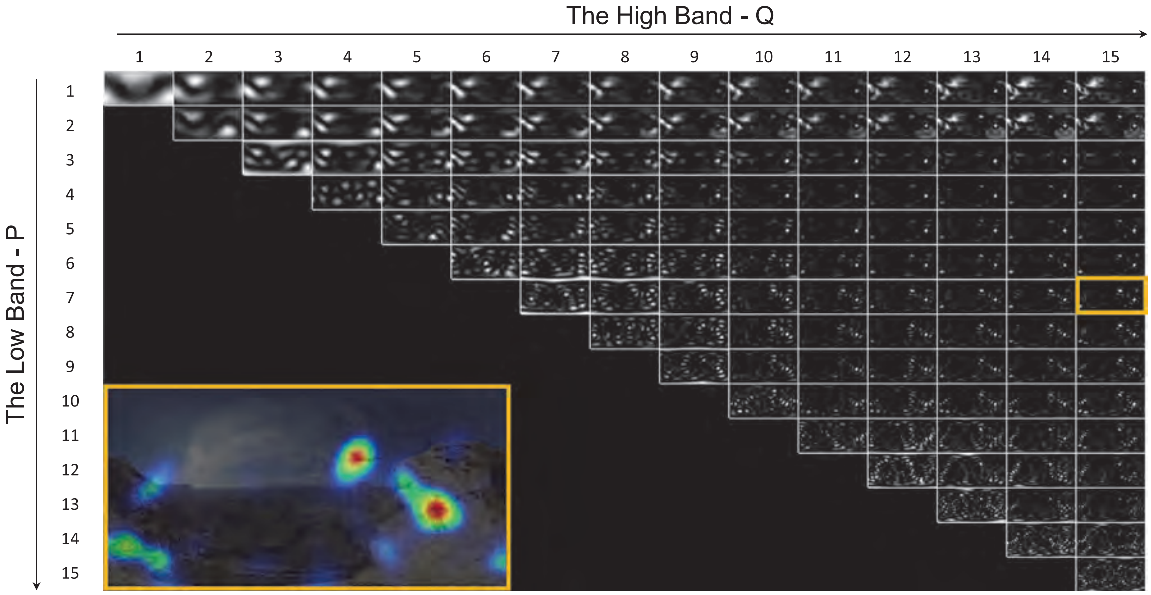


FIGURE 4. Spectral residual maps between different frequency bands of SH. The number along the horizontal axis indicates the high band Q, while the vertical axis indicates the low band P. Note that the saliency maps within or close to the orange box successfully detect the two people in the frame (lower left).

Temporal Saliency

In addition to intensity and color features, we further extract temporal saliency in the SH domain.

For the SH coefficients extracted from the three feature maps, we maintain two sliding temporal windows to estimate temporal contrast. The smaller window w_0 stores the more recent SH coefficients from the feature maps, and the larger window w_1 stores the SH coefficients over a longer term. For each frame, we calculate the estimated SH coefficients $\bar{c}_l^m, \bar{\bar{c}}_l^m$ from both windows, using two probability density functions from the Gaussian distribution ($|w_0| = 5, |w_1| = 25, \sigma = 7.0$). We use a formulation similar to the spatial saliency to measure the SSR between two temporal windows

$$\mathfrak{R}(F_{\text{temporal}}, \theta, \phi) = \left| \sum_{l=P+1}^{Q} \sum_{m=-l}^{l} \left(\bar{\bar{c}}_l^m(\theta,\phi) - \bar{c}_l^m(\theta,\phi)\right) \cdot Y_l^m(\theta,\phi) \right|. \tag{7}$$

We again apply (6) to compute the smoothed temporal saliency maps.

Saliency Maps With Nonlinear Normalization

Following Itti *et al.*,[5] we apply the nonlinear normalization operator $\mathcal{N}(\cdot)$ to all six saliency maps: intensity, red-green, and blue-yellow contrasts, both statically and temporally. This operator globally promotes maps, which contain a few peak responses and suppresses maps with a large number of peaks. After the nonlinear normalization, we linearly combine all saliency maps into the final saliency map

$$\mathcal{S} = \tfrac{1}{N} \bigoplus_{i=1}^{N} \mathcal{N}(\mathbf{S}(F_i)). \tag{8}$$

Empirically, we choose $Q = 15, P = 7$. The final composed result is shown at the bottom left corner in Figure 4, as well as in the accompanying video: https://youtu.be/jwv5hjlg-MI.

Comparison Between Itti *et al.* and SSR Model

As shown in Figure 6, our SSR model is visually better than the Itti *et al.* model. In addition, our experimental results below compare the classic Itti *et al.* model and our model.

We use six videos from the Insta360[b] and the 360Rize.[c] The video resolutions vary from 1920 × 1080 to 7680 × 3840 pixels.

TABLE 2. Timing comparison between the Itti *et al.* model and our SSR model.

Resolution	Average Timing Per Frame		
	Itti *et al.* (CPU)	SSR (CPU)	SSR (GPU)
1920×1080	104.46 ms	21.34 ms	10.81 ms
4096×2048	314.94 ms	48.18 ms	13.20 ms
7680×3840	934.26 ms	69.53 ms	26.58 ms

Our results are obtained on a workstation with an NVIDIA GTX 1080 and an Intel Xeon E5-2667 2.90 GHz CPU with 32 GB RAM. Both the Itti *et al.* model and the SSR model are implemented in C++ and OpenCV. The GPU version of the SSR model is developed using CUDA 8.0. We measure the average timing of saliency computation as well as the visual results between the Itti *et al.* model and our SSR model. Note that the timings do not include the uploading time for each frame from system memory to GPU memory. We expect our algorithms would map well to products such as NVIDIA DrivePX[d] in which videos are directly loaded onto the GPU memory.

We measure the average computational cost of the initial 600 frames across three resolutions: 1920 × 1080, 4096 × 2048, and 7680 × 3840, as shown in Table 2. All frames are preloaded into the CPU memory to eliminate the I/O overhead. Both the CPU and GPU versions of our SSR model outperform the classic Itti *et al.* model, with the speedups ranging from 4.8× to 13.4 × , depending on various resolutions. We show example input and the output from both models in Figure 5.

SALIENCY-GUIDED VIRTUAL CINEMATOGRAPHY

We now present a saliency-guided virtual cinematography system for navigating 360° videos. Inspired by prior art on camera path selection and interpolation,[4,14] we formulate a spatiotemporal model to ensure large saliency coverage while reducing the camera movement jitter.

We compute our saliency maps by linearly combining the saliency maps based on intensity, color, and motion, and then performing a nonlinear normalization, as explained in the previous section. However, for 360° videos, the most salient objects may vary from frame to frame, due to the varying

[b] Insta360: https://www.insta360.com
[c] 360 Rize: http://www.360rize.com

[d] https://NVIDIA.com/en-us/self-driving-cars/drive-px

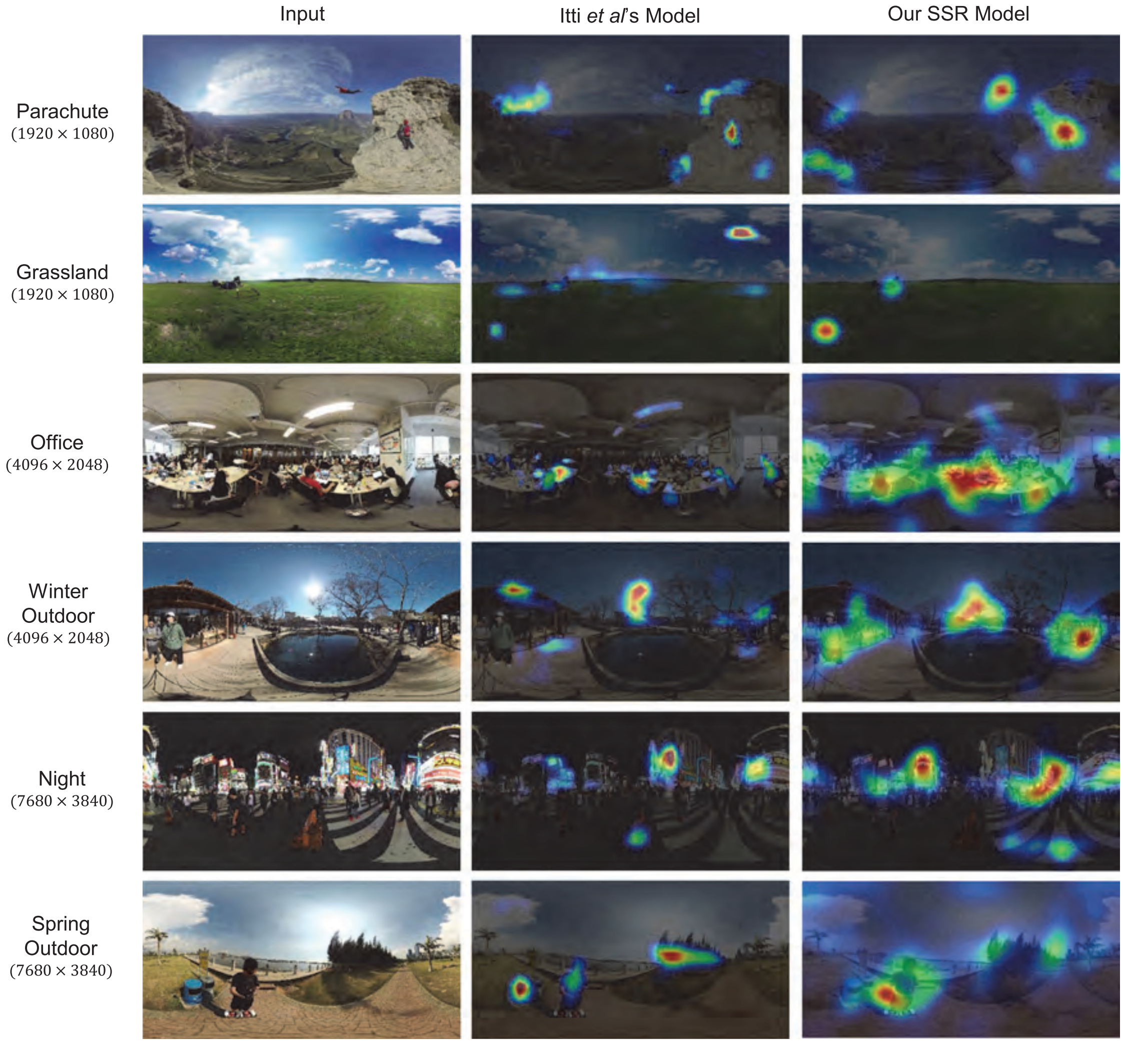


FIGURE 5. Visual comparison between the Itti *et al.* model and our SSR model. Note that while the results are visually similar, our SSR model is 5× to 13× faster than the Itti *et al.* model.

occlusions, colors, and self-movement. As a result, an approach that relies on just tracking the most salient objects may incur rapid motion of the virtual camera, and worse still, may induce motion sickness in virtual reality. Hence, we devise a spatio-temporal optimization model of the virtual camera's discrete control points and further employ a spline interpolation among the control points to achieve smooth camera navigation.

Optimization of the Camera's Control Points

To estimate the virtual camera's control points, we formulate an energy function $\mathbf{E}(C)$ in terms of camera location $C = (\theta, \phi)$. The energy function

$$\mathbf{E}(C) = \lambda \mathbf{E}_{\text{saliency}}(C) + \mathbf{E}_{\text{temporal}}(C) \tag{9}$$

consists of a saliency coverage term $\mathbf{E}_{\text{saliency}}$ and a temporal motion term $\mathbf{E}_{\text{temporal}}$, thus taking both saliency coverage and temporal smoothness into consideration. Empirically, we assign $\lambda = 2$.

Saliency Coverage Term

This spatial term $\mathbf{E}_{\text{saliency}}$ penalizes the coverage of the saliency values beyond the FoV. As for a specific virtual camera location C, this term would be written as

$$\mathbf{E}_{\text{saliency}}(C) = \frac{\sum_{\theta,\phi} \mathbf{S}(\theta, \phi) \cdot \mathbf{O}(C, \theta, \phi)}{\sum_{\theta,\phi} \mathbf{S}(\theta, \phi)} \tag{10}$$

where $\mathbf{O}(C, \phi, \theta)$ indicates whether an arbitrary spherical point (ϕ, θ) is observed by the virtual camera

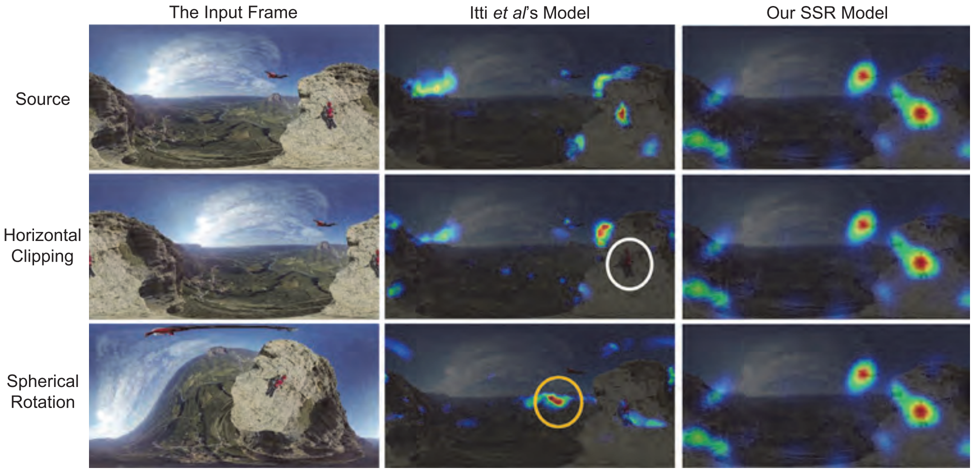


FIGURE 6. Comparison between the Itti *et al.* model and our SSR model with horizontal translation and spherical rotation in the 360° video frame. White circles indicate the false negative result from Saliency Toolbox and orange ones indicate false positive result from Saliency Toolbox. Meanwhile, the results from our SSR model remain consistent, regardless of horizontal clipping and spherical rotation.

centered at the location C_i

$$\mathbf{O}(C,\theta,\phi)=\begin{cases}1, & (\theta,\phi) \text{ is observed by camera at } C\\ 0, & \text{otherwise.}\end{cases} \tag{11}$$

Thus, $\mathbf{E}_{\text{saliency}}(C)$ measures the coverage of the saliency values beyond the FoV of the virtual camera centered at C. To reduce the computation, we compute the saliency coverage term over 2048 points (θ,ϕ), which are uniformly distributed over the sphere.

Temporal Motion Term

For the ith frame in the sequence of the discrete control points, $\mathbf{E}_{\text{temporal}}(C)$ measures the temporal motion of the virtual camera as follows:

$$\mathbf{E}_{\text{temporal}}(C)=\begin{cases}\|C_{i-1},C_i\|_2, & i\geq 1\\ 0 & ,i=0.\end{cases} \tag{12}$$

The Optimization Process

Based on this spatiotemporal model, we evaluate the energy functions over 32×64 pairs of discrete (θ,ϕ). This process is highly parallel, and can be efficiently implemented on the GPU. For each frame, we compute the optimal camera point as follows:

$$C=\underset{C}{\text{argmin}}\,\mathbf{E}(C). \tag{13}$$

In this way, we extract a subsequence of discrete spherical coordinates $Seq=\{C_i|C_i=(\phi_i,\theta_i)\}$ of the optimal camera location in the saliency maps every K frames, $K=5$ in our examples. Since these locations are discrete and sampled at a lower frame rate, we further perform spline interpolation with C^2 continuity.

Interpolation of Quaternions

To interpolate between 3-D rotations of the surrounding sphere over time through our calculated salient positions, we convert the spherical coordinates to quaternions:

$$Q(\theta,\phi)=(0,\ \sin(\theta)\cos(\phi),\ \sin(\theta)\sin(\phi),\ \cos(\theta)). \tag{14}$$

We then use spherical spline curves with C^2 continuity to compute the smooth trajectory of the camera

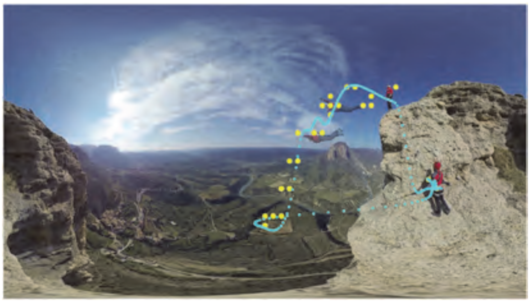


FIGURE 7. Interpolation among the global maximas of the saliency maps in the spherical space. The yellow dots show the discrete optimal locations using the energy function, and the blue dots show the interpolation using the spherical spline curve.

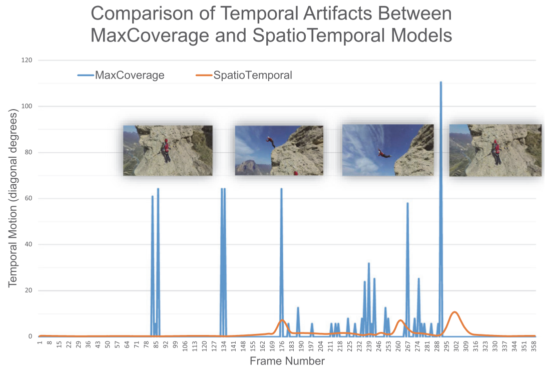


FIGURE 8. Quantitative comparison between the MaxCoverage model and the SpatioTemporal Optimization (STO) model. Compared with the MaxCoverage model, the STO model significantly reduces the temporal jitters.

cruise path over the quaternions. Figure 7 shows the locations of the global maximas, as well as the interpolated spline path over the sphere. An alternative method is to use spherical linear interpolation[e] for the interpolation.

Evaluation of Virtual Cinematography Models

We compare our method, denoted as SpatioTemporal Model (STO) with a MaxCoverage model, which determines the camera position for the maximal coverage of the saliency map. We evaluate the temporal motion terms for the same video sequence and plot the data in Figure 8.

From the quantitative evaluation, as well as the accompanying https://youtu.be/jwv5hjlg-MIvideo, we have validated that the STO model reduces the temporal jittering of the camera motion compared to MaxCoverage model for virtual cinematography in 360° videos.

FUTURE DIRECTIONS

With new datasets of stereoscopic 360° videos and eye-tracking data, one may extend our model and optimize the foveated streaming[1] and automatic camera navigation. We believe our spherical representation of saliency maps will inspire more research to think out of the rectilinear space. We envision our techniques will be widely used for live streaming of events, video surveillance of public areas,[11] as well as templates for directing the camera path for immersive storytelling. Future research may explore how to naturally place 3-D objects with SH irradiance in 360° videos, how to employ SH for foveated rendering in 360° videos, and the potential of compressing and streaming 360° videos with SH.

[e] https://boost.org/doc/libs/1_67_0/libs/qvm/doc/slerp.html

REFERENCES

1. D. Li, R. Du, A. Babu, C. D. Brumar, and A. Varshney, "A log-rectilinear transformation for foveated 360-degree video streaming," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 5, pp. 2638–2647, May 2021.
2. V. Sitzmann *et al.*, "Saliency in VR: How do people explore virtual environments?," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1633–1642, Apr. 2018.
3. R. Du *et al.*, "Geollery: A. mixed reality social media platform," in *Proc. CHI Conf. Human Factors Comput. Syst.*, no. 685, 2019, pp. 1–13.
4. Y.-C. Su *et al.*, "Pano2Vid: Automatic cinematography for watching 360 videos," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 154–171.
5. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
6. S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.
7. Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Process.*, vol. 93, no. 6, pp. 1401–1407, 2013.
8. G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5455–5463.
9. X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
10. R. Green, "Spherical harmonic lighting: The gritty details," in *Proc. Archives Game Developers Conf.*, vol. 56, 2003, Art. no. 47.
11. R. Du, "Fusing multimedia data into dynamic virtual environments," Ph.D. dissertation, Univ. Maryland, College Park., MD, USA, Nov. 2018.
12. D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.
13. G. E. Blelloch, "Scans as primitive parallel operations," *IEEE Trans. Comput.*, vol. 38, no. 11, pp. 1526–1538, Nov. 1989.
14. T. Oskam *et al.* "OSCAM-optimized stereoscopic camera control for interactive 3D," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–8, 2011.

RUOFEI DU is currently a Senior Research Scientist at Google. He is the corresponding author of this article. Contact him at me@duruofei.com.

AMITABH VARSHNEY is the Dean of the College of Computer, Mathematical and Natural Sciences, and a Professor of Computer Science, University of Maryland at College Park. Contact him at varshney@cs.umd.edu.

Contact department editor Mike Potel at potel@wildcrest.com.

IEEE COMPUTER SOCIETY

PURPOSE: The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.

MEMBERSHIP: Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

COMPUTER SOCIETY WEBSITE: www.computer.org

OMBUDSMAN: Direct unresolved complaints to ombudsman@computer.org.

CHAPTERS: Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

AVAILABLE INFORMATION: To check membership status, report an address change, or obtain more information on any of the following, email Customer Service at help@computer.org or call +1 714 821 8380 (international) or our toll-free number, +1 800 272 6657 (US):

- Membership applications
- Publications catalog
- Draft standards and order forms
- Technical committee list
- Technical committee application
- Chapter start-up procedures
- Student scholarship information
- Volunteer leaders/staff directory
- IEEE senior member grade application (requires 10 years practice and significant performance in five of those 10)

PUBLICATIONS AND ACTIVITIES

***Computer*:** The flagship publication of the IEEE Computer Society, *Computer* publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.

Periodicals: The society publishes 12 magazines and 17 journals. Refer to membership application or request information as noted above.

Conference Proceedings & Books: Conference Publishing Services publishes more than 275 titles every year.

Standards Working Groups: More than 150 groups produce IEEE standards used throughout the world.

Technical Committees: TCs provide professional interaction in more than 30 technical areas and directly influence computer engineering conferences and publications.

Conferences/Education: The society holds about 200 conferences each year and sponsors many educational activities, including computing science accreditation.

Certifications: The society offers three software developer credentials. For more information, visit www.computer.org/certification.

BOARD OF GOVERNORS MEETING

1-3 February 2022

IEEE

revised 16 December 2021

EXECUTIVE COMMITTEE

President: William D. Gropp
President-Elect: Nita Patel
Past President: Forrest Shull
First VP: Riccardo Mariani; **Second VP:** David S. Ebert
Secretary: Jyotika Athavale; **Treasurer:** Michela Taufer
VP, Membership & Geographic Activities: Andre Oboler
VP, Professional & Educational Activities: Hironori Washizaki
VP, Publications: David S. Ebert
VP, Standards Activities: Annette Reilly
VP, Technical & Conference Activities: Grace Lewis
2021–2022 IEEE Division VIII Director: Christina M. Schober
2022-2023 IEEE Division V Director: Cecilia Metra
2022IEEE Division VIII Director-Elect: Leila De Floriani

BOARD OF GOVERNORS

Term Expiring 2022: Nils Aschenbruck, Ernesto Cuadros-Vargas, David S. Ebert, Grace Lewis, Hironori Washizaki, Stefano Zanero
Term Expiring 2023: Jyotika Athavale, Terry Benzel, Takako Hashimoto, Irene Pazos Viana, Annette Reilly, Deborah Silver
Term Expiring 2024: Saurabh Bagchi, Charles (Chuck) Hansen, Carlos E. Jimenez-Gomez, Daniel S. Katz, Shixia Liu, Cyril Onwubiko

EXECUTIVE STAFF

Executive Director: Melissa A. Russell
Director, Governance & Associate Executive Director: Anne Marie Kelly
Director, Conference Operations: Silvia Ceballos
Director, Finance & Accounting: Sunny Hwang
Director, Information Technology & Services: Sumit Kacker
Director, Marketing & Sales: Michelle Tubb
Director, Membership & Education: Eric Berkowitz
Director, Periodicals & Special Projects: Robin Baldwin

COMPUTER SOCIETY OFFICES

Washington, D.C.: 2001 L St., Ste. 700, Washington, D.C. 20036-4928; **Phone:** +1 202 371 0101; **Fax:** +1 202 728 9614; **Email:** help@computer.org
Los Alamitos: 10662 Los Vaqueros Cir., Los Alamitos, CA 90720; **Phone:** +1 714 821 8380; **Email:** help@computer.org

MEMBERSHIP & PUBLICATION ORDERS

Phone: +1 800 678 4333; Fax: +1 714 821 4641; Email: help@computer.org

IEEE BOARD OF DIRECTORS

President: K.J. Ray Liu
President-Elect: TBD
Past President: Susan K. "Kathy" Land
Secretary: TBD
Treasurer: TBD
Director & President, IEEE-USA: TBD; **Director & President, Standards Association:** TBD; **Director & VP, Educational Activities:** TBD; **Director & VP, Membership & Geographic Activities:**TBD; **Director & VP, Publication Services & Products:** TBD; **Director & VP, Technical Activities:** TBD

EDITOR: Robert Blumen, SalesForce, robert@robertblumen.com

This article originally appeared in
Software
vol. 38, no. 2, 2021

DEPARTMENT: SOFTWARE ENGINEERING RADIO

Matt Lacey on Mobile App Usability

Gavin Henry

FROM THE EDITOR

Matt Lacey, author of *Usability Matters*, discusses usability for consumers and business or in-house users. Host Gavin Henry spoke with Lacey about the six components of great app experiences, things every app should do, native apps, password managers, accessibility, feedback, telemetry, locations, non-mobile devices, examples of good and bad apps, testing, connectivity, user involvement during development, and usability and software engineering. We provide summary excerpts below; to hear the full interview, visit http://www.se-radio.net or access our archives via RSS at http://feeds.feedburner.com/se-radio.—*Robert Blumen*

Gavin Henry: What is usability?

Matt Lacey: It means software is simple and does everything the person using it needs it to do without trouble.

Why is usability important?

For a consumer app, customers have alternatives. For a business app used internally, the experience of using it affects productivity and how people feel about their work. More attention is usually paid to consumer-facing apps than to internal apps. But internal software affects productivity, morale, and the ability to work remotely.

Is usability the same as an experience?

An experience is what it is like to use the app. That's usability—how does the person using it feel, and how productive are they? Can they do what they want and achieve their goals? An app can be perfectly functional but horrible to use. Functionality is not the same as usability.

What are the six components of great app experiences covered in your book?

Context, input, output, responsiveness, connectivity, and resources. These are things beyond code that are important and easy to overlook.

Context is understanding where the app is used in addition to how it actually works: who uses it, what they want to do with it, how they use it, what devices they use it on, where they are located, what languages they speak, and so on.

Input is how data get into the mobile app, both from the person using it and from other sources. Information from sensors on a device or from a remote source

Digital Object Identifier 10.1109/MS.2020.3042424
Date of current version: 11 February 2021

is input in addition to what gets typed in through the keyboard, and you need to think about it.

Output is what I put on the screen and more, such as sound and sending emails. And where will those emails be received? On the same device as the app? I've seen apps where you sign up through the app and it sends you an email, and then you can't view the app and the email easily at the same time on the device.

The fourth component is *responsiveness*. How do we respond, or how is our response perceived? Do users get a quick response when they click a button? Did the mobile app immediately start doing something, or did it sit and spin and make the user wonder if it was all okay? How does the user feel about the speed of the response? You have to make sure that things happen fast enough for the person using it. It comes back to context.

Must the developer know the context up front?

I always like to know context, to be sure I'm building something that's appropriate for users. We can measure, test, and ask for feedback during development. I do user-based testing, getting software in front of real users as soon as possible. The feedback I need varies depending on the application, but I make sure that people can provide feedback and that the software is being used by users in the real world, as much and as soon as possible. In those early stages, there can't be too much feedback before you get to the App Store. Once you go public, there will be a lot of new feedback coming—the real world can always find new things for you.

Fifth is *connectivity*. Not everyone always has a high-speed Internet connection; being occasionally connected is the norm. You have to think about whether you can do without connectivity, what to do when it goes down, and how you can work around it not being there or if it keeps dropping and coming back—how do you live with those scenarios? To provide a good experience using an app, you should always make sure the user knows what's going on. They may think, *Did it send? I'm offline, what's happened, what does that mean?* When they're asking questions like these, their experience has been negatively affected. Uncertainty in the user is always bad.

SOFTWARE ENGINEERING RADIO

Visit www.se-radio.net to listen to these and other insightful hour-long podcasts.

RECENT EPISODES

- 437—Tim Sneath, product manager for Flutter and Dart at Google, discusses with host Gavin Hendry what Flutter is, why it was created, and where Dart came from, as well as what the different layers of Flutter are and why it's so popular.
- 436—Yi Pan, lead maintainer of Apache Samza, discusses the internals of the Samza project as well as the Stream Processing ecosystem with host Adam Conrad.
- 435—Julie Lerman discusses Object–Relational Mappers and Entity Framework with host Jeremy Jung.

UPCOMING EPISODES

- Host Justin Beyer discusses cyprography with Jean-Philippe Aumonsson.
- James Smith talks with host Priyanka Raghavan about software bugs.
- Host Kanchan Shringi discusses UX for Enterprise Software with Arin Bhowmick.

Sixth are *resources*. The biggest resource you need to consider on a mobile device is power—will what I'm doing drain the battery? No one will use your app if it makes their battery run down quickly. The next biggest one is disk space. Mobile phones are highly constrained in terms of the space they can use. The availability of disk space for users who have many images on their phones might be something you need to consider.

Can you use telemetry data to monitor and understand your app's battery usage or storage needs? Or do premium phones on modern operating systems already limit battery usage by apps anyway?

Measuring power consumption or how quickly you drain the battery can be harder to test for than other things. The operating system can take care of some of that, and there are settings reports on how much of the bandwidth, power, and screen time the app is using. But you can't rely on the operating system to take care of this for you. The operating system will do what's best for the operating system, but you need to do what's best for your app and your users.

You don't need telemetry for disk space, but you can check how much disk space is there before trying to save images locally, save data, or make a copy of the database, which will take up a lot of space. You have to be aware of the resources you're using and be sensitive to the needs of the device as well as the needs of your app. Resources also include things such as the device camera—how many, and how do I use them?—and the location sensor.

What is one important thing that every app should do?

Basic analytics and error reporting—if something goes wrong, you want to know about it. Don't rely on the user telling you because they won't know enough to give you the information you need. Analytics tell you what devices and operating systems people are using. You might want to use this new feature, but it's available only in the latest version of the operating system. If your users don't update to the latest operating system quickly, then that's a waste of your effort.

GAVIN HENRY is the founder and managing director of SureVoIP, an Internet telephony service in the U.K. Contact him at ghenry@surevoip.co.uk.

EDITOR: Gabriela Marcu, gmarcu@umich.edu

DEPARTMENT: PERVASIVE HEALTHCARE

This article originally appeared in
IEEE pervasive COMPUTING
MOBILE SYSTEMS | UBIQUITOUS COMPUTING | INTERNET OF THINGS
vol. 19, no. 4, 2020

Empowering Communities With a Smartphone-Based Response Network for Opioid Overdoses

Gabriela Marcu, *University of Michigan*

David G. Schwartz, *Bar-Ilan University*

Janna Ataiants and Alexis Roth, *Drexel University*

Inbal Yahav, *Tel Aviv University*

Benjamin Cocchiaro, *University of Pennsylvania*

Michael Khalemsky, *Bar-Ilan University*

Stephen Lankenau, *Drexel University*

In a Philadelphia neighborhood where opioid overdoses are frequent, neighbors used a smartphone app to request and give help for victims of suspected overdose. A one-year study demonstrated the feasibility of this approach, which empowered the local community to save lives and even respond to overdoses faster than emergency medical services.

The opioid epidemic continues to devastate families, ravage communities, and cause a large number of overdose deaths across the United States. Drug overdoses were the leading cause of injury-related death nationally in 2018, and 70% of those involved an opioid.[1] Philadelphia has reported the highest annual rate of overdose deaths among large cities, at 46.8 per 100 000 individuals compared to Chicago's 15.4 and New York City's 11.2.[2] In addition, thousands of nonfatal drug overdoses each year test Philadelphia's available resources such as hospitals and emergency medical services.

One of the efforts to address this complex public health issue is promoting bystander response among community members in order to prevent overdose fatalities. For every overdose death, several lives are being saved every day by use of the overdose reversal medication naloxone. Naloxone, also known by the brand name Narcan, is administered intranasally to reverse an overdose and prevent death. Laypersons can administer naloxone as a nasal spray with few medical risks, making it a feasible intervention for wide use by the public. In Philadelphia, considerable resources have been channeled into increasing access to naloxone among those who use drugs and members of their social networks.

With increasing naloxone saturation in communities, coordinating response during overdose events could also improve response times and further reduce deaths. Policy efforts are encouraging the development of smartphone applications to connect laypersons carrying naloxone with nearby overdose events, and research is testing such applications. In this column, we summarize the research results reported in an article[3] on a pilot study of our UnityPhilly app and an article[4] on formative research that informed the pilot study. We also reflect on some potential implications of pervasive computing on community and public health efforts.

Digital Object Identifier 10.1109/MPRV.2020.3019947
Date of current version 9 November 2020.

INTEGRATION WITH DISPATCH SYSTEMS

Philadelphia EMS is one of the busiest systems in the United States. While some U.S. cities have adopted text-based emergency messaging to call centers (text-to-911), this service is not supported in Philadelphia. Therefore, we had to find another way for our app to automate, or at least facilitate, communicating information about a reported suspected overdose to EMS.

An initial version of the UnityPhilly app featured an automatic computer-generated voice call to EMS in which the system "spoke" to a human EMS dispatcher and provided the address+GPS coordinates, and a message that an opioid overdose had occurred – without any direct interaction between EMS dispatch and the human signaler. The text of the original message was "Hi, I am reporting an overdose incident that is happening now. This automated message was generated by the UnityPhilly app and will repeat twice. The overdose is occurring at {Street Address}, {City}. The GPS coordinates are {location GPS x} and {location GPS y}. Please send an ambulance with naloxone."

During consultations with Philadelphia EMS in preparation for our pilot study, concerns for situational assessment and control were raised leading to an EMS request that this functionality be removed and that a direct person-to-person voice call be established. Therefore, we modified the app to comply with Philadelphia EMS requirements, which enabled a process whereby standard EMS caller interrogation protocols could be followed irrespective of the additional layperson support provided through the UnityPhilly app. These phone calls were initiated immediately and without delay when a participant, having encountered a suspected opioid overdose, pressed the button to signal the overdose event occurring at that location.

Text-based communication with call centers may become more common, and has the potential to ease integration of apps designed as Emergency Response Communities. This may have advantages for empowering communities to provide peer support, but could mean an influx of activity for call centers, which must also be considered.

THE UNITYPHILLY APP

UnityPhilly is an app designed as an Emergency Response Community, to support laypersons in signaling and responding to opioid overdose incidents.[3,4] Volunteers signal an overdose incident with a single button press, initiating an automated alert to other nearby volunteer app users who answer the alert if they can respond to the scene. Design of the app was informed by needs assessment in the form of interviews and focus groups with end users.[4]

Signalling an Overdose

App users signal an alert when they encounter a suspected opioid overdose, administer naloxone if they have any, and speak with 911 through a phone call initiated by the app. Concurrent with sending the alert, a call is initiated from the signaler's smartphone to a dedicated phone number connecting to the Philadelphia Police EMS (Emergency Medical Services) dispatch unit. Smartphone operating system constraints result in slightly different EMS call behavior for Android and Apple-based phones. On Android handsets, calls are placed immediately when the signaling button is pressed. On Apple iOS handsets, a pop-up with the EMS phone number appears requiring the caller to confirm the dial request.

Location data from the volunteer's smartphone are transmitted to UnityPhilly servers which automatically check for other nearby volunteers and send dispatch alerts with the overdose location to the four closest. For our pilot study, 'nearby' was defined as within a 15-minute estimated time of arrival to the overdose site, calculated dynamically based on the participants' declared transport mode (foot, car, etc.).

Responding to an Overdose

Volunteers receiving the alert can use the app to indicate they are responding or declining to respond to the alert; navigate to the overdose site; communicate with the signaler and other responding volunteers; and review salient overdose information including instruction for recognizing overdose, administering naloxone, and rescue breathing.

The system sends alerts to additional volunteers if an alerted volunteer does not acknowledge within 2 min. In this manner, additional volunteers are notified of the incident until either all four have confirmed they are en-route, or there are no additional volunteers within the set radius. Volunteer locations are automatically updated every 15 min by a message sent from the app to the server. Signalers are automatically informed when nearby volunteers have been found, when volunteers indicate they are responding, and when a volunteer is arriving on scene.

PILOT STUDY IN KENSINGTON

We conducted a pilot study[3] of UnityPhilly in the Philadelphia neighborhood of Kensington, from March 2019 to February 2020. Kensington, where fentanyl, heroin, prescription opioids, and other illegal drugs are openly sold, has Philadelphia's highest concentration of overdose deaths.[4] Kensington is also home to Prevention Point Philadelphia, the only city-sanctioned syringe exchange program in Philadelphia, and one of only two in the state of Pennsylvania. We therefore selected Kensington due to its high number of overdoses, its population density, and the opportunity to leverage the sense of community built around Prevention Point's harm reduction services and support.

Prevention Point is a local leader providing naloxone training, working to meet demand by accommodating a high volume of requests for training sessions. Prevention Point distributed about 5500 doses of naloxone in 2016.[2] The organization's efforts have been recognized as helping to ease the demand on the city's emergency services, and they played a key role on the Mayor's Task Force to Combat the Opioid Epidemic.[2] Our formative research also showed that those who use opioids and are affected by opioid use in Kensington have high levels of trust in Prevention Point.[4]

One hundred twelve community members participated in our pilot study, which they heard about through touchpoints with services provided by Prevention Point. In line with efforts to distribute naloxone to those who are likely to witness an overdose, our model for UnityPhilly is to create a network of people who are actively using opioids, as well as other members of the local community who report they have not had any nonmedical opioid use in the past 30 days. The 112 participants were almost equally divided between these groups, 57 and 55, respectively.

While not everyone in the community will have reliable access to a smartphone, our formative research found that we could reach a critical mass for our Emergency Response Community. Participants were therefore required to have their own smartphone with a data plan, and we installed the UnityPhilly app on their phone before providing training and practice with using the app. We notified participants that their location/movements would be tracked by the app. We also provided naloxone training and two doses of naloxone to each participant.

KEY APP USE OUTCOMES

Participants signaled 291 suspected opioid overdose alerts during the one-year study period.[3]

- Eighty-nine (30.6%) signaled events were determined to be false alarms, i.e., canceled by the signaler within 2 min of the alert being sent or the signaler entering an app chat message to the effect that this was a “false alarm.” Every signaled event initiated a phone call to EMS, irrespective of the alert being true or false, enabling EMS to execute their follow up protocol regardless of layperson responder engagement.
- In 74 (36.6%) of the remaining 202 cases, at least one dose of naloxone was administered by a layperson participating in the study (whether the person signaling or responding through the app). A successful reversal was reported in 71 (95.9%) of these cases.
- In the remaining 128 (63.4%) cases, 911 was called but no naloxone administration or follow up by laypersons was reported by incident survey respondents.
- The first dose of naloxone was provided by a nearby volunteer responding to the alert in 22/74 (29.7%) of cases and by the signaling volunteer in 52/74 (70.3%) of cases.
- One on-scene death was reported (1.35%) and two intervention outcomes were unreported (2.7%).

HOW CAN AN APP HELP?

Locations of Overdoses

The Kensington community environment is characterized by an open-air drug market and about 30% of study participants were homeless. Most incidents (58%) were reported as occurring on the street. We also observed a significant number of in-home overdose signaling (23%) indicating the relevance of this approach in providing at-home support for caregivers and family members of opioid users. Allowing entry of layperson responders into homes or businesses in this study was at the discretion of the person who signaled the alert. Other locations reported included in a business, vehicle, and abandoned building. Monitoring location trends of where suspected overdoses have been reported could inform interventions tailored to the needs of the neighborhood.

Participants also suggested real-time geographic monitoring, to help identify when and where especially lethal batches of opioids are infiltrating a community. In addition to varying strengths across different batches, a surge in fatalities has been attributed to batches being mixed with fentanyl, a substance 50–100 times more potent than morphine.

Informed Voluntary Response

In interviews before and after our pilot study, we heard consistent suggestions for the app to provide contextual information that can help someone who has been signaled about an overdose to make an informed decision about whether to respond. During the pilot study, the current version of UnityPhilly did not provide additional information outside of an address on a map. Participants described several types of risks that responders to an app signal could be exposing themselves to, and felt that responders may not even be cognizant of these risks in the moment. An app could either use context-awareness to provide pertinent information, or facilitate question and answer with the person who signaled the overdose. For example, helping a responder determine if the location is in someone's home, or an abandoned building, and whether or not they are comfortable entering such a location. Or, provide an understanding of the victim's condition, such as how long they may have been unconscious, or whether they were breathing.

Each Minute Matters

During the 52-week study, naloxone was administered at 74 overdose events (1.42 times per week on average), and was done more than 5 min in advance of EMS arrival in 59.46% of cases. Without timely reversal, opioid overdose causes respiratory depression that may deteriorate into apnea, leading to anoxic injury. In the minutes immediately following opioid overdose, "time is brain."

Interaction With Emergency Services

Participants who initiated calls to EMS using the app, in addition to alerting other volunteers, reported staying with the victim until EMS arrival in 89.19% of cases. One study of behavior during drug overdoses found that no more than half of those who respond to an overdose event sought help from emergency services.[5] During our own formative research in Kensington, participants reported an aversion to communicating with EMS or other authorities, especially if there was a chance of interaction with law enforcement—they did not trust police, and feared that they or others could be arrested.[4] That a majority of app users stayed with the victim could be a bias based on the types of users who self-selected to participate in the study and agreed to respond to a signal of an overdose through the app. Regardless, this article demonstrates the feasibility of a smartphone-based network of layperson responders as part of the ecosystem of emergency services.

Community-Based Peer Help

The name of the app, UnityPhilly, was drawn from insights during our initial qualitative research in the Kensington community. Trust was high in one's peer groups, such as fellow opioid users or members of their neighborhood community, and we found a deep camaraderie and desire to support one another. Smartphone-based networks could therefore have the potential to empower members of a marginalized group and underserved neighborhood to unite and help one another. Community members' distrust of institutions such as emergency services, and the perception that outsiders' prejudice renders them less helpful to overdose victims, enhances their reliance on one another.

Designing apps like UnityPhilly to not only facilitate logistical support, but also promote shared identity and social capital, could have meaningful impact on social support within the community. Moreover, this type of design may help members of a community with coping and mental health in the context of significant death and suffering.

Understanding and Managing Personal Risk

At the same time, participants' concerns about a smartphone application tracking their location stemmed from some distrust of fellow community members. In a community deeply affected by opioid use, theft is commonplace, and bartering with naloxone was even reported. Privacy concerns therefore related to one's own community members misusing a smartphone application, with the risk of theft or assault.

Interestingly, despite a distrust of police coming up in our formative interviews, participants made no mention of their location data possibly being obtained by law enforcement. Techniques for preserving privacy while using location-based services should be explored, and future work could help community members understand various potential risks as well as eliciting their preferences for how their privacy could be protected.

CONCLUSION

The distribution of naloxone to those who are likely to witness an overdose is a key evidence-based strategy for addressing the opioid epidemic. Smartphone applications are a novel medium for facilitating naloxone distribution and administration, and policy efforts are encouraging their development. Our findings support the benefits of equipping community members with naloxone and an emergency response community smartphone app, for alerting EMS and nearby laypersons to provide additional naloxone. Individuals affected by opioid use and overdose reacted positively to the concept and use of our UnityPhilly app, which they perceived as a useful tool that could help combat the high rate of opioid overdose fatalities in their neighborhood. A sense of unity with others who have shared their experiences could be leveraged to connect willing volunteers with victims of overdose, but risk must be mitigated for layperson responders.

ACKNOWLEDGMENTS

The authors would like to thank our development partner team at Verint/Nowforce for helping to customize and support the UnityPhilly app. We thank our community partners, including Prevention Point Philadelphia, Esperanza, and Angels in Motion, as well as collaborators in the city of Philadelphia, including the Police Department and Fire/Emergency Medical Services. This project was supported by NIH through the National Institute on Drug Abuse under Grant 5R34DA044758.

REFERENCES

1. Centers for Disease Control and Prevention, Opioid Overdose, May 5, 2020. [Online]. Available: https://www.cdc.gov/drugoverdose/index.html
2. Mayor's Task Force, "The Mayor's task force to combat the opioid epidemic in Philadelphia: Final report & recommendation," 2017. [Online]. Available: http://dbhids.org/wp-content/uploads/2017/05/OTF_Report.pdf, Accessed: Aug. 12, 2020.
3. D. G. Schwartz, et al., "Layperson reversal of opioid overdose supported by smartphone alert: A prospective observational cohort study," *EClinicalMedicine*, 2020. [Online]. Available: https://doi.org/10.1016/j.eclinm.2020.100474
4. G. Marcu, R. Aizen, A. M. Roth, S. Lankenau, and D. G. Schwartz, Acceptability of smartphone applications for facilitating layperson naloxone administration during opioid overdoses," *JAMIA Open*, vol. 3, no, 1, pp. 44–52, Apr. 2020.
5. L. K. Lim, et al., "Factors associated with help seeking by community responders trained in overdose prevention and naloxone administration in Massachusetts," *Drug Alcohol Depend*, vol. 204, Nov. 2019, Art. no. 107531.

GABRIELA MARCU is an assistant professor with the School of Information, University of Michigan. Contact her at gmarcu@umich.edu.

DAVID G. SCHWARTZ is a professor of information systems with the Graduate School of Business Administration, Bar-Ilan University. Contact him at david.schwartz@biu.ac.il.

JANNA ATAIANTS is a postdoctoral research fellow with the

Dornsife School of Public Health, Drexel University. Contact her at ja633@drexel.edu.

ALEXIS ROTH is an associate professor with the Dornsife School of Public Health, Drexel University. Contact her at amr395@drexel.edu.

INBAL YAHAV is a senior lecturer with the Coller School of Management, Tel Aviv University. Contact her at inbalyahav@tauex.tau.ac.il.

BENJAMIN COCCHIARO is an adjunct fellow with the Center for Public Health Initiatives, University of Pennsylvania. Contact him at ben.cocchiaro@gmail.com.

MICHAEL KHALEMSKY is currently working toward the PhD degree in information systems with the Graduate School of Business Administration, Bar-Ilan University. Contact him at khalemsky@gmail.com.

STEPHEN LANKENAU is a professor and the associate dean for research with the Dornsife School of Public Health, Drexel University. Contact him at sel59@drexel.edu.

EDITOR: Schahram Dustdar, dustdar@dsg.tuwien.ac.at

This article originally appeared in IEEE Internet Computing vol. 25, no. 3, 2021

DEPARTMENT: INTERNET OF THINGS, PEOPLE, AND PROCESSES

Toward Distributed, Global, Deep Learning Using IoT Devices

Bharath Sudharsan, *National University of Ireland Galway, Galway H91 TK33, Ireland*
Pankesh Patel, *National University of Ireland Galway, Galway H91 TK33, Ireland*
John Breslin, *National University of Ireland Galway, Galway H91 TK33, Ireland*
Muhammad Intizar Ali, *Dublin City University, Dublin 9, Ireland*
Karan Mitra, *Lulea University of Technology, 97187 Lulea , Sweden*
Schahram Dustdar, *TU Wien, 1040 Vienna, Austria, Cardiff University, Cardiff CF10 3AT, U.K.*
Omer Rana, *Cardiff University, Cardiff CF24 3AA, U.K.*
Prem Prakash Jayaraman, *Swinburne University of Technology, Hawthorn VIC 3122, Australia*
Rajiv Ranjan, *Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.*

Deep learning (DL) using large scale, high-quality IoT datasets can be computationally expensive. Utilizing such datasets to produce a problem-solving model within a reasonable time frame requires a scalable distributed training platform/system. We present a novel approach where to train one DL model on the hardware of thousands of mid-sized IoT devices across the world, rather than the use of GPU cluster available within a data center. We analyze the scalability and model convergence of the subsequently generated model, identify three bottlenecks that are: high computational operations, time consuming dataset loading I/O, and the slow exchange of model gradients. To highlight research challenges for globally distributed DL training and classification, we consider a case study from the video data processing domain. A need for a two-step deep compression method, which increases the training speed and scalability of DL training processing, is also outlined. Our initial experimental validation shows that the proposed method is able to improve the tolerance of the distributed training process to varying internet bandwidth, latency, and Quality of Service metrics.

IoT datasets are now being produced at an ever increasing rate, as emerging IoT frameworks and libraries have simplified the process of continuous monitoring, real-time edge-level processing, and encrypted storage of the generated multimodal image, audio, and sensor data. Such data are generated by a variety of hardware systems operating in indoor and outdoor infrastructures, including smart factory floors, AR/VR experience centers, smart city sensors, etc. In order to complete training in a reasonable time when using such large scale, high-quality IoT datasets that have been collected over decades, we need a scalable distributed training system that can efficiently harness the hardware resources of millions of IoT devices. Particularly, such a system should take account of current network connectivity between these devices, and able to collectively train to produce the final problem-solving deep learning (DL) models at very high speeds.

Digital Object Identifier 10.1109/MIC.2021.3053711
Date of current version 18 June 2021.

Instead of following the traditional approach that loads such datasets and trains a model locally within a GPU cluster or a data center, we utilize distributed training on multiple IoT devices as:

i) Considering the GPU to IoT devices ratio, IoT devices are much greater in number, i.e., market estimates show that roughly 50 Billion Micro-Contoller Units (MCU) and small CPU chips were shipped in 2020, which far exceed other processors like GPUs (only 100 Million units sold);
ii) Every modern household does not compulsorily own a GPU, yet it roughly has around a dozen medium resource IoT devices which when efficiently connected together can, within a home network, train machine learning models without depending on Cloud or GPU servers that can perform the same training task at very high speeds, but at additional cost;
iii) In most real-life IoT scenarios, the training dataset used to produce a learned model can often be hard to source due to GDPR and privacy concerns. In such cases, we need an algorithm to directly utilize capability of the IoT device hardware without disturbing routine operation of the device. This algorithm when deployed across user devices should make use of locally generated data to "collectively" train a model without storing live data on a central server. Thus, locally producing learned models from data without violating the privacy protection regulations;
iv) Training advanced DL models on a single GPU might consume days or even weeks to converge. Hence, if we design and use an intelligent algorithm that can tolerate high latency and low bandwidth constraints, we can collectively harness the idle hardware resource of thousands of mid-sized IoT devices and complete training at very high speeds. For example, at the time of writing, the latest GEFORCE RTX 2080 Ti GPU has 11GB RAM but costs ~US $1500. Whereas one Alexa smart speaker device has 2 GB RAM and efficiently connecting 20 such devices can collectively pool 40 GB of RAM. In this way, we can complete training faster on such resources, if coordinated correctly, compared to expensive GPU and at a comparatively smaller investment—particularly by utilizing idle capacity of smart IoT devices that exist across the world.

The hardware of IoT devices is not designed for DL workloads. Resource-friendly model training algorithms like Edge2Train[1] could be used in distributed setups for training models MCUs and limited capacity CPUs of IoT devices. We identify challenges involved with DL model training on hardware of common IoT devices such as video doorbells, smart speakers, cameras, etc. To overcome some of the challenges, we also present a two-step deep compression method that increases the training speed and scalability of DL training processing.

Outline. For globally distributed DL model training scenarios, in section "Distributed Global Training: Research Challenges," we present our bottleneck analysis. Section "Proposed two-step Deep Compression Method and Initial Experimental Results" contains our solution to address the challenges highlighted in Section "Distributed Global Training: Research Challenges." In section "Discussion," we conclude by providing greater context for future work.

DISTRIBUTED GLOBAL TRAINING: RESEARCH CHALLENGES

In the large-scale distributed/collaborative learning domain, distributed training has seen limited adoption, especially when the target is to train a DL model than can perform video analytics tasks such as object detection (e.g., detect FedEx, USPS vehicles, etc.) for package theft prevention, detect, and recognize unsafe objects such as a gun to reduce crime, identify known/unknown faces. This is because:

i) Models that can learn from video datasets have a dense (i.e., large number of parameters and layers) architecture design that requires significant computational resources when compared to models designed to learn from image or audio datasets. For example, the popular ResNet-50 model trained using a 2-D image dataset consumes around 4 GFLOPs, whereas a ResNet-50 Inflated 3-D model contains 3-D convolutional kernels to model temporal information in a video, consuming 30 GFLOPs, i.e., more than 7× times larger than the previous case;
ii) These datasets can be significant in size, hence consuming high internet bandwidth when loading video from a (central) data server to training devices that are geographically distributed. For example, the ImageNet dataset has 1.28M images, whereas the Kinetics-400 video dataset has 63M frames, i.e., 50× times larger; and
iii) Finally, complex models trained on such datasets can have millions of parameters and gradients that need to be quickly exchanged (with minimum latency) among devices during

distributed training, which again increases internet traffic (and charges to consumers) and more critically can lead to slow convergence when devices involved in training suffer from network latency issues. In short, the bottlenecks are due to the demand for high computational power, time overhead associated with dataset loading I/O, and slow exchange of model gradients. In the rest of this section, each of these three bottlenecks are explained in more detail.

High FLOPs Consumption

Unlike for 2-D image recognition models, the input/activation of video analytics DL networks has [N, T, C, H, W] as its five dimensions, where: N is the batch size, T refers to temporal timestamps, the channel number is C, and spatial resolution H & W. To reduce computational overhead and network congestion, we can train using the same target dataset by applying 2-D CNN to each image frames from the video. Using such an approach, the temporal relationship between the frames cannot be modeled/learned, which is crucial to understanding the scenes (labeled) from the video datasets. Hence, inflating the 2-D to 3-D convolution layer results in producing an I3-D model, which grows the model size by *k* times. For distributed learning of spatio-temporal data, the models with 3-D convolutions, in addition to model size demands, also suffers from having a large number of parameters, which is the main reason to slow down the training and communication process even within a GPU cluster and in real-world networks. Consequently, training will stall when unexpected network issues are encountered.

Expensive Dataset Loading I/O

Video network architectures available in ML Hubs and marketplaces (Google AI Hub and TensorFlow Hub) usually sample many frames from video datasets and use them as input during learning (i.e., top models[2] sample 32 and 64 frames). Then, they progressively reduce the temporal resolution by realizing "temporal pooling" techniques.[3] Another orthogonal approach is to design networks that sample and use fewer frames (i.e., eight frames) during learning and maintain the same temporal resolution to retain information from the video dataset. In both designs, the overall computational requirements are similar, but the former involves additional sampling and full dataset loading steps, increasing the dataset loading I/O at the data server, while making data loading on many distributed IoT devices challenging when considering the limited memory and internet bandwidth available in practice.

Slow Exchange of Model Gradients

During training, maintaining good scalability, low latency, and high bandwidth internet connection is mandatory at least during gradients exchange.[3] Existing large-scale distributed learning studies and frameworks require high-end Infiniband network infrastructure where bandwidth ranges from 10 to 100 Gb/s, with a $\sim$1 μs latency. Even if we increase bandwidth by stacking (aggregating) hardware, latency improvements are still difficult to achieve. In contrast to our assumption, latency in real-world scenarios can be further exacerbated due to queueing delay in switches and indirect routing between service providers. This bottleneck makes distributed training scale poorly in real-world network conditions, particularly when transmitting datasets in addition to the gradients.

Handling Dataset I/O Efficiency

Video datasets are usually stored in a high-performance storage system (HPSS) or a central data server that is shared across all worker nodes—in our case these are IoT devices distributed across the world. Although HPSS systems have good sequential I/O performance, their random-access performance is inferior, causing bottlenecks for large data traffic. Most existing I3-D models a high frame rate (within a video), then perform a downsampling to reduce overall data size. Given the distributed training scenario being considered, we argue that such designs waste bandwidth. Consequently, research needs to consider novel data approximation, sampling and filtering methods. For example, in the context of video datasets, one can develop a method to identify videos that have multiple similar frames (i.e., we say that nearby frames contain similar information), then load and share only nonredundant frames during distributed training. Similarly, for other datasets associated with images and sensor readings, we recommend filtering or downsampling the data without losing information, then distributing it during training. Therefore any approximation, sampling, and filtering method will need to be correctly parameterized while considering the resource-constrained nature of IoT devices.

Variable Training and Convergence Speed

Research has shown that naive synchronous stochastic gradient descent (SSGD) achieves poor scalability in real time and distributed DL model training, making training using 100 distributed GPUs slower than training on 1 GPU. Unlike SSGD, asynchronous SGD (ASGD) relaxes synchronization enabling its use across many real-world applications. D2[10] and AD-PSGD[11] perform only partial synchronization in each update to overcome latency issues. Such large-scale training takes advantage of data parallelism by increasing the number of contributing devices, but at the cost of data transfer between devices (e.g., exchange of parameters), which can be time consuming, especially when many devices are pooled. This dwarfs the savings in computation time and producing a low computation-to-communication ratio. However, such distributed learning approaches do not scale well when network latency is high. Additionally, lower network bandwidth, expensive/limited mobile data plan, and intermittent network connection, which are all common across use of mobile devices, also impact our training scenarios. Hence, if we use SSGD, ASGD, D2, AD-PSGD (or any such native algorithms) across a large number of medium-resource IoT devices, the target DL model might never converge to a suitable level of accuracy. Hence, there is a need for a method that can efficiently communicate with a large number of heterogeneous IoT devices, even under real-world internet latency and bandwidth constraints, and complete training at high speeds. As SSGD, ASGD, D2, AD-PSGD can be adapted to learn a globally distributed model, there is a need to develop benchmarking techniques that compare them against common evaluation metrics including average accuracy, training time, and convergence speed. Eventually, these evaluation metrics will need to be formulated as a unified distributed-training performance model. Metaheuristic techniques such as genetic programming and particle swarm optimization could be used to solve and find feasible (Pareto optimal) solutions for improving performance model.

Handling Network Uncertainties

Distributed learning can be impacted by properties of access links that connect IoT devices (sensors and actuators) to edge gateways and/or cloud nodes. These uncertainties include time-varying connectivity, network unavailability, and time-varying traffic patterns research has indicated that wireless network bandwidth and availability fluctuates dramatically due to weather conditions, signal attenuation, and channel interference. For instance, consider the use of SSGD during a distributed training process, where only one gradient transmission occurs in one iteration. This aspect can worsen with an increase in the number of transmissions, and if the previously sent gradients arrive late along with recent gradients (late arrival due to network congestions). The second issue we expect is the large variance in latency, which is common in real-world IoT networks, especially where devices have long-distance connections and communicate via a range of networks, e.g., long-range/low-power communications using LoRa-WAN and NarrowBand-IoT and more powerful high-bandwidth WiFi, 4G/5G radio. While we can aim to maintain a low average latency by choosing and involving only IoT devices with stable internet connection, changes in device network connectivity due to mobility (e.g., when the IoT device is placed in a car) can cause variable latency.

Handling Staleness Effects

Most popular distributed model training techniques (e.g., SSGD, ASGD, D2, AD-PSGD) adopt a nonsynchronous execution approach for alleviating network communication bottleneck that produces *stale* parameters, i.e., the model parameters arrive late, not reflecting the latest updates. Staleness not only slows down convergence but also degrades model performance. Despite notable contributions in distributed learning,[12,13] the effects of staleness during training can lead to model instability,[14] because it is practically not feasible to monitor and control staleness in the current complex IoT environments containing heterogeneous devices using different network protocols. This challenge can be addressed by designing *accuracy guaranteeing dynamic error compensation and network coding techniques*—primarily a light-weight technique that adopts a two-step process. In the first step gradient synchronization is not performed, instead each participating IoT device updates their part of the model with locally available gradients (e.g., local learning). In the second step, IoT devices perform gradient synchronization based on the computed averaged gradients, which takes account of the designed error compensations.

PROPOSED TWO-STEP DEEP COMPRESSION METHOD AND INITIAL EXPERIMENTAL RESULTS

In this section, we present an initial approach to handle network uncertainties and data staleness challenges in the context of distributed training of DNNs. In our distributed training scenario, we model the communication time t_c as

$$t_c = \text{latency} + (\text{model size}/\text{bandwidth}). \qquad (1)$$

Both latency and bandwidth are dynamic and depend on the network condition, which we cannot control. Instead, in the following, we present model size reduction techniques that can be applied to various parts of the DL model to save communication time and networking traffic.

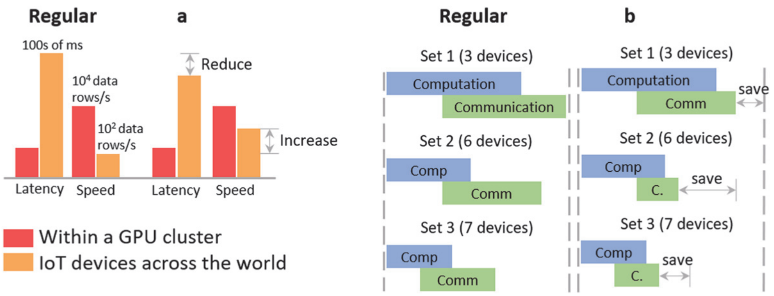


FIGURE 1. Comparing distributed training within a GPU cluster versus training using geographically distributed IoT devices. Our proposed two-step deep compression method can (a) tolerate latency and increase training speed, (b) reduce the communication-to-computation ratio to improve scalability and reduce communication costs.

To reduce the communication bandwidth, we recommend quantizing the model gradients to low-precision values, then transmitting these to other IoT devices or servers. The popular methods are: 1-bit SGD,[4] which achieves a 10× speedup for speech datasets. In QSGD,[5] the tradeoff between model accuracy and gradient precision was balanced. Other work demonstrate the convergence of quantized gradient training of various CNNs and RNNs. A few quantize the entire model, including gradients then perform training and a few studies use different bit sizes (e.g., DoReFa-Net[7] uses 2-bit gradients with 1-bit weights). Threshold quantization method[6] transmits gradients only when they exceeds a set threshold, which in practice is hard to choose. To improve this, a fixed proportion of positive and negative gradient was chosen[8] to update separately.

Since the theoretical quantization limit cannot exceed 32, to address this limitation, gradient sparsification methods are being applied and investigated in this distributed training setting. In the studies that sparsify the gradients by gradient dropping, the method from Ba *et al.*[9] saved 99% of gradient exchange while only compromising 0.3% of the BLEU score for a machine translation dataset. Some studies automatically tune this compression rate based on gradient activity and show 200x compression of fully connected layers for the ImageNet dataset.

From our discussions in Section "Distributed Global Training: Research Challenges," it is apparent that scalability is essential when connecting a large number of devices. To improve scalability, we need to significantly reduce communication frequency, where the communication cost is determined by network bandwidth and latency [see (1)]. All conventional studies focus on reducing the bandwidth requirements, as the latency between GPUs inside a cluster or servers inside a data center is usually low. In contrast, in our use case, since we propose to perform the same training but on IoT device hardware that is geographically distributed, latency still remains an issue due to physical device separation. For instance, if we can achieve X times training speedup on Y machines, the overall distributed training scalability (defined as X/Y) increases. Next, if we can also tolerate latency, the speedup will improve further since high latency severely reduces scalability.

We propose a two-step method to improve live model compression during training, yet without altering the DL model architecture and also without compromising the model accuracy. Our two-step deep compression method jointly aims to increase the training speed and scalability. Particularly, the first step aims to tolerate variation in real-world latency and bandwidth issues by sparsely transmitting only the important gradients. The second step aims to reduce communication-to-computation ratio and improve scalability by locally accumulating gradients, then encode and perform transmission only after crossing the gradient threshold. In the rest of this section, we describe each of these steps.

In the First step, we identify the important gradients, using gradient magnitude as the simple heuristic (users can also choose other selection criteria). We accumulate these important gradients locally to not forget the learned information. Since this step reduces the gradient synchronization frequency by not allowing to transmit all the gradients, as shown in Figure 1 (a) the training process can tolerate latency (does not reduce the dynamic real-world latency since it is practically not possible). This results in increasing training scalability, enabling the participation of more IoT devices to complete training at higher speeds.

In the Second step, after the set threshold (dynamically derived for the model in use) for the accumulated gradients is crossed, we encode the gradients (not quantizing like previous works) then transmit them to other contributing devices involved in the training process or to the parameter server. As shown in Figure 1(b). this step improves scalability by reducing the communication-to-computation ratio by sending all the important gradients, not at defined intervals, but only when required.

Briefly, during training, both the steps jointly work to improve training speed and scalability by *accumulating, encoding, and sparsely transmitting only the important gradient*s.

CONCLUSION

In this article, we presented an approach for training DL models on idle IoT devices, millions of which exist across the world. With an increase in mechanisms to connect such devices to a network, the potential for using such devices to support learning on locally collected data has increased. As data are maintained locally (and never transferred to a server), user privacy is also maintained—as the developed model can then be aggregated with other models (without the need to transfer raw data). We have identified and studied challenges associated with building such machine learning models, and presented a two-step deep compression method to improve distributed training speed and scalability.

The proposed approach can be used to interconnect DL frameworks executed on large scale resources (such as TensorFlow on GPU clusters) with proposals from the TinyML community (studies that design resource-friendly models for embedded systems) since we enable distributed training of computationally demanding models on distributed idle IoT devices. TinyML and related approaches often only undertake inference on IoT devices and assume that a model is constructed at a data center. A learned model is subsequently modified (e.g., using quantization) to execute on a resource constrained device (e.g., using TensorFlow-Lite). Support for performing training on resource limited devices is still limited at present—with general approaches provided in frameworks such as "Federated Learning," where a surrogate model is constructed on each remote resource, and models are then aggregated at on a cloud server. There is also an assumption within Federated Learning that each dataset (from a participating IoT device) follows the IID distribution (identical, independently distributed).

Since our method can significantly compress gradients during the training of a wide range of NN architectures such as CNNs and RNNs, the proposed approach can also be utilized alongside TF-Lite and Federated Learning approaches thereby providing the basis for a broad-spectrum of decentralized and collaborative learning applications.

ACKNOWLEDGMENTS

This publication has emanated from research supported by grants from the European Union's Horizon 2020 research and innovation program under grant agreement number 847577 (SMART 4.0 Marie Sklodowska-Curie actions COFUND) and from Science Foundation Ireland (SFI) under Grants SFI/16/RC/3918 and SFI/12/RC/2289_P2 cofunded by the European Regional Development Fund.

REFERENCES

1. B. Sudharsan, J. G. Breslin, and M. I. Ali, "Edge2train: A framework to train machine learning models (SVMS) on resource-constrained IoT edge devices," in *Proc. 10th Int. Conf. Internet Things*, 2020, Art. no. 6.
2. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1711.07971
3. J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," 2018. [Online]. Available: https://arxiv.org/abs/1811.08383
4. F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs." in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014.
5. D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," 2016. [Online]. Available: https://arxiv.org/abs/1610.02132
6. N. Strom, "Scalable distributed DNN training using commodity GPU cloud computing," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1488–1492.
7. S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 2016. [Online]. Available: https://arxiv.org/abs/1606.06160
8. N. Dryden, T. Moon, S. A. Jacobs, and B. V. Essen, "Communication quantization for data-parallel training of deep neural networks," in *Proc. Workshop Mach. Learn. High Perform. Comput. Environ.*, 2016, pp. 1–8.
9. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," 2016. [Online]. Available: https://arxiv.org/abs/1607.06450
10. H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D2: Decentralized training over decentralized data," 2018. [Online]. Available: https://arxiv.org/abs/1803.07068

11. X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," 2017. [Online]. Available: https://arxiv.org/abs/1710.06952
12. I. Mitliagkas, C. Zhang, S. Hadjis, and C. Ré, "Asynchrony begets momentum, with an application to deep learning," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput.*, 2016, pp. 997–1004.
13. B. Qian *et al.*, "Orchestrating the development lifecycle of machine learning-based IoT applications: A taxonomy and survey," *ACM Comput. Surv.*, vol. 53, 2020, Art. no. 82.
14. W. Dai, Y. Zhou, N. Dong, H. Zhang, and E. P. Xing, "Toward understanding the impact of staleness in distributed machine learning," 2018. [Online]. Available: https://arxiv.org/abs/1810.03264

BHARATH SUDHARSAN is currently working toward the Ph.D. degree with the CONFIRM SFI Centre for Smart Manufacturing, Data Science Institute, National University of Ireland Galway, Ireland. His research areas are resource-constrained IoT devices, edge intelligence and analytics, real-time machine training. He is the corresponding author of this article. Contact him at bharath.sudharsan@insight-centre.org.

PANKESH PATEL is Senior Researcher with the CONFIRM SFI Centre for Smart Manufacturing, Data Science Institute, National University of Ireland Galway, Ireland. His academic background and research work focus on building software development tools to easily develop applications in the cross-section of the Internet of Things/Industry 4.0, artificial intelligence, edge computing, and cloud computing. Contact him at pankesh.patel@insight-centre.org.

JOHN BRESLIN is a Personal Professor (Personal Chair) in electronic engineering with the College of Science and Engineering, National University of Ireland Galway, Ireland, where he is the Director of the TechInnovate/AgInnovate programmes. Contact him at john.breslin@insight-centre.org.

MUHAMMAD INTIZAR ALI is an Assistant Professor with the School of Electronic Engineering, Dublin City University, Dublin, Ireland. His research interests include data analytics, Internet of Things, stream query processing, data integration, distributed and federated machine learning, and knowledge graphs. Contact him at ali.intizar@dcu.ie.

KARAN MITRA is an Assistant Professor with Luleå University of Technology, Luleå, Sweden. His research interests include quality of experience modelling and prediction, context-aware computing, cloud computing and mobile and pervasive computing systems. Contact him at karan.mitra@ltu.se.

SCHAHRAM DUSTDAR is a Professor of Computer Science and head of the Distributed Systems Group at TU Wien, Vienna, Austria. He was named Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in 2016 for contributions to elastic computing for cloud applications. Contact him at dustdar@dsg.tuwien.ac.at.

OMER RANA is a Professor of Performance Engineering and previously led the Complex Systems research group, School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research interests lie in the overlap between intelligent systems and high-performance distributed computing. He is particularly interested in understanding how intelligent techniques could be used to support resource management in distributed systems, and the use of these techniques in various application areas. Contact him at ranaof@cardiff.ac.uk.

PREM PRAKASH JAYARAMAN is a Senior Lecturer and Head of the Digital Innovation Lab in the Department of Computer Science and Software Engineering, Faculty of Science, Engineering and Technology at Swinburne University of Technology, Melbourne, Australia. Contact him at pjayaraman@swin.edu.au.

RAJIV RANJAN is an Australian-British computer scientist, of Indian origin, known for his research in Distributed Systems (Cloud Computing, Big Data, and the Internet of Things). He is the University Chair Professor for the Internet of Things research with the School of Computing, Newcastle University, Newcastle upon Tyne, U.K. Contact him at raj.ranjan@ncl.ac.uk.

This article originally appeared in
Computer
vol. 53, no. 5, 2020

COLUMN: AFTERSHOCK

Attacking Machine Learning Systems

Bruce Schneier

The field of machine learning security is progressing rapidly, and new risks have been detected. Machine learning technologies and solutions are expected to become prominent features in the information security landscape.

The field of machine learning (ML) security—and corresponding adversarial ML—is rapidly advancing as researchers develop sophisticated techniques to perturb, disrupt, or steal the ML model or data. It's a heady time; because we know so little about the security of these systems, there are many opportunities for new researchers to publish in this field. In many ways, this circumstance reminds me of the cryptanalysis field in the 1990s. And there is a lesson in that similarity: the complex mathematical attacks make for good academic papers, but we mustn't lose sight of the fact that insecure software will be the likely attack vector for most ML systems.

We are amazed by real-world demonstrations of adversarial attacks on ML systems, such as a 3D-printed object that looks like a turtle but is recognized (from any orientation) by the ML system as a gun.[1] Or adding a few stickers that look like smudges to a stop sign so that it is recognized by a state-of-the-art system as a 45 mi/h speed limit sign.[2] But what if, instead, somebody hacked into the system and just switched the labels for "gun" and "turtle" or swapped "stop" and "45 mi/h"? Systems can only match images with human-provided labels, so the software would never notice the switch. That is far easier and will remain a problem even if systems are developed that are robust to those adversarial attacks.

At their core, modern ML systems have complex mathematical models that use training data to become competent at a task. And while there are new risks inherent in the ML model, all of that complexity still runs in software. Training data are still stored in memory somewhere. And all of that is on a computer, on a network, and attached to the Internet. Like everything else, these systems will be hacked through vulnerabilities in those more conventional parts of the system.

This shouldn't come as a surprise to anyone who has been working with Internet security. Cryptography has similar vulnerabilities. There is a robust field of cryptanalysis: the mathematics of code breaking. Over the last few decades, we in the academic world have developed a variety of cryptanalytic techniques. We have broken ciphers we previously thought secure. This research has, in turn, informed the design of cryptographic algorithms. The classified world of the National Security Agency (NSA) and its foreign counterparts have been doing the same thing for far longer. But aside from some special cases and unique circumstances, that's not how encryption systems are exploited in practice. Outside of academic papers, cryptosystems are largely bypassed because everything around the cryptography is much less secure.

I wrote this in my book, *Data and Goliath:*

Digital Object Identifier 10.1109/MC.2020.2980761
Date of current version: 7 May 2020

> *The problem is that encryption is just a bunch of math, and math has no agency. To turn that encryption math into something that can actually provide some security for you, it has to be written in computer code. And that code needs to run on a computer: one with hardware, an operating system, and other software. And that computer needs to be operated by a person and be on a network. All of those things will invariably introduce vulnerabilities that undermine the perfection of the mathematics…*[3]

This remains true even for pretty weak cryptography. It is much easier to find an exploitable software vulnerability than it is to find a cryptographic weakness. Even cryptographic algorithms that we in the academic community regard as "broken"—meaning there are attacks that are more efficient than brute force—are usable in the real world because the difficulty of breaking the mathematics repeatedly and at scale is much greater than the difficulty of breaking the computer system that the math is running on.

ML systems are similar. Systems that are vulnerable to model stealing through the careful construction of queries are more vulnerable to model stealing by hacking into the computers they're stored in. Systems that are vulnerable to model inversion—this is where attackers recover the training data through carefully constructed queries—are much more vulnerable to attacks that take advantage of unpatched vulnerabilities.[5]

But while security is only as strong as the weakest link, this doesn't mean we can ignore either cryptography or ML security. Here, our experience with cryptography can serve as a guide. Cryptographic attacks have different characteristics than software and network attacks, something largely shared with ML attacks. Cryptographic attacks can be passive. That is, attackers who can recover the plaintext from nothing other than the ciphertext can eavesdrop on the communications channel, collect all of the encrypted traffic, and decrypt it on their own systems at their own pace, perhaps in a giant server farm in Utah. This is bulk surveillance and can easily operate on this massive scale.

On the other hand, computer hacking has to be conducted one target computer at a time. Sure, you can develop tools that can be used again and again. But you still need the time and expertise to deploy those tools against your targets, and you have to do so individually. This means that any attacker has to prioritize. So while the NSA has the expertise necessary to hack into everyone's computer, it doesn't have the budget to do so. Most of us are simply too low on its priorities list to ever get hacked. And that's the real point of strong cryptography: it forces attackers like the NSA to prioritize.

This analogy only goes so far. ML is not anywhere near as mathematically sound as cryptography. Right now, it is a sloppy misunderstood mess: hack after hack, kludge after kludge, built on top of each other with some data dependency thrown in. Directly attacking an ML system with a model inversion attack or a perturbation attack isn't as passive as eavesdropping on an encrypted communications channel, but it's using the ML system as intended, albeit for unintended purposes. It's much safer than actively hacking the network and the computer that the ML system is running on. And while it doesn't scale as well as cryptanalytic attacks can—and there likely will be a far greater variety of ML systems than encryption algorithms—it has the potential to scale better than one-at-a-time computer hacking does. So here again, good ML security denies attackers all of those attack vectors.

We're still in the early days of studying ML security, and we don't yet know the contours of ML security techniques.[4,6,7] There are really smart people working on this and making impressive progress,[8] and it'll be years before we fully understand it. Attacks come easy, and defensive techniques are regularly broken soon after they're made public. It was the same with cryptography in the 1990s, but eventually the science settled down as people better understood the interplay between attack and defense. So while Google, Amazon, Microsoft, and Tesla have all faced adversarial ML attacks on their production systems in the last three years,[9] that's not going to be the norm going forward.

All of this also means that our security for ML systems depends largely on the same conventional computer security techniques we've been using for decades. This includes writing vulnerability-free software, designing user interfaces that help resist social engineering, and building computer networks that aren't full of holes. It's the same risk-mitigation techniques that we've been living with for decades. That we're still mediocre at it is cause for concern, with regard to both ML systems and computing in general.

I love cryptography and cryptanalysis. I love the elegance of the mathematics and the thrill of discovering a flaw—or even of reading and understanding a flaw that someone else discovered—in the mathematics. It feels like security in its purest form. Similarly, I am starting to love adversarial ML and ML security, and its tricks and techniques, for the same reasons.

I am not advocating that we stop developing new adversarial ML attacks. It teaches us about the systems being attacked and how they actually work. They are, in a sense, mechanisms for algorithmic understandability. Building secure ML systems is important research and something we in the security community should continue to do.

There is no such thing as a pure ML system. Every ML system is a hybrid of ML software and traditional software. And while ML systems bring new risks that we haven't previously encountered, we need to recognize that the majority of attacks against these systems aren't going to target the ML part. Security is only as strong as the weakest link. As bad as ML security is right now, it will improve as the science improves. And from then on, as in cryptography, the weakest link will be in the software surrounding the ML system.

REFERENCES

1. K. Martineau, "*Why did my classifier just mistake a turtle for a rifle*?" MIT News, July 31, 2019. [Online]. Available: https://news.mit.edu/2019/why-did-my-classifier-mistake-turtle-for-rifle-computer-vision-0731
2. K. Eykholt et al., Robust physical-world attacks on deep learning models. Apr. 10, 2018. [Online]. Available: https:arXiv:1707.08945v5
3. B. Schneier, *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. New York: Norton, Mar. 2015.
4. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, *Concrete problems in AI safety*. July 25, 2016. [Online]. Available: https:arXiv:1606.06565
5. T. Gu, B. Dolan-Gavitt, and S. Garg, BadNets: Identifying vulnerabilities in the machine learning model supply chain, machine learning and security. Mar. 11, 2019. [Online]. Available: https:arXiv:1708.06733
6. R. S. Siva Kumar, D. O. Obrien, K. Albert, S. Viljoen, and J. Snover, Failure modes in machine learning systems. Nov. 25, 2019. [Online]. Available: https:arXiv:1911.11034
7. S. Herpig, "*Securing artificial intelligence – Part 1: The attack surface of machine learning and its implications*," Think Tank at the Intersection of Technology and Society, Stiftung Neue Verantwortung, Berlin, Oct. 2019. [Online]. Available: https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf
8. G. McGraw, H. Figueroa, V. Shepardson, and R. Bonett, "*An architectural risk analysis of machine learning systems: Toward more secure machine learning*," Berryville Inst. of Machine Learning, San Francisco, 2020. [Online]. Available: https://www.garymcgraw.com/wp-content/uploads/2020/02/BIML-ARA.pdf
9. R. S. Siva Kumar et al., Adversarial machine learning: Industry perspectives. Feb. 4, 2020. [Online]. Available: https:arXiv:2002.05646

BRUCE SCHNEIER is a security technologist, fellow, and lecturer at the Harvard Kennedy School and chief of security architecture at Inrupt, Inc. Contact him at www.schneier.com.

EDITORS: Jelena Mirkovic, mirkovic@isi.edu
Bill Newhouse, newhouse@nist.gov

DEPARTMENT: EDUCATION

This article originally appeared in
IEEE SECURITY & PRIVACY
vol. 18, no. 5, 2020

Preparing America's Cyber Intelligence Workforce

Randy Borum and Ron Sanders, *University of South Florida*

Cyber intelligence requires analytic and technical skills. It is a pillar of U.S. national security and is needed in the private sector. We outline a framework for cyber intelligence education, including a dual-track model to balance the technical and analytic requirements.

Cyber intelligence has become an essential and integral part of cybersecurity and cyber defense. The intelligence function should complement our technical tools and countermeasures by focusing on adversaries and strategic competitors, the people behind and above those at the keyboards. These malign actors and groups are driving the threats, and insights into their intentions, motivations, objectives, knowledge, and capabilities allow cybersecurity to move from a reactive to a proactive posture.

At the national level, cyber intelligence (sometimes called *cyber threat intelligence*) is now a key mission objective in the U.S. National Intelligence Strategy (NIS), which defines it as:

> *… the collection, processing, analysis, and dissemination of information from all sources of intelligence on foreign actors' cyber programs, intentions, capabilities, research and development, tactics, targets, operational activities and indicators, and their impact or potential effects on US national security interests. Cyber threat intelligence also includes information on cyber threat actor information systems, infrastructure, and data; and network characterization, or insight into the components, structures, use, and vulnerabilities of foreign cyber program information systems.*[7]

Building an effective cyber intelligence capability not only requires state-of-the-art technical equipment and personnel but also disciplined and specialized analytic personnel and tradecraft. Very few of today's cybersecurity education programs, however, provide systematic training in intelligence collection, analysis, and management. A review of academic cybersecurity programs in the United States concluded that "[t]he training paths to become a qualified cyber intelligence analyst are inconsistent or nonexistent in some cases."[1] While some schools offer a specific course in cyber intelligence, only a couple offer a coherent program of study or concentration within a related undergraduate or graduate degree program.

This article begins by outlining some key challenges to developing cyber intelligence capabilities in the cybersecurity workforce and scoping the intelligence function in cybersecurity contexts. Then, it charts a path forward by proposing a competency-based framework for cyber intelligence education to balance the technical and analytic dimensions and suggesting how cyber intelligence education might be better integrated into basic cybersecurity curricula and developed as a specialty area as well.

NATIONAL CHALLENGES IN DEVELOPING CYBER INTELLIGENCE CAPABILITIES IN THE CYBERSECURITY WORKFORCE

Cybersecurity comprises a dynamic and evolving range-of-work roles and functions. "Effective cybersecurity workforce planning requires a clear understanding of the gaps between the workforce of today and the needs of tomorrow."[2] Cyber intelligence

Digital Object Identifier 10.1109/MSEC.2020.3005035
Date of current version: 3 September 2020

capabilities are among the most significant of those gaps. Organizations and government agencies are using the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework (NICE Framework)[3,4] to inventory their current cyber workforce and to map out their future needs. While the NICE Framework encompasses the entire spectrum of cybersecurity-related work roles and tasks (not just those for cyber intelligence), effective intelligence capabilities can be found in all seven categories. One of the seven NICE Framework categories is titled "Analyze," and is composed of the following "specialty areas":[4]

> ***All-Source Analysis***
> *Analyzes threat information from multiple sources, disciplines, and agencies across the Intelligence Community. Synthesizes and places intelligence information in context; draws insights about the possible implications.*
>
> ***Exploitation Analysis***
> *Analyzes collected information to identify vulnerabilities and potential for exploitation.*
>
> ***Language Analysis***
> *Applies language, cultural, and technical expertise to support information collection, analysis, and other cybersecurity activities.*
>
> ***Targets***
> *Applies current knowledge of one or more regions, countries, non-state entities, and/or technologies.*
>
> ***Threat Analysis***
> *Identifies and assesses the capabilities and activities of cybersecurity criminals or foreign intelligence entities; produces findings to help initialize or support law enforcement and counterintelligence investigations or activities.*

Cyber intelligence as a specialty may include these five domains but intelligence-related functions operate across other NICE Framework categories/specialty areas as well; for example, in the "Collect and Operate," "Operate and Maintain," and "Protect and Defend" categories.

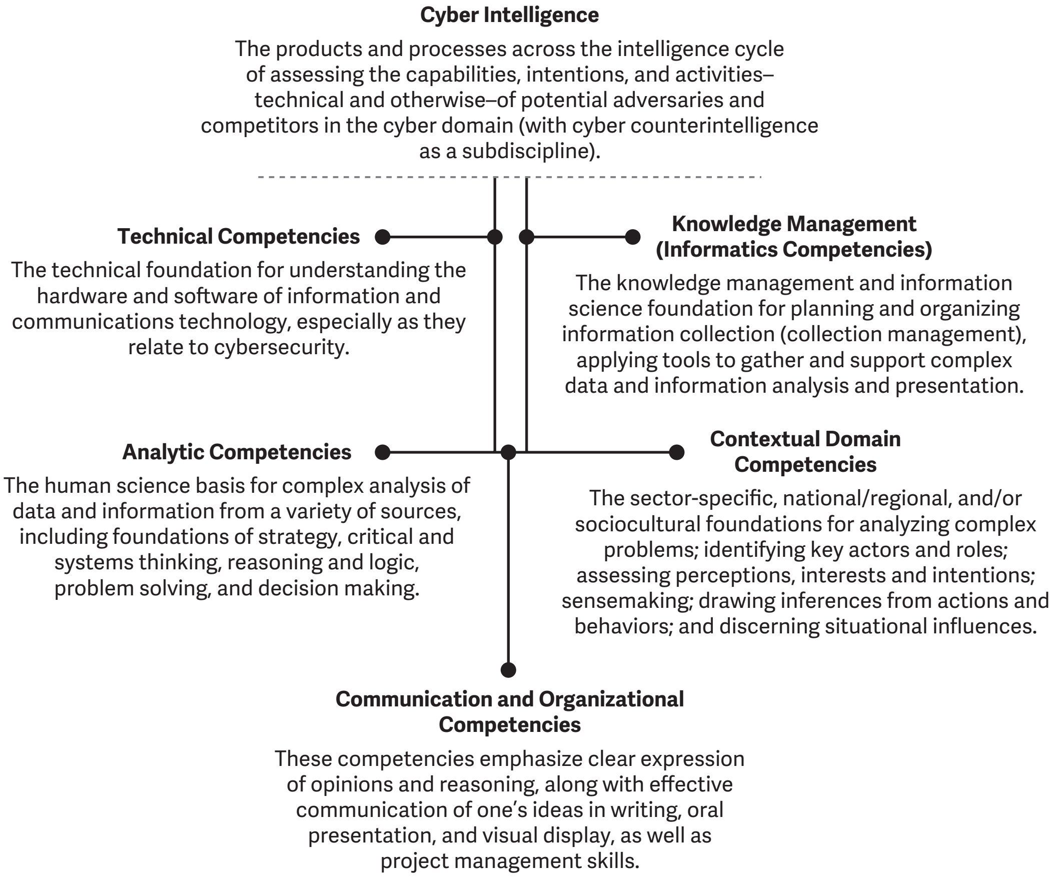
The U.S. government and the private sector will need cyber intelligence analysts within their cybersecurity workforce, along with a cadre technical cybersecurity professionals who collectively can perform some of the critical cyber intelligence functions, such as:[5]

- Maintaining a "deep and current knowledge" of an organization or agency's "attack surface…"
- Identifying "its most vulnerable and valued targets…"
- Discovering and suggesting ways to protect its "technical vulnerabilities"…
- Maintaining "situational awareness on what malicious actors are using and targeting…"
- Developing "techniques and program custom tools to detect local changes, identify suspect interactions," and to identify and "respond to what the malicious actors are doing."
- Assessing attackers' intentions, "motivation, language organization, and social behaviors" and grouping "threat actors logically to create effective 'cyber' profiles of groups, actors, and campaigns…."

While the tradecraft of analysis and intelligence have been around for years, the discipline of cyber intelligence is continuing to evolve and mature, and so too are the training and education requirements necessary to prepare individuals for the discipline.

A FRAMEWORK FOR TRAINING AND EDUCATION IN CYBER INTELLIGENCE

With few specified career paths or systematic



Cyber Intelligence

The products and processes across the intelligence cycle of assessing the capabilities, intentions, and activities–technical and otherwise–of potential adversaries and competitors in the cyber domain (with cyber counterintelligence as a subdiscipline).

Technical Competencies

The technical foundation for understanding the hardware and software of information and communications technology, especially as they relate to cybersecurity.

Knowledge Management (Informatics Competencies)

The knowledge management and information science foundation for planning and organizing information collection (collection management), applying tools to gather and support complex data and information analysis and presentation.

Analytic Competencies

The human science basis for complex analysis of data and information from a variety of sources, including foundations of strategy, critical and systems thinking, reasoning and logic, problem solving, and decision making.

Contextual Domain Competencies

The sector-specific, national/regional, and/or sociocultural foundations for analyzing complex problems; identifying key actors and roles; assessing perceptions, interests and intentions; sensemaking; drawing inferences from actions and behaviors; and discerning situational influences.

Communication and Organizational Competencies

These competencies emphasize clear expression of opinions and reasoning, along with effective communication of one's ideas in writing, oral presentation, and visual display, as well as project management skills.

FIGURE 1. A framework for education in cyber intelligence. (Source: [6]; adapted with permission.)

curricula in cyber intelligence, the current array of roles, skills, preparation, and standards among professionals engaged in cyber intelligence activity is fuzzy and varied. As a foundation for standardizing cyber intelligence training and education standards, the field of cyber intelligence would benefit from a common framework that outlines "what a competent professional in the field must know."[8] Because, like most intelligence disciplines, cyber intelligence is as much or more of an analytic discipline than a purely technical one, professionals in the field must be capable of formulating complex hypotheses, conducting and/or critically evaluating research, acquiring and managing new knowledge, generating and analyzing courses of inquiry and action (that is, collection), framing and solving complex problems, expressing clearly reasoned opinions, and communicating effectively in writing, oral presentation, and visual display—in addition to understanding computing and information security fundamentals and technical exploitation (though to a lesser extent than technical specialists).

Drawing on the foundations of knowledge in cyber intelligence's parent disciplines—cybersecurity and intelligence studies—it is possible to construct the beginnings of a competency-based knowledge framework for cyber intelligence and to parse those models into levels that distinguish between knowledge competencies (awareness and understanding) and proficiencies (skills and abilities). Such a framework would help government agencies and the private sector to conduct a needs assessment, craft appropriate job qualifications and position descriptions, and begin to effectively assess the proficiencies and performance of those analysts who perform those roles. They might also be applied to develop a model curriculum structure for cyber intelligence education. The five categories of competencies (Figure 1) would include the following:[6]

***Technical Competencies:** The technical foundation for understanding the hardware and software of information and communications technology, especially as they relate to cybersecurity, including the operation and underlying mechanisms of [information and communications technology, data storage (e.g., in the cloud) and networks, and operating systems; the mechanisms of] technical (e.g., malware) and human (e.g., social engineering) vulnerabilities and exploitations; and applied principles and tools of information security, including risk assessment/ management, intrusion detection, cryptography, network defense, incident response and recovery.*

***Analytic Competencies:** The social-scientific basis for complex analysis of data and information from a variety of sources, including foundations of strategy, [critical and] systems thinking, reasoning and logic, problem [framing/]solving, and decision making. Emphasis is placed on hypothesis generation and testing, and formulating, selecting, and applying appropriate qualitative and/or quantitative analytic methodologies, including collection strategies and methods [(both cyber-related and, more broadly, to include other intelligence disciplines)]. This includes recognition and application of ethical and professional/ tradecraft standards in choosing and communicating about those methods. When applying analytic competencies to cyber intelligence it is important for the analyst to understand and consider [and where necessary, be able to articulate] the culture, leadership, behavior, and background of adversaries [in other words, their social biases] as well as consumers.*

***Knowledge Management (Informatics) Competencies:** [The term informatics is used broadly here to refer to the study and practice of applying data, information, and knowledge to improve problem solving and decision making.] The knowledge management and information science basis for planning and organizing information collection [collection management], developing and applying tools to gather and support complex data and information analysis from heterogeneous sources, information visualization, and understanding, utilizing, and [with the above technical competencies as a foundation,] evaluating various information storage and retrieval systems.*

***Contextual Domain Competencies:** The sector-specific, national/regional, psychosocial, and/or sociocultural foundations for analyzing complex problems; identifying key actors and roles; assessing perceptions, interests and intentions; sense-making; drawing inferences from actions and behaviors; and discerning situational influences. [And where applicable, f]oreign language capability and regional/cultural competence (at the strategic, operational, and/or tactical level) may also be included as a domain competency.*

***Communication and Organizational Competencies:** These competencies emphasize clear expression of opinions and reasoning, along with effective communication of one's ideas in writing, oral presentation, and visual display. It also comprises the project management skills necessary to plan, organize, evaluate, motivate, mobilize and control resources, processes and outcomes to achieve specific goals.*

A DUAL-TRACK MODEL: BALANCING THE ANALYTIC-TECHNICAL SPECTRUM

A competency-based framework informs the areas in which a cyber intelligence analyst should be knowledgeable, but it also stands to reason that not every analyst should be expected to demonstrate deep expertise in all of them. The requisite depth of knowledge will vary for different cyber intelligence professionals—particularly as roles and responsibilities differentially emphasize the technical and analytic competencies. For instance, some analysts may require only a foundational understanding of cryptography's terms and basic concepts, whereas a truly proficient cryptographer would be able to independently apply those concepts and mathematical algorithms to create an authentication protocol. Thus, setting the right balance of foundational and more specialized analytic and technical skills is an imperative for effective cyber intelligence workforce development, and consequently, for any training and education framework.

As noted previously, very few integrated and specialized education opportunities exist in cyber intelligence. Thus, the current workforce is composed primarily of people who come from a background of education or experience either in intelligence or in technical computer/systems security, with some degree of cross-training between the disciplines. People with a technical/computer science background and experience will pick up some intelligence concepts, terminology, and tradecraft in the course of their work, just as individuals with a nontechnical intelligence analytic background will learn some cybersecurity concepts, terminology, and technical skills.

That cross-training, however, should not be unsystematic. The technical and analytic knowledge, skills, and abilities represent broad classes of competencies in cyber intelligence; however, it is the integration of these "hard" and "soft" competencies that distinguishes cyber intelligence from other cyber specialty areas. Thus, a dual-track training and development model—with one track for technically trained cybersecurity professionals who wish to specialize in the intelligence function (Technical/Analytic) and another for the classically trained intelligence analysts who wish to specialize in the cyber domain (Analytic/Technical)—is a good first step. Both tracks might be grounded in a common foundation of knowledge and skills but with each branching to develop distinct roles, required skills, preparation, and practices.

This is not to say we should not try to effectively train analysts with more technical skills and technical professionals with more analytic skills but that we should more thoughtfully consider where and how to employ their relative expertise. In each case, greater training and education is required in both analytic and technical areas. As a nation, we must seek to effectively develop and employ the right person for the right role while improving learning in both areas.

THE WAY AHEAD FOR CYBER INTELLIGENCE TRAINING AND EDUCATION

An entry-level cyber intelligence analyst must be equipped to demonstrate basic cyber intelligence competencies when entering the workforce, acquired mainly through a balance of technical and analytic training. Given a common set of foundational knowledges and skills, such training could be dual tracked by weighting content more heavily toward either the technical or the analytic, as appropriate (arriving at, in essence, the technical analyst versus the analytic technician). Track content and focus also could vary based on near-term organizational roles and expectations.

With so few integrated and specialized training opportunities available in cyber intelligence, how might we better integrate the analytic components of the intelligence function into cybersecurity/cyber defense education at all levels, while also advancing cyber intelligence as a specialty area? We suggest using the mechanism of the National Security Agency (NSA)/Department of Homeland Security (DHS) Centers for Academic Excellence in Cyber Defense (CAE-CD).

While not an accreditation program per se, the CAE-CD program establishes guidelines for curricula in cybersecurity and cyber defense, and degree programs that conform to these guidelines are certified as CAE-designated programs. Those standards are organized around knowledge units (KUs) and focus areas. Programs applying for their designation must map their course offerings to the KU topics and outcomes.

Applicant institutions have the option of applying for one or more "CAE-CD Specialization" designations for their programs, with each specialization having its own subset of KUs (technical core, nontechnical core, and optional) that must be met. At present, none of CAE's KUs or Specializations focus on cyber intelligence as we have defined it here, but the framework for doing so exists.[6] The KUs all follow a basic model structure: Name, Description, Outcomes, KU Topics, Vocabulary, and NICE Framework categories. Some also include reference to "Related Knowledge Units" and a list of "Specializations" that use that particular KU.

We recommend that NSA/DHS CAE Program add two optional KUs to the current list—one for intelligence analysis (to cover analytic competencies) and another for knowledge management (to cover the Informatics competencies)—and develop a specialization for cyber intelligence in the CAE-CD framework. A draft of these two KUs is outlined next as a starting point for further development.

Intelligence Analysis

Description. The intent of this KU is to provide students with sufficient understanding of strategy, critical and systems thinking, intelligence research, decision making, hypothesis generation and testing, and intelligence analytic methodologies and tradecraft such that they can identify, analyze, and communicate about the capabilities, intentions, and activities of a potential adversary or competitor in the cyber domain.

Outcomes

- Students will be able to use technologies, methods, and tradecraft to retrieve, aggregate, and organize information and to develop and evaluate new knowledge.
- Students will be able to analyze data and apply information to an organization's mission and strategic objectives by generating and analyzing courses of action.
- Students will be able to express clearly reasoned opinions and communicate effectively in writing, oral presentation, and visual display.
- Students will be able to recognize and apply ethical and professional standards in choosing and communicating about their analytic methodology.

Topics

- foundations of strategy
- critical and systems thinking
- problem definition, scope management, and problem solving
- judgment and decision making
- hypothesis generation and testing
- qualitative and/or quantitative analytic methodologies
- structured intelligence analytic techniques
- analytic communication
- ethics and standards in intelligence analysis.

Vocabulary

- intelligence, counterintelligence, intelligence analysis, intelligence cycle, intelligence discipline, structured analytic techniques, inductive reasoning, deductive reasoning, analytic tradecraft, probability, strategy, systems thinking.

NICE Framework Categories

- Primary: Analyze

Specializations

- Cyber Intelligence.

Knowledge Management

Description. The intent of this KU is to provide students with the ability to plan and organize information collection (collection management), develop and apply tools to gather and support complex data and information analysis from heterogeneous sources, adapt how data are visually presented to maximize understanding, and to understand, use, and evaluate information storage and retrieval systems.

Outcomes

- Students will be able to plan and organize information collection by developing a collection management plan.
- Students will be able to demonstrate advanced use of search engines and to use specialized tools for data and information analysis.
- Students will be able to adapt the presentation of data (including in visual form) to maximize understanding for the audience/decision maker.

Topics

- foundations of informatics
- information needs and information seeking
- information access and retrieval
- advanced search strategies
- knowledge evaluation and organization
- collection management
- open source collection using online tools
- data and information analysis tools
- information analytics
- data visualization.

Vocabulary

- analytics, collection management, competitive advantage, critical knowledge function, expectational knowledge, information management, information requirements/needs, knowledge analysis, situational awareness.

NICE Framework Categories

- Primary: Analyze.

Specializations

- Cyber Intelligence.

The focus area definition would be "KUs necessary to impart the necessary skills and abilities for assessing, analyzing, and communicating the capabilities, intentions, and activities—technical and otherwise—of potential adversaries and competitors in the cyber domain."

To form the new specialization in cyber intelligence, these two new KUs—Intelligence Analysis and Knowledge Management—would be combined with the foundational KUs (Cybersecurity Foundations, Cybersecurity Principles, and IT System Components) and the following existing KUs:

- Basic Networking
- Network Defense
- Cyber Threats
- Basic Cyber Operations
- Security Risk Analysis
- Data Administration
- Intelligence Analysis
- Knowledge Management.

In addition to creating an architecture for both foundational and specialized training and education, efforts should be made to integrate cyber intelligence into cybersecurity/cyber defense education at all levels. Cyber intelligence should be addressed, for example, in discussions of threat assessment and risk management, encouraging the use of a risk-based, intelligence-driven approaches to information and system security, as well as other, broader national security objectives. General discussions of cyber intelligence should include definitions; scope and sources of information (including nontechnical information from "beyond the network"); collection management; threat analysis to include adversary intentions, capabilities, and activity; and alignment of threat information with asset valuation for effective security planning.

Intelligence (and counterintelligence as a critical subset) is an essential component of cybersecurity and cyber defense and an important area for workforce development activity in the cyber realm. While numerous, sometimes overlapping efforts have sought to establish educational standards and workforce competencies for the various cybersecurity professions, less attention has been given to the emerging field of cyber intelligence and the knowledge and skills needed to prepare individuals and teams to perform this critical function. Cyber intelligence requires a blend of technical expertise and classic analytic tradecraft, and it is the integration of these hard and soft competencies that fundamentally distinguishes cyber intelligence from other cyber specialty areas.

This article draws on existing efforts to outline five clusters of competencies for the discipline of cyber intelligence: technical, analytic, knowledge management (informatics), contextual domain, and communication and organizational. But this rationally derived framework is just a start. It is necessary to validate the framework, and that effort is currently underway. Researchers at the University of South Florida have recently surveyed cyber intelligence practitioners, educators, and thought leaders to explore the dual-track model and to discern the importance and level of proficiency needed for each knowledge element to vet and validate the field's body of knowledge

The newly proposed KUs—Intelligence Analysis and Knowledge Management—for the CAE-CD curriculum structure can accelerate cyber intelligence's integration into cybersecurity education and practice and create a foundation for ultimately developing standards for the cyber intelligence discipline.

ACKNOWLEDGMENTS

This article was adapted, in part, from a white paper titled, "Cyber Intelligence Training and Education: Preparing Today's Talent for Tomorrow's Threats," prepared for the Intelligence and National Security Alliance (INSA). We gratefully acknowledge the assistance of our INSA editing team: Chuck Alsup, Ron Carback, Adam C. De Millio, Joe DeTrani, John Felker, Bob Gourley, Geoff Hancock, Mark Jarek, Joe Mazzafro, Ryan Pretzer, and Jerome Webb.

REFERENCES

1. M. Ludwick, T. Townsend, and J. P. Downing, "*White Paper: CITP training and education*," Software Engineering Inst., Carnegie Mellon Univ., Pittsburgh, White Paper, Sept. 2013, p. 11.
2. S. Donovan, B. Cobert, M. Daniel, and T. Scott, "*Strengthening the federal cybersecurity workforce*,"

The White House, Washington, D.C., 2016. Accessed on: Nov. 1, 2016. [Online]. Available: https://www.whitehouse.gov/blog/2016/07/12/strengthening-federal-cybersecurity-workforce

3. W. Newhouse, S. Keith, B. Scribner, and G. Witte, "*National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework*," NIST Special Publication 800-181, Gaithersburg, MD, Aug. 2017. doi: 10.6028/NIST.SP.800-181.
4. "*NICE Cybersecurity Workforce Framework online database version*," NIST, Gaithersburg, MD. Accessed on: May 21, 2019. [Online]. Available: https://niccs.us-cert.gov/workforce-development/cyber-security-workforce-framework
5. Homeland Security Advisory Council, "*CyberSkills Task Force report*," Department of Homeland Security, Washington, D.C., 2012, pp. 7–8.
6. R. Borum and R. Sanders, "*Cyber intelligence: Preparing today's talent for tomorrow's threats*," Intelligence and National Security Alliance (INSA), Washington, D.C., Sept. 2013, pp. 10–11.
7. Office of the Director of National Intelligence, "*National Intelligence Strategy of the United States of America, 2019*," Office of the Director of National Intelligence, Washington, D.C., 2019, p. 11.
8. W. Bandara, P. Harmon, and M. Rosemann, "Professionalizing business process management: Towards a body of knowledge for BPM," in *Business Process Management Workshops*. Germany: Springer-Verlag, 2011, pp. 759–774.

RANDY BORUM is the director of intelligence studies in the School of Information, University of South Florida. Contact him at borum@usf.edu.

RON SANDERS is the staff director for Cyber Florida, University of South Florida. Contact him at rpsanders@usf.edu.

EDITOR: James Bret Michael, IEEE Senior Member, bmichael@nps.edu

This article originally appeared in
Computer
vol. 53, no. 10, 2020

DEPARTMENT: CYBERTRUST

Security Awareness Training for the Workforce: Moving Beyond “Check-the-Box” Compliance

Julie Haney, *National Institute of Standards and Technology*

Wayne Lutters, *University of Maryland*

Security awareness training requirements set a minimum baseline for introducing security practices to an organization’s workforce. But is simple compliance enough to result in behavior change?

Given the high stakes and rapidly changing threat landscape of cybersecurity today, orienting an entire organization toward security practices is an important, but nontrivial, undertaking. A starting point is security awareness training, which is twofold: awareness seeks to change organizational attitudes, while training gives employees the skills and tools to practice good security hygiene.[1]

It is common for various public and private industry sectors to mandate security awareness training for their workforce. For example, within the United States, the Gramm–Leach–Bliley Act for the financial sector and the Federal Information Security Modernization Act of 2014 for federal agencies mandate security awareness training. The European Union’s General Data Protection Regulation requires organizations to provide similar training. A good example from the private sector is the Payment Card Industry (PCI) Security Standards Council, which makes awareness training available to people who must comply with the PCI Data Security Standard.

These mandates, policies, and standards establish a measurable, minimum baseline. The hope is that compliance with these training requirements will result in long-term positive impacts on security behaviors, such as securely handling cardholders’ data in the case of PCI, thus improving the overall security posture of organizations. But does compliance-based training live up to its promise?

WHEN COMPLIANCE IS NOT ENOUGH

Despite its noble intent, security awareness training can develop a bad reputation within an organization, sometimes with good reason. Training may be stereotypically boring, such as “death by PowerPoint” presentations and their corresponding computer-based quizzes, with the same generic content year after year. Furthermore, this annual refresh is likely susceptible to diminishing learning effects when not regularly reinforced with practice.

In addition to the lackluster way in which training is presented, those tasked with managing it may be at a disadvantage. Often plucked straight from the ranks of a firm’s security professionals with little understanding of “people issues,” they may be given insufficient guidance or resources to perform these additional duties.[2]

Most importantly, the training’s impact may never really be fully known. Some organizations view training simply as a “check-the-box” exercise, measuring success solely by training completion rates. However, this reveals little about how effective the training is in changing and sustaining attitudes and behaviors.

TAKING IT TO THE NEXT LEVEL

Although security awareness training may get a bad rap, this type of compliance activity can be beneficial

Digital Object Identifier 10.1109/MC.2020.3001959
Date of current version: 5 October 2020

as it ensures the workforce is at least exposed to security concepts and practices. But recognizing that training compliance sets just a minimum bar, what are the next, evolutionary steps?

We suggest that the goal of security awareness training should never be just to check the box but rather to move employees toward intrinsic motivation, where they see the value of security, develop the curiosity to learn more on their own, feel a sense of ownership and empowerment, want to do the right thing, and as a result, actually practice good behaviors. If resources allow, consider bringing in an outside consulting firm with expertise in security awareness training to help your program progress and measure success. Otherwise, there is plenty your own security awareness team can do. Based on best practices gleaned from seasoned security awareness professionals and our prior research studying security advocacy and awareness,[3] we offer the following suggestions to prompt organizations to rethink existing security awareness training and take their programs to the next level.

Become an advocate

Instead of viewing the security awareness team as merely compliance managers, consider that their primary job is advocacy—promoting and facilitating an understanding of security considerations and the adoption of security best practices. Security advocacy necessitates a different set of competencies beyond the technical skills possessed by most security professionals. Nontechnical competencies, such as interpersonal skills, communication skills, an appreciation of the audience, a customer-service orientation, and boundless creativity, may be essential for this role.[4] The security awareness team should also have a keen sense of the organization and its workforce, such as its goals, culture, constraints, and skill levels. Armed with this contextual knowledge, advocates then need to tailor security awareness communications and translate technical concepts into a language best understood by the workforce. This may require different messaging to the various roles within the organization. A usable security advocate, talking about what it takes to do security advocacy well, stated the following: "If you're a computer scientist, and all you know is the computer science, and you don't have the empathy, you don't have the skills to listen,...you don't have that psychological side, I don't think you can make it work."

INSTEAD OF VIEWING THE SECURITY AWARENESS TEAM AS MERELY COMPLIANCE MANAGERS, CONSIDER THAT THEIR PRIMARY JOB IS ADVOCACY–PROMOTING AND FACILITATING AN UNDERSTANDING OF SECURITY CONSIDERATIONS AND THE ADOPTION OF SECURITY BEST PRACTICES.

Building a multidisciplinary team can be particularly valuable. In addition to those who understand the technology, programs should also leverage the talents of individuals possessing much-needed skills in communications, marketing, behavior change, event planning, and graphic design. In addition, if resources are an issue, consider having liaisons who are members of different organizational groups to serve as extensions to the team.

Make security relatable

Employees need a reason to care about security. Training should communicate the business value of security best practices to the organization: how it enables the mission, ensures revenue, or protects assets and reputation. Framing the importance of security awareness training in business terms can be especially important for gaining management buy-in. But perhaps most importantly, people will be more apt to thoughtfully make security decisions when they have a sense of personal responsibility and view security as relevant to their day-to-day lives. Therefore, security awareness training should show the linkage between security and the duties of all roles in the organization, from frontline staff to senior executives.

Security communications should be topical, for example, related to contemporary topics in the news, pressing organizational issues, or seasonal activities. For instance, one federal agency has a December training event that educates the workforce on holiday-relevant topics, such as safe gift shopping, including security and privacy considerations for smart-home devices, interactive toys, and fitness trackers.

> *SECURITY COMMUNICATIONS SHOULD BE TOPICAL, FOR EXAMPLE, RELATED TO CONTEMPORARY TOPICS IN THE NEWS, PRESSING ORGANIZATIONAL ISSUES, OR SEASONAL ACTIVITIES.*

Another recent trend in security awareness training is an increasing emphasis on the work–home connection. With more employees teleworking in some capacity, they need to stay vigilant no matter their location. Good security habits are easier to form when they permeate someone's entire life and do not just end when they leave the office or turn off their work computer. To highlight this connection, in addition to topics of interest related to secure work habits, consider providing information employees can take home to educate their families, such as details on the secure use of smartphone apps and social media.

Get their attention

To engage the workforce and reinforce training concepts, the security awareness team should go beyond the typical, once-and-done canned presentations to disseminate security information using a variety of communication channels and techniques periodically throughout the year. For example, we have seen organizations bring in high-quality speakers for security day events, produce concise handouts with security tips, hold security information fairs, create visually appealing posters, and enable remote broadcasting for those who cannot attend training events in person. Whatever the media, the approach should pique interest while being mindful of employee limitations on time, interest, and skill level. A security awareness professional stated, "You want to just put a different spin on it because people just see stuff all the time: 'Have a good password. Lock your computer'…Be creative and think outside the box."

Security awareness training should ideally be tailored to the local culture of the organization, memorable, and entertaining when appropriate. In our studies, we have come across numerous examples of creative approaches: a security-themed food truck event, complete with security trivia games while patrons wait in line to order; security-themed coloring books and calendars; a Shakespeare-themed play titled "To Send or Not to Send" that educated employees about proper email use; and a late-night show parody with a cyber-themed comedic monologue and guests who talked about security topics. Employing a variety of communication methods provides something for everyone because employees will have different preferences on how they receive and best retain security awareness information.

Empower them with tools

Raising awareness of security threats is important, but it does not necessarily lead to behavior change. Doing so without advice or the appropriate tools on how to confront those threats may leave employees feeling anxious, unsatisfied, and powerless. Therefore, employees should be provided with practical, prioritized, and actionable steps they can take to protect themselves and their organization.

When providing recommendations, meet people where they are. Training topics should include recommendations that are achievable given employees' skill sets, described in terms they understand, and accompanied by pointers to helpful resources. Remember that "perfect security" is an impossible goal. Security is more of a journey, so start off by giving employees small steps they can immediately implement that have a large impact.

Measure impact

Once this evolved security awareness training is in place, you need to determine whether it actually makes a difference. Compliance metrics are easy to collect and analyze, but they are only part of the story. Unfortunately, it can be difficult to develop meaningful measures of aspects that matter most to your organization. To get started, we suggest a few approaches that other organizations have found helpful.

If in-person or remote training events are held, attendance can be an indicator of reach. But be careful to focus not just on the numbers; also look at who is attending. This can lend insight into those buying into the importance of the training, whether your program is reaching the right people, and where additional effort should be focused.

Employee feedback is another way in which you can quantitatively and qualitatively assess the effectiveness of your program. Informal break room conversations are valuable, but anonymous, postevent surveys reach a broader audience and provide more structured, honest data. However, it is important to note that response rates can be low and are often subject to self-selection bias. For instance, only those with strong positive or negative opinions may respond. Still, these can help the team gauge overall satisfaction, track perceived takeaways, and identify suggestions for future topics or formats.

But perhaps the most telling measure of effectiveness comes in the form of trends in user-generated security incident data aggregated from multiple sources. For example, after security awareness training regarding the sending of sensitive information via email, are the number of personal data disclosures going down? Don't just focus on where employees fall short; look at indicators of positive behaviors as well, such as increased reporting of suspicious emails or other security incidents to the help desk.

This holistic approach requires two-way communication with the cybersecurity and incident handling arms of the organization. Collaboration with physical security staff can likewise offer interesting insights into security mechanisms that have physical components, for example, smart cards used for both facility access and computer login. These partnerships have an added benefit: common threats to the organization observed by other security groups can help inform areas that may warrant additional workforce awareness and training. A security awareness team lead stated the following: "We're trying to…be able to tie in together the people who take their training to the people who get caught with phishing exercises…with people who are losing their badges to people who send out information they shouldn't to see what's the correlation here. Are these people just too busy? Are they not paying attention? Is there a training problem?"

Finally, because we all know that statistics can be misleading, be sure to contextualize the data you collect and consider possible explanations with targeted solutions. For instance, if a particular department within the organization is more susceptible to certain security threats, how can that population be better trained? If click rates (the number of people falling for a simulated phish) for phishing training exercises go up one month, were the phishing emails more sophisticated than usual, or was the email premise more aligned with the functions of the organization?

Be positive and constructive

At some point, all employees will have a security slipup, and they usually hear about it. But what about commending them when they do something good? The threat of negative consequences has been found to have a limited impact on decisions to implement security,[5] but positive and constructive feedback can be effective in encouraging and maintaining desired behaviors.[6] To better incentivize employees to learn from their slipups, take an educational rather than a punitive approach when something goes wrong. Also, try to recognize employees who make good security decisions, for example, those who report suspicious emails or promote security best practices to their colleagues. Recognition doesn't have to be anything big. Sometimes a simple, but personal, "thank you" can be enough.

Strive toward continuous improvement

You will likely not get security awareness training "right" from the start. Therefore, you should commit to improving the program incrementally over time. To ensure training stays fresh and keeps up with relevant threats to the organization, consider regular updates to your training material. This includes not just changes to topics but also letting measures of effectiveness inform any necessary adaptations of communication channels to better accommodate employees and reach broader populations within the organization.

Finally, an organization's security awareness program should not be an island. Learn from others. Talk to security awareness professionals working in similar organizations about what works for them. Consider participation in online security awareness communities, which can be a wellspring of valuable resources (e.g., SANS,[7] EDUCAUSE,[8] and National Cyber Security Alliance[9]). Ideas can also be found at events that are focused on or have tracks related to security awareness training, for example, the annual Federal Information Security Educators[10] or RSA[11] conferences.

For cybertrust, there are baseline benefits of mandated security awareness training. However, organizations should be cautious about the potential pitfalls of slipping into a strict compliance mentality. Compliance metrics do not tell the whole story and fail to measure the effectiveness of the program in a sustained change in employee attitudes and behaviors. While compliance-based training is a start, security awareness programs should strive to go beyond, engaging and empowering employees to be informed, responsible cybercitizens in and outside of work.

DISCLAIMER

Any mention of commercial products or reference to commercial organizations is for information only. It does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products mentioned are necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of their employers. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotations thereon.

REFERENCES

1. *"Awareness, training, and education controls,"* NIST, Gaithersburg, MD. Accessed: Aug. 19, 2020. [Online]. Available: https://csrc.nist.gov/glossary/term/Awareness_Training_and_Education_Controls
2. *"2018 SANS security awareness report: Building successful security awareness programs,"* SANS, Bethesda, MD, 2018. [Online]. Available: https://www.sans.org/security-awareness-training/reports/2018-security-awareness-report
3. J. M. Haney and W. G. Lutters, "It's scary … It's confusing … It's dull: How cybersecurity advocates overcome negative perceptions of security," in *Proc. Symp. Usable Privacy and Security*, 2018, pp. 411–425. doi: 10.5555/3291228.3291261.
4. B. Woelk, *"The successful security awareness professional: Foundational skills and continuing education strategies,"* ECAR, Louisville, CO, 2016. [Online]. Available: https://library.educause.edu/~/media/files/library/2016/8/erb1608.pdf
5. H. S. Rhee, C. Kim, and Y. U. Ryu, "Self-efficacy in information security: It's influence on end users' information security practice behavior," *Comput. Secur.*, vol. 28, no. 8, pp. 816–826, 2009. doi: 10.1016/j.cose.2009.05.008.
6. C. Hadnagy and M. Fincher, *Phishing Dark Waters: The Offensive and Defensive Sides of Malicious Emails*. Hoboken, NJ: Wiley, 2015.
7. *"Resources,"* SANS, Bethesda, MD. Accessed: Aug. 19, 2020. [Online]. Available: https://www.sans.org/security-awareness-training/resources
8. *"Awareness campaigns,"* Educause, Louisville, CO. Accessed: Aug. 19, 2020. [Online]. Available: https://www.educause.edu/focus-areas-and-initiatives/policy-and-security/cybersecurity-program/awareness-campaigns
9. *"Resources,"* STOP. THINK. CONNECT., National Cyber Security Alliance, Cambridge, MA. Accessed: Aug. 19, 2020. [Online]. Available: https://www.stopthinkconnect.org/resources
10. *"FISSEA—Federal information security educators,"* NIST, Gaithersburg, MD. [Online]. Available: https://csrc.nist.gov/projects/fissea
11. RSA Conference. Accessed: Aug. 19, 2020. [Online]. Available: https://www.rsaconference.com/

JULIE HANEY is a computer scientist and usable security researcher at the National Institute of Standards and Technology. Contact her at julie.haney@nist.gov.

WAYNE LUTTERS is an associate professor in the College of Information Studies at University of Maryland. Contact him at lutters@umd.edu.

IEEE COMPUTER SOCIETY D&I FUND

Drive Diversity & Inclusion in Computing

Supporting projects and programs that positively impact diversity, equity, and inclusion throughout the computing community.

Do you have a great idea for new programs that will positively impact diversity, equity, and inclusion throughout the computing community?

The IEEE Computer Society Diversity & Inclusion Committee seeks proposals for projects, programs, and events that further its mission. New programs that deliver education, outreach, and support, including, but not limited to, mentoring programs at conferences, panel discussions, and webinars, are welcomed.

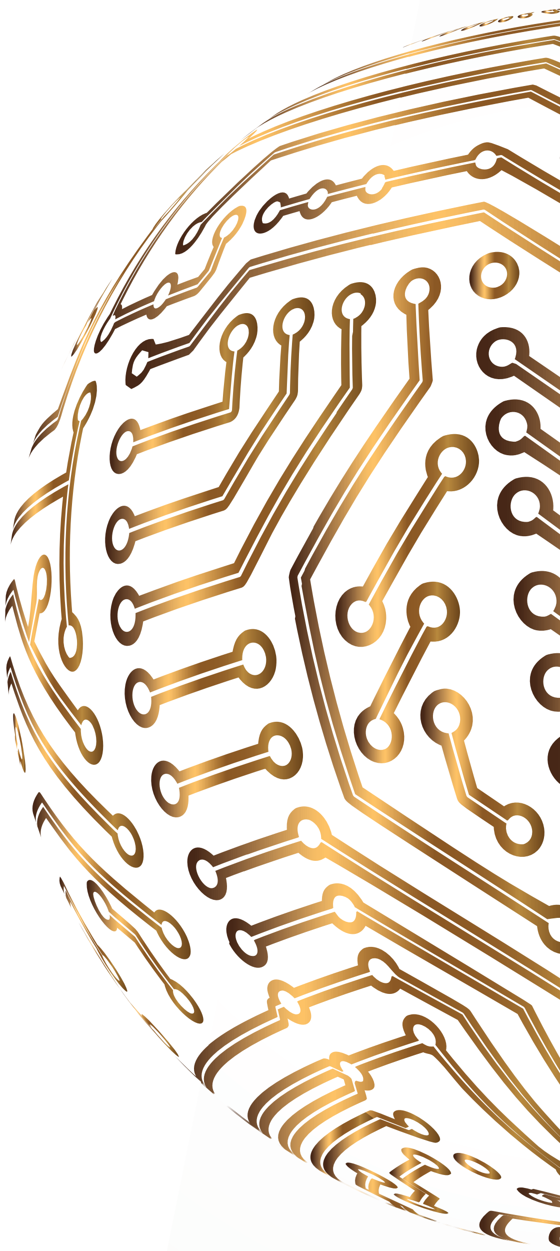
Help propel the Computer Society's D&I programs—submit a proposal today!

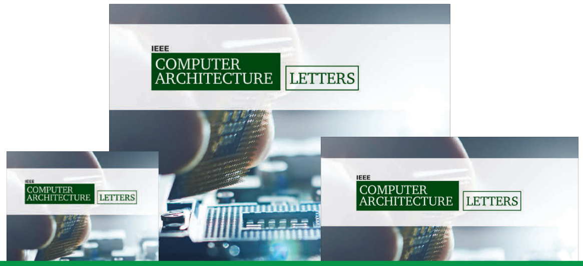
https://bit.ly/CS-Diversity-CFP

Donations to the IEEE Computer Society D&I Fund are welcome!

IEEE COMPUTER SOCIETY

IEEE Foundation




IEEE
COMPUTER ARCHITECTURE
LETTERS

E-Currency and Fairness • Ransomware Defense • A National Cybersecurity Policy
IEEE SECURITY & PRIVACY
SECURITY & PRIVACY
SECURITY & PRIVACY
Privacy and Automated Airport Screening
SECURITY & PRIVACY
CYBERSECURITY AND PRIVACY ISSUES IN BRAZIL

Conference Calendar

IEEE Computer Society conferences are valuable forums for learning on broad and dynamically shifting topics from within the computing profession. With over 200 conferences featuring leading experts and thought leaders, we have an event that is right for you. Questions? Contact conferences@computer.org.

MARCH

12 March
- ICSA (IEEE Int'l Conf. on Software Architecture), Honolulu, USA
- VR (IEEE Conf. on Virtual Reality and 3D User Interfaces), Christchurch, New Zealand

14 March
- DATE (Design, Automation & Test in Europe Conf. & Exhibition), Antwerp, Belgium

15 March
- CSASE (Int'l Conf. on Computer Science and Software Eng.), Duhok, Iraq
- SANER (IEEE Int'l Conf. on Software Analysis, Evolution, and Reengineering), Honolulu, USA

21 March
- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Pisa, Italy

30 March
- WONS (Wireless On-Demand Network Systems and Services Conf.), Oppdal, Norway

APRIL

2 April
- HPCA (IEEE Int'l Symposium on High-Performance Computer Architecture), Seoul, South Korea

4 April
- ICST (IEEE Int'l Conf. on Software Testing, Verification and Validation), virtual

11 April
- PacificVis (IEEE Pacific Visualization Symposium), Tsukuba, Japan

25 April
- VTS (IEEE VLSI Test Symposium), San Diego, USA

MAY

4 May
- ICCPS (ACM/IEEE Int'l Conf. on Cyber-Physical Systems), Milano, Italy
- RTAS (IEEE Real-Time and Embedded Technology and Applications Symposium), Milano, Italy

9 May
- ICDE (IEEE Int'l Conf. on Data Eng.), virtual

15 May
- FCCM (IEEE Int'l Symposium on Field-Programmable Custom Computing Machines), New York, USA

16 May
- ICFEC (IEEE Int'l Conf. on Fog and Edge Computing), Messina, Italy

17 May
- ISORC (Int'l Symposium On Real-Time Distributed Computing), Västerås, Sweden

18 May
- ISCV (Int'l Conf. on Intelligent Systems and Computer Vision), Fez, Morocco
- ISMVL (IEEE Int'l Symposium on Multiple-Valued Logic), Dallas, USA

19 May
- AQTR (Int'l Conf. on Automation, Quality and Testing, Robotics), Cluj-Napoca, Romania
- SELSE (IEEE Workshop on Silicon Errors in Logic – System Effects), virtual

21 May
- ICSE (IEEE/ACM Int'l Conf. on Software Eng.), Pittsburgh, USA

22 May
- ISPASS (IEEE Int'l Symposium on Performance Analysis of Systems and Software), Singapore
- SP (IEEE Symposium on Security and Privacy), San Francisco, USA

23 May
- ETS (IEEE European Test Symposium), Barcelona, Spain
- SEAMS (Int'l Symposium on Software Eng. for Adaptive and Self-Managing Systems), Pittsburgh, USA

2469-7087/22 © 2022 IEEE

25 May

- SERA (IEEE/ACIS Int'l Conf. on Software Eng., Management and Applications), Las Vegas, USA

30 May

- DCOSS (Int'l Conf. on Distributed Computing in Sensor Systems), Los Angeles, USA
- IPDPS (IEEE Int'l Parallel & Distributed Processing Symposium), Lyon, France

JUNE

6 June

- EuroS&P (IEEE European Symposium on Security and Privacy), Genoa, Italy
- MDM (IEEE Int'l Conf. on Mobile Data Management), Paphos, Cyprus

11 June

- ISCA (ACM/IEEE Int'l Symposium on Computer Architecture), New York, USA

14 June

- WoWMoM (IEEE Int'l Symposium on a World of Wireless, Mobile and Multimedia Networks), Belfast, UK

19 June

- CVPR (IEEE/CVF Conf. on Computer Vision and Pattern Recognition), New Orleans, USA

25 June

- CSCLOUD (IEEE Int'l Conf. on Cyber Security and Cloud Computing), Xi'an, China

26 June

- ICIS (IEEE/ACIS Int'l Conf. on Computer and Information Science), Zhuhai, China

27 June

- COMPSAC (IEEE Computers, Software, and Applications Conf.), Torino, Italy
- DSN (IEEE/IFIP Int'l Conf. on Dependable Systems and Networks), Baltimore, USA
- HOST (IEEE Int'l Symposium on Hardware Oriented Security and Trust), McLean, Virginia, USA

JULY

1 July

- ICALT (IEEE Int'l Conf. on Advanced Learning Technologies), Bucharest, Romania

6 July

- ISVLSI (IEEE Computer Society Symposium on VLSI), Nicosia, Cyprus

10 July

- ICDCS (IEEE Int'l Conf. on Distributed Computing Systems), Bologna, Italy

11 July

- ICME (IEEE Int'l Conf. on Multimedia and Expo), Taipei, Taiwan

21 July

- CBMS (IEEE Int'l Symposium on Computer-Based Medical Systems), Shenzhen, China

AUGUST

1 August

- ICCP (Int'l Conf. on Computational Photography), Pasadena, USA

2 August

- MIPR (IEEE Int'l Conf. on Multimedia Information Processing and Retrieval), virtual

4 August

- BCD (IEEE/ACIS Int'l Conf. on Big Data, Cloud Computing, and Data Science Eng.), Danang, Vietnam

7 August

- CSF (IEEE Computer Security Foundations Symposium), Haifa, Israel

9 August

- IRI (IEEE Int'l Conf. on Information Reuse and Integration for Data Science), virtual

15 August

- RE (IEEE Int'l Requirements Eng. Conf.), Melbourne, Australia

SEPTEMBER

6 September

- CLUSTER (IEEE Int'l Conf. on Cluster Computing), Heidelberg, Germany

CALL FOR SPECIAL ISSUE PROPOSALS

Computer solicits special issue proposals from leaders and experts from a broad range of computing communities. Proposed themes/issues should address timely, emerging topics that will be of broad interest to *Computer*'s readership. Special issues are an important component of *Computer*, as they deliver essential research insights and well-developed perspectives on new and established technologies and computing strategies.

We encourage submissions of high-quality proposals for the 2023 editorial calendar. Of particular interest are proposals centered on:

- offsite educational and business continuity technology challenges,
- privacy related to personal location tracking and surveillance (digital and physical),
- artificial intelligence and machine learning,
- technology's role in disrupted supply chains,
- misinformation and disinformation (fake information—malicious or non-malicious), and
- cyberwarfare/cyberterrorism

Proposal guidelines are available at:
www.computer.org/csdl/magazine/co/write-for-us/15911

IEEE COMPUTER SOCIETY

IEEE