Augmenting Cognition via Edge Computing

Mahadev Satyanarayanan
Carnegie Mellon University

Nigel Davies
Lancaster University

***Abstract*—**

Augmented cognition can transform human capabilities, but delivering its benefits in real-time will require low-latency wireless access to powerful infrastructure resources from lightweight wearable devices. Edge computing is the only viable approach to meeting these stringent requirements. In this paper, we explore the symbiotic relationship between augmented cognition and edge computing. We show how off-the-shelf wearable hardware, standard AI technologies such as computer vision, and edge computing can be combined to create a system that is much greater than the sum of its parts. Augmenting human cognition thus emerges as a prime example of a new class of *edge-native applications* that can become "killer apps" for edge computing.

I. Introduction

Viewed as autonomous mobile computing systems with builtin sensing, processing and persistent storage, humans are the result of over a billion years of evolution. Our chances of improving upon nature in a short time (say 10 years) are negligible if we are bound by the same rules as biological evolution. However, we have a unique opportunity that is not available to nature: namely, *to amplify human cognition in real time through low-latency wireless access to infrastructure resources.* These resources can be larger, heavier, more energy-hungry and more heat-dissipative than could ever be carried or worn by a human user. Distributed sensing can also offer real-time inputs from vantage points other than the first-person viewpoint of a human. By seamlessly integrating these resources with human perception and cognition, we could achieve a whole that is much greater than the sum of parts.

This vision was first articulated in 2004 [19], but only now have the necessary building blocks reached a level of maturity that they can be viewed as "off-the-shelf technologies". These include wearable computers with rich arrays of sensors (such as video cameras, microphones, accelerometers, and gyroscopes) and cognitive algorithms based on deep neural networks (DNNs) for computer vision, speech recognition, and natural language processing that have now reached near-human levels of accuracy. A further crucial building block is the ability to wirelessly access cloud-like computing resources at such low end-to-end latency and high bandwidth that we are able to seamlessly integrate them into the "inner loop" of human cognition. This is the essence of *edge computing,* which is emerging as a new disruptive force [21], [20].

In this paper, we share the experience and insights that we have gained so far from exploring two distinct ways of augmenting human cognition:

- The first is providing just-in-time guidance and error detection for a user who is performing an unfamiliar task. Prompt detection of errors can be valuable even on familiar tasks, when the user is working under conditions of fatigue, stress, or cognitive overload. Informally, this is like having "an angel on your shoulder" [18].
- The second is amplifying the bandwidth and fidelity of the long-term persistent memory of a human user. Human memory is notoriously fallible but contemporary psychology theories suggest that traces captured and displayed using pervasive devices can be used to both reinforce and attenuate human memories, opening up the possibility of a very wide range of new applications for memory augmentation devices [7].

Using our insights from these two styles of augmentation, we seek to lay the foundations for edge-based augmented processing, storage, and retrieval in humans. Our work spans a wide swath of computer science, including operating systems, wireless networks, computer vision, human-computer interaction, augmented reality, data science, and health systems. In contrast to replacing the human, which is the goal of classic artificial intelligence (AI), our goal is to enhance and extend the capabilities of a human.

II. Why Edge Computing Is Essential

Human performance on cognitive tasks is remarkably fast and accurate. For example, face recognition takes between 370 and 620 ms, depending on familiarity [17]. Speech recognition takes 300 to 450 ms for short phrases, and only 4 ms to tell that a sound is a human voice [1]. Virtual reality applications that use head-tracked systems require latencies less than 16 ms to achieve perceptual stability [10]. Humans are acutely sensitive to delays in the critical path of interaction. This is apparent to anyone who has used a geosynchronous satellite link for a telephone call. The nearly 500 ms round-trip delay is distracting, and leads to frequent conversational errors.

Cognitive augmentation requires sensing to be superhuman in speed, without loss of accuracy. Only then will there be time left within a very tight budget for additional processing to provide augmentation. An end-to-end latency target of a few tens of milliseconds is a safe and conservative goal, with 10 ms as the ideal. Larger delays may distract and annoy a mobile user who is already attention challenged. Since jitter is also annoying and distracting, it is important to avoid long-tailed distributions of end-to-end latency.

The most accurate cognitive algorithms are typically processing- and memory-intensive. Their execution speed on mobile devices tends to be slow, relative to execution on a server. Figure 1 illustrates this point with 2018 data from Wang et al [23], corroborating 2013 results from Ha et al [13]

	MobileNet (milliseconds)	ResNet (milliseconds)
Nexus 6 smartphone	353 (67)	983 (141)
NVIDIA Jetson TX2	13 (0)	92 (2)
Rack-mounted Server	4 (0)	33 (0)

Source: Adapted from Figure 3 of Wang et al [23]

Times above are per-image, averaged across 100 random images. Numbers in parentheses are standard deviations. Full experimental details can be found in the source document [23]. MobileNet is a DNN that is optimized for mobile devices. It has a smaller memory footprint and processing demand than ResNet, but is less accurate.

Fig. 1. Inference Speed on Image Classification Task

Year	Typical Server Processor	Typical Server Speed	Typical Mobile Device Device	Typical Mobile Device Speed
1997	Pentium II	266 MHz	Palm Pilot	16 MHz
2002	Itanium	1 GHz	Blackberry 5810	133 MHz
2007	Intel Core 2	9.6 GHz (4 cores)	Apple iPhone	412 MHz
2011	Intel Xeon X5	32 GHz (2x6 cores)	Samsung Galaxy S2	2.4 GHz (2 cores)
2013	Intel Xeon E5-2697v2	64 GHz (2x12 cores)	Samsung Galaxy S4	6.4 GHz (4 cores)
			Google Glass	2.4 GHz (2 cores)
2016	Intel Xeon E5-2698v4	88.0 GHz (2x20 cores)	Samsung Galaxy S7	7.5 GHz (4 cores)
			HoloLens	4.16 GHz (4 cores)
2017	Intel Xeon Gold 6148	96.0 GHz (2x20 cores)	Pixel 2	9.4 GHz (4 cores)

Source: Adapted from Chen [3] and Flinn [12]
"Speed" metric = number of cores times per-core clock speed.

Fig. 2. Long-Term Impact of Mobility Constraints

and 2016 results from Hu et al [14]. Over a 20-year period from 1997 to 2017, mobile devices have consistently lagged far behind server hardware, as shown in Figure 2. This stubborn performance gap is due to the fact that mobile users value light weight, small size, long battery life, comfort and aesthetics, and tolerable heat dissipation over speed, memory size, or storage capacity. While mobile devices will improve in the future, so will server hardware — the gap will remain. One can view this gap as a "mobility penalty." It is the price one pays in performance for the benefit of mobility [22].

Wirelessly offloading compute-intensive operations to servers in the infrastructure helps to bridge the gap shown in Figure 2. However, using servers in the public cloud is unsatisfactory because the cloud is typically far away. The high level of consolidation necessary for economies of scale in cloud computing implies that there can only be a few large data centers worldwide. Li et al. [15] report that average RTT from 260 global vantage points to their optimal Amazon EC2 instances is nearly 74 ms. A wireless first hop would add to this amount. This makes it impossible to meet tight end-to-end latency goals of just a few tens of milliseconds.

Edge computing creates the illusion of "bringing the cloud

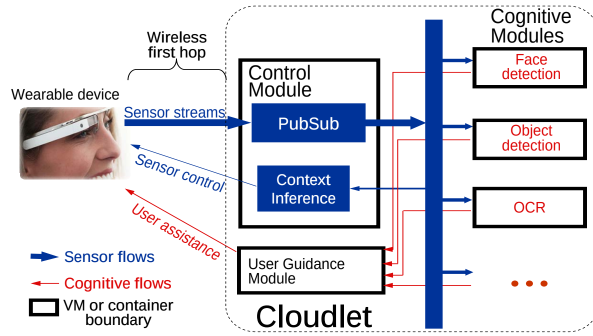


Fig. 3. Gabriel Architecture (Source: Adapted from Chen et al [4])

closer" [20]. Server hardware in edge computing is comparable to that in cloud computing, but engineered differently. Instead of extreme consolidation into a few large data centers, servers in edge computing are organized into small, dispersed data centers that we call *cloudlets*. A cloudlet can be viewed as "a data center in a box." By offloading to a cloudlet rather than to the cloud, a resource-challenged wearable device can simultaneously meet the goals of low end-to-end latency and resource-intensive processing. This is a crucial capability for augmenting cognition.

The next two sections illustrate how edge computing can be used to enable two different types of augmented cognition. Section III describes how the ease and accuracy of task performance can be improved, especially when a user is performing a task for the first time. Section IV describes how notoriously fallible human memory can be made more accurate. We hope that our success in these efforts will stimulate and encourage research on many other forms of augmented cognition.

III. Augmenting Task Performance

GPS navigation systems have transformed our driving experience. They guide you step-by-step to your destination, offering you helpful just-in-time voice guidance about upcoming actions that you need to take. If you make a mistake (e.g., miss an exit), this is promptly recognized and corrected. The difficult task of navigating an unfamilar city has been transformed into a trivial exercise in following directions.

Wearable Cognitive Assistance broadens the metaphor of GPS navigation. It can be viewed as an "angel on your shoulder" that silently observes what you are doing, and offers helpful hints just in time. This concept lies at the convergence of wearable computers, edge computing, and cognitive algorithms (such as computer vision, speech recognition, natural language understanding, and other derivatives of machine learning). The wearable device provides a first-person viewpoint of a user's task. Sensor streams from the device (such as video, audio, accelerometer and gyroscope) are transmitted over a wireless network to a nearby cloudlet for task-specific processing. The cognitive algorithms in this processing typically have memory, CPU and GPU demands that cannot be sustained on the wearable device. Based on inferred task state, guidance in visual, verbal or tactile form is generated, transmitted over the wireless network, and presented to the user on the wearable device.

Gabriel, shown in Figure 3, is an extensible PaaS (Platform as a Service) layer that we have created for Wearable Cognitive Assistance. The front-end on a wearable device performs preprocessing of sensor data (e.g. compression and encoding), and then streams it over a wireless network to a cloudlet. The Gabriel back-end on the cloudlet is organized as a collection of cognitive modules. The *control module* is the focal point for all interactions with the wearable device. A publish-subscribe (PubSub) mechanism decodes and distributes the incoming sensor streams to multiple *cognitive modules* (e.g., task-specific computer vision algorithms) for concurrent processing. Cognitive engine outputs are integrated by a task-specific *User Guidance module*. This code performs higher-level cognitive processing such as inferring task state and detecting errors. From time to time, it generates guidance.

On this platform, using a diversity of wearable devices such as Google Glass, Microsoft HoloLens, Vuzix Glass, and ODG R7, we have implemented over a dozen applications [4], [20]. Some of these are summarized in Figure 4. The implementations of these applications show considerable high-level similarity in terms of cloudlet workflow. This workflow has two major phases.

The first phase analyzes the current video frame to extract a *symbolic representation* of the current state of the task. This phase has to be tolerant of considerable real-world variation in the video frame due to variable lighting levels, varying light sources, varying positions of the viewer, task-unrelated clutter in the background, and so on. The symbolic representation is an idealized representation of the current task state that excludes all irrelevant detail. One can view this phase as a task-specific "analog-to-digital" conversion of an input video frame—the enormous state space of the input is simplified to the much smaller state space of the symbolic representation.

The second phase operates exclusively on the symbolic representation. It compares the symbolic representation to the expected task state to determine whether user guidance is needed, and if so what that guidance should be. The guidance may have video, static images, plain text and audio components that are streamed back to the wearable device for presentation to the user. Further details can be found in the paper by Chen et al [4], which also analyzes the sources of end-to-end latency in this class of applications.

Today, Gabriel applications depend entirely on first-person sensing from a user-worn device. No use is made of additional sensing from off-body viewpoints, which has been shown to be valuable from our memory augmentation work described in Section IV. Exploring how such additional sensing could help Gabriel applications is part of our future work.

IV. Augmenting Memory

In the previous section we have focused primarily on decision making—the human equivalent of processing. In this section we focus on augmenting human memory, the equivalent to upgrading the size, speed, and indexing of storage available to a computer. The importance of addressing our current inability to augment human memory cannot be overstated—for example, 47.5 million people worldwide are currently living with dementia. The loss of memory, and with it a sense of identity, is often cited as one of the most distressing aspects of the disease. Even for otherwise healthy individuals, memory augmentation offers the potential to deliver significant benefits in productivity and quality of life.

Technology has always had a direct impact on how and what humans remember. This impact is both inevitable and fundamental—technology radically changes the nature and scale of the cues that we can preserve outside our own memory in order to trigger recall. We have previously argued in [7] that recent developments in three separate strands of technology together enable entirely new ways of augmenting human memory:

- near-continuous collection of memory cues has become possible through the use of lifelogging technologies, social networks and interaction logs.
- advances in audio and image processing now enable widespread mining of stored cues for proactive presentation.
- the pervasive nature of displays (both mobile and fixed) provides many new opportunities for displaying memory cues to trigger recall.

These building blocks provide the foundation for a new technology eco-system that can transform the way humans remember in order to measurably and significantly improve functional capabilities while maintaining individual control. Example applications of memory augmentation include support for learning new skills, affecting behaviour change by helping users recall previous positive (or negative) experiences, and helping address many of the everyday cognitive failures we all experience [5].

Memory augmentation will obviously make use of mobile devices such as life-logging cameras. However, they are not sufficient and edge computing will be crucial in delivering the sensing, processing, and cuing required for effecting memory augmentation. For example, in terms of sensing, wearable devices (and their associated first-person views) have significant limitations as a platform for augmenting human cognition [6]. Lifelogging cameras are often difficult to place comfortably on the body whilst still maintaining clear and meaningful coverage of the environment (common problems include occlusion by hair/clothes and poor viewing angle). Moreover, these cameras are typically static and therefore capture a poor representation of what was actually "seen" at the time (see Figure 5). Furthermore, psychological literature indicates that despite seeing the "first person" view, individuals may experience detachment from their current perspective leading them to "see" things from the view of an onlooker. More significantly, such observer (third-person) views are a not uncommon feature of recalled memories [16]. Being able to capitalise on sensors in the environment offers a number of advantages for memory augmentation systems including

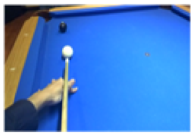
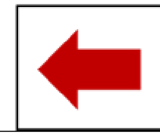
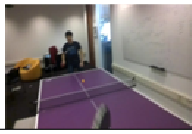
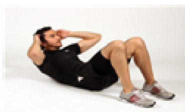
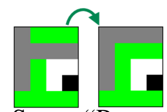
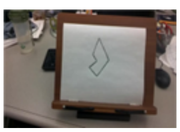
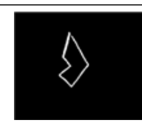
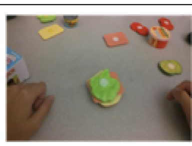
App Name	Example Input Video Frame	Description	Symbolic Representation	Example Guidance
Pool		Helps a novice pool player aim correctly. Gives continuous visual feedback (left arrow, right arrow, or thumbs up) as the user turns his cue stick. The symbolic representation describes the positions of the balls, target pocket, and the top and bottom of cue stick.	<Pocket, object ball, cue ball, cue top, cue bottom>	
Ping-pong		Tells novice to hit ball to the left or right, depending on which is more likely to beat opponent. Uses color, line and optical-flow based motion detection to detect ball, table, and opponent. Video URL: *https://youtu.be/_lp32sowyUA*	<InRally, ball position, opponent position>	Whispers “Left!”
Work-out		Counts out repetitions in physical exercises. Classification is done using Volumetric Template Matching on a 10-15 frame video segment. A poorly-performed repetition is classified as a distinct type of exercise (e.g. “good pushup” versus “bad pushup”).	<Action, count>	Says “8 ”
Face		Jogs your memory on a familiar face whose name you cannot recall. Detects and extracts a tightly-cropped image of each face, and then applies a state-of-art face recognizer. Whispers the name of the person recognized.	ASCII text of name	Whispers “Barack Obama”
Lego		Guides a user in assembling 2D Lego models. The symbolic representation is a matrix representing color for each brick. Video URL: *https://youtu.be/7L9U-n29abg*	[[0, 2, 1, 1], [0, 2, 1, 6], [2, 2, 2, 2]]	 Says “Put a 1x3 green piece on top”
Draw		Helps a user to sketch better. Builds on third-party app for desktops. Our implementation preserves the back-end logic. A new Glass-based front-end allows a user to use any drawing surface and instrument. Displays the error alignment in sketch on Glass. Video URL: *https://youtu.be/nuQpPtVJC6o*		
Sand-wich		Helps a cooking novice prepare sandwiches according to a recipe. Since real food is perishable, we use a food toy with plastic ingredients. Object detection uses faster-RCNN deep neural net approach. Video URL: *https://youtu.be/USakPP45WvM*	Object: “E.g. Lettuce on top of ham and bread”	 Says “Put a piece of bread on the lettuce”

Fig. 4. Example Wearable Cognitive Assistance Applications (Source: Adapted from [20]

access to improved quality data, professional maintenance of the sensor infrastructure, and more cost effective solutions as the cost of sensing can be shared across multiple users.

Edge resources will also be needed to support storage and processing needs of memory augmentation systems – it is simply not possible to store and process a lifetime's memories on a mobile device so cloud and edge support will be required. Similarly, while mobile devices such as Google Glass can be used to deliver memory cues future systems are likely to make use of the full device eco-system and present information via pervasive displays, audio devices and environmental control (e.g. stimulating recall by recreating the environmental context of a memory).

In our work we have begun to develop architectures for memory augmentation (see Figure 6). The architecture highlights the three distinct points of intervention for memory augmentation systems (encoding, rehearsal, and retrieval), a range of presentation options (spanning both mobile and infrastructure), and examples of sensing systems that provide the raw data from memory cues. For example, we have built systems that capture image data using lifelogging cameras (Narrative Clips), process these images to produce summaries that can then be presented to users via in-home ambient displays (supporting rehearsal). The same conceptual architecture

Fig. 5. Example of the same event photographed by a lifelogging device and an infrastructure camera [6]

has also been used to support a diverse set of applications including automatic summarisation of meetings and delivery of lecture summaries as students walk to lectures—both applications being designed to help users restore context before the next meeting/lecture.

Our architecture is underpinned by the notion of a "memory vault" in which a user's lifetime of memories reside. The architecture highlights that much of the functionality for memory augmentation is likely to reside at the edge – extensive use is made of infrastructure sensors and it is clearly not possible to store a user's entire memories on a single mobile device.

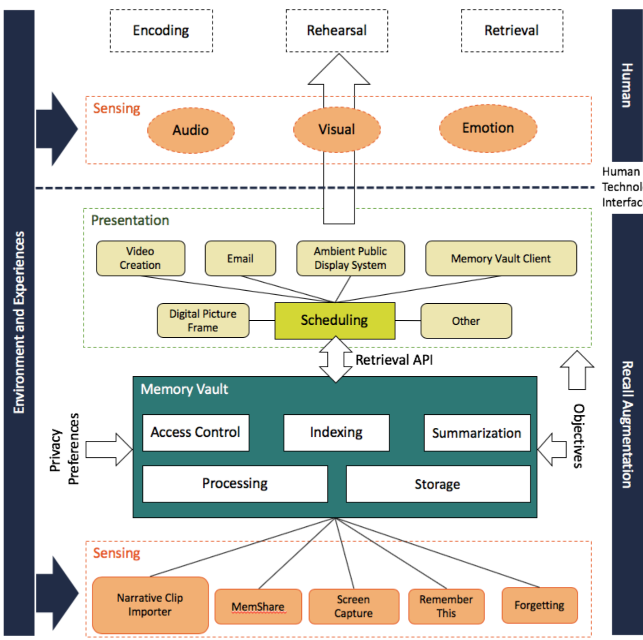


Fig. 6. The RECALL Memory Augmentation Architecture

V. The Road Ahead

Augmenting cognition using edge computing represents a perfect example of "transformative computing" in which existing technologies (e.g. lifelogging cameras, head-mounted displays and image processing algorithms) are leveraged to provide entirely new applications. However, it is clear that before augmented cognition is available at scale a number of challenges must be overcome.

A. Achieving Widespread Deployment of Edge Computing

Edge computing is key to augmenting cognition. It provides the low-latency processing, storage and sensing infrastructure that is essential for this demanding class of applications. Although actual deployments of edge computing are minimal today, there is intense industry interest and it is believed that we are on the cusp of major industry investments [9].

Cognitive augmentation applications have the potential to play the role of "killer apps" for edge computing. Even imperfect implementations of these applications can provide such high value to the end user, without facing any competing alternatives, that they have the potential to create demand for edge computing. These are examples of *Edge-native applications:* i.e., applications that simply cannot function satisfactorily without edge computing. This is in contrast to *Edge-accelerated Cloud-native applications,* where edge support is optional. The 20-year history of CDNs for web access is a good example of edge acceleration. Industry today is focused on identifying new edge-accelerated use cases rather than edge-native use cases, because they involve less investment risk. Edge-accelerated use cases involve much less software development, and their markets are much larger since they can function acceptably even in the absence of edge computing.

However, we believe that it is the creation of new edge native applications that will drive edge computing. The history of technology is replete with examples of rudimentary implementations of "killer apps" (e.g., automobiles, aircraft) driving the creation of the necessary ecosystem, and rapidly establishing a virtuous cycle that leads to continuous long-term improvements in both the core technology and the sustaining ecosystem. In computing, there is strong evidence that the development of the spreadsheet circa 1982-1983 was a major driver in the adoption of personal computing by small businesses. It was the low latency of human interaction (relative to timesharing) that made PCs indispensable for spreadsheets. The crucial role of low latency in cognitive augmentation applications suggests that they have the potential to play an analogous role for edge computing.

B. Unique Security and Privacy Challenges

Augmented memory and decision-making applications raise a number of significant security and privacy concerns that will need addressing prior to widespread adoption. For example:

Experience Provenance Traditional experience capture systems typically use wearable devices that are assumed to be trusted and the data produced is considered to accurately describe (within the constraints of the technology) the experience of the wearer. As edge computing is used to provide external data streams this represents an obvious point of attack. For example, if a microphone in a meeting room is used to capture audio associated with a meeting how do users know (without carrying out a manual review) that the audio captured is indeed an accurate reflection of what occurred in the meeting? [7]

Memory and Decision Manipulation Both contemporary psychology theories and recent experiments suggest that cued recall can be used to both re-enforce *and attenuate* human memories [2]—with immense security implications. A key challenge is how can users tell if their memories or decisions are being manipulated? In prior work [7] we have suggested that this will necessitate users monitoring cues that are delivered to them to look for unusual patterns of activity—akin to a virus checker for human memories or decision making processes. Such systems are likely to be both computational demanding and need access to contextual data as ground-truth, suggesting that edge support will be required. If a user's memory vault exists at the edge then care will need to be taken to ensure that different user's vaults are appropriately protected using, e.g. techniques borrowed from the field of application sandboxing.

Privacy Mediation The widespread use of augmented cognition applications that collect substantial data is likely to significantly impact the privacy of bystanders. One possible approach is to attempt to denature data streams before they are processed by augmented cognition applications. Such denaturing (e.g. face blurring or processing audio to hide speaker identities) is also likely to computationally demanding. Earlier work [8] has proposed that denaturing could be performed on cloudlets.

C. Resolving Ethical Dilemmas

Augmenting cognition also raises a number of non-technical challenges. While not necessarily the focus of scientific or engineering research these are important considerations for the community as they will significantly impact on the use of the technology.

Managing Shared Memories While we often think of memories as intensely personal, much of the data that underpins these memories often relates to other people. The challenges of designing appropriate security mechanisms increase significantly when sharing memories is considered. For example, in a meeting involving three people who owns the memory of the event? Is it necessary for each of the people to keep their own copy of the memory and then manage their own access controls or is it possible for a single copy to be maintained with appropriate shared ownership ? As the various participants chose to delete their copies of the memory what happens when the last interested party deletes the memory? Such challenges are compounded when we consider the case of managing memories after death or their use as part of the grieving process [11].

Avoiding a New Digital Divide Traditional technologies for human augmentation such as glasses and hearing aids generally aim to provide their users with abilities that approximate the norm. As a result they raise few ethical challenges. However, if cognitive assistance becomes widely available they raise important questions of fairness and equality when comparing augmented and non-augmented humans. Crucially, care will need to be taken to ensure that a new "digital divide" is not created between those that can afford to augment their capabilities and those that can not.

VI. Closing Thoughts

When our eyesight fails we are fitted for glasses. When our hearing fails we buy a hearing-aid. What do we do when our decision making or memory is no longer sufficient for the tasks in hand? The prospect of augmented cognition is truly tantalising, yet to achieve the vision will require leveraging a wide range of technologies to support these transformative applications. In this paper we have argued that edge computing will have a key role to play—enabling us to draw on a wide range of environmental sensing and processing resources while meeting the low-latency demands of cognitive augmentation.

Acknowledgments

The work of Nigel Davies was in part funded by the UK EPSRC under grant EP/N028228/1 (PACTMAN). Nigel would like to acknowledge the contributions of members of the RECALL project in shaping many of the ideas presented in this paper. The work of Mahadev Satyanarayanan was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001117C0051 and by the National Science Foundation (NSF) under grant number CNS-1518865. Additional support was provided by Intel, Vodafone, Deutsche Telekom, Verizon, Crown Castle, NTT, VMware, Seagate and the Conklin Kistler family fund. Satya would like to thank his students, staff and faculty colleagues in the Elijah and Gabriel projects for their many contributions to the work presented here. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view(s) of their employers or the funding sources mentioned above.

References

[1] T. Agus, C. Suied, S. Thorpe, and D. Pressnitzer. Characteristics of human voice processing. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, Paris, France, June 2010.

[2] M. C. Anderson, E. L. Bjork, and R. A. Bjork. Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, 7(3):522–530, 2000.

[3] Z. Chen. *An Application Platform for Wearable Cognitive Assistance*. PhD thesis, Computer Science Dept., Carnegie Mellon Univ., 2018.

[4] Z. Chen, W. Hu, J. Wang, S. Zhao, B. Amos, G. Wu, K. Ha, K. Elgazzar, P. Pillai, R. Klatzky, D. Siewiorek, and M. Satyanarayanan. An Empirical Study of Latency in an Emerging Class of Edge Computing Applications for Wearable Cognitive Assistance. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, Fremont, CA, October 2017.

[5] S. Clinch and C. Mascolo. Learning from our mistakes: Identifying opportunities for technology intervention against everyday cognitive failure. *IEEE Pervasive Computing*, 17(2):22–33, Apr 2018.

[6] S. Clinch, P. Metzger, and N. Davies. Lifelogging for 'observer' view memories: An infrastructure approach. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 1397–1404, United States, 2014. Association for Computing Machinery.

[7] N. Davies, A. Friday, S. Clinch, C. Sas, M. Langheinrich, G. Ward, and A. Schmidt. Security and privacy implications of pervasive memory augmentation. *IEEE Pervasive Computing*, 14(1):44–53, 1 2015.

[8] N. Davies, N. Taft, M. Satyanarayanan, S. Clinch, and B. Amos. Privacy mediators: Helping iot cross the chasm. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications*, HotMobile '16, pages 39–44, New York, NY, USA, 2016. ACM.

[9] Editorial. Take it to the edge. *Nature Electronics*, 2, January 2019. https://doi.org/10.1038/s41928-019-0203-8.

[10] S. R. Ellis, K. Mania, B. D. Adelstein, and M. I. Hill. Generalizeability of Latency Detection in a Variety of Virtual Environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 48, 2004.

[11] S. Ellis Gray and P. Coulton. Living with the dead: Emergent post-mortem digital curation and creation practices. In C. Maciel and V. Carvalho Pereira, editors, *Digital legacy and interaction*, Human-Computer Interaction Series, pages 31–47. Springer, 2013.

[12] J. Flinn. *Cyber Foraging: Bridging Mobile and Cloud Computing via Opportunistic Offload*. Morgan & Claypool Publishers, 2012.

[13] K. Ha, P. Pillai, G. Lewis, S. Simanta, S. Clinch, N. Davies, and M. Satyanarayanan. The Impact of Mobile Multimedia Applications on Data Center Consolidation. In *Proceedings of the IEEE International Conference on Cloud Engineering*, 2013.

[14] W. Hu, Y. Gao, K. Ha, J. Wang, B. Amos, Z. Chen, P. Pillai, and M. Satyanarayanan. Quantifying the Impact of Edge Computing on Mobile Applications. In *Proceedings of the 7th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys 2016)*, Hong Kong, China, 2016.

[15] A. Li, X. Yang, S. Kandula, and M. Zhang. CloudCmp: Comparing Public Cloud Providers. In *Proceedings of the 10th Annual Conference on Internet Measurement*, Melbourne, Australia, 2010.

[16] G. Nigro and U. Neisser. Point of view in personal memories. *Cognitive Psychology*, 15:467–482, 10 1983.

[17] M. Ramon, S. Caharel, and B. Rossion. The speed of recognition of personally familiar faces. *Perception*, 40(4), 2011.

[18] T. Ray. An Angel on Your Shoulder: Who Will Build A.I.? *Barron's*, February 2018.

[19] M. Satyanarayanan. Augmenting Cognition. *IEEE Pervasive Computing*, 3(2), April-June 2004.

[20] M. Satyanarayanan. The Emergence of Edge Computing. *IEEE Computer*, 50(1), 2017.

[21] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies. The Case for VM-Based Cloudlets in Mobile Computing. *IEEE Pervasive Computing*, 8(4), 2009.

[22] M. Satyanarayanan, W. Gao, and B. Lucia. The Computing Landscape of the 21st Century. In *The 20th International Workshop on Mobile Computing Systems and Applications (HotMobile'19)*, Santa Cruz, CA, February 2019.

[23] J. Wang, Z. Feng, Z. Chen, S. George, M. Bala, P. Pillai, S.-W. Yang, and M. Satyanarayanan. Bandwidth-efficient Live Video Analytics for Drones via Edge Computing. In *Proceedings of the Third IEEE/ACM Symposium on Edge Computing (SEC 2018)*, October 2018.

About the Authors

Mahadev Satyanarayanan is the Carnegie Group Professor of Computer Science at Carnegie Mellon University. He received the PhD in Computer Science from Carnegie Mellon, after Bachelor's and Master's degrees from the Indian Institute of Technology, Madras. He is a Fellow of the ACM and the IEEE. Contact him at satya@cs.cmu.edu.

Nigel Davies is a Distinguished Professor of Computer Science at Lancaster University, UK, and co-director of Lancaster's Data Science Institute. His research is characterized by an experimental approach involving large-scale deployments of novel systems with end-users. Contact him at n.a.davies@lancaster.ac.uk.