A Quintillion Live Pixels: The Challenge of Continuously Interpreting and Organizing the World's Visual Information

Kayvon Fatahalian
Carnegie Mellon University

I estimate that by 2030, cameras across the world will have an aggregate sensing capacity exceeding one quintillion (10^{18}) pixels. These cameras, which will be embedded in vehicles, worn on our bodies, and positioned throughout public and private everyday environments, will generate a worldwide visual data stream that is over eight orders of magnitude greater than the rate of video ingest by Youtube today. *A vast majority of these images will never be observed by a human eye*—doing so would require every human on the planet to spend their life watching the equivalent of ten high-definition video feeds! Instead, future computer systems will be tasked to automatically observe, understand, and extract value from this dense sampling of life's events. Some applications of this emerging capability trigger clear privacy and oversight concerns, and will rightfully undergo great public debate. However, many others clearly have the potential to have critical impact on central human challenges of the coming decades. Sophisticated image analysis, deployed at scale, will play a role in realizing efficient autonomous transportation, optimizing the daily operation of future megacities, enabling fine-scale environmental monitoring, and advancing how humans access information and interact with information technology. Our ability to develop new image understanding techniques (the topic of Kristen Grauman's talk in the same Frontiers of Engineering session), architect large-scale systems to efficiently execute these computations (the subject of my own research), and deploy these systems transparently and responsibly to improve worldwide quality-of-life is a key engineering challenge of the coming decade.

To understand the potential impact of these quintillion pixels, let’s examine the role of image understanding in three contexts: cameras in vehicles, on the human body, and in urban environments:

Continuous capture on vehicles. It is estimated that there will be over two billion cars in the world by 2030 [1]. Regardless of the extent to which autonomous capability is present in these vehicles, a vast majority of them will feature high-resolution image sensing. (High-resolution cameras, augmented with high-performance image processing systems, will be a low-cost and higher information-content alternative to more expensive active sensing technologies such as Lidar.) Image analysis systems are critical for localizing vehicles within their expected surroundings and comprehending dynamic environments to predict and detect obstacles as they arise. In short, they are critical to the development of vehicles that drive more safely and use roads more efficiently than human drivers. Today, researchers in academia and industry are racing to develop efficient image processing systems that can execute image understanding tasks simultaneously on multiple high-resolution video feeds and with low latency. Hundreds of tera-ops of processing capability (present only in top supercomputers only a decade ago) will soon be commonplace in vehicles, and computer vision algorithms are being re-thought to meet the needs of these systems. High-performance analysis of vehicular video feeds will be a key technology that enables significant advances in transportation efficiency.

Continuous capture on humans. Although on-body cameras, such as Google Glass, have thus far failed to realize widespread social acceptance, there are compelling reasons for cameras to capture the world from the perspective of a human (“egocentric” video). For example, enabling mobile augmented reality (AR) requires systems to precisely know where a person is and what a headset wearer is looking at. (Microsoft’s Hololens headset is one example of promising recent advances in practical AR technology. [2]) Achieving the goal of commodity, pervasive AR demands continuous, low-energy egocentric video capture and analysis.

More ambitiously, for computers to take on a more expansive role of augmenting human capabilities (e.g., the ever-present life assistant), they must understand much more about us than our present location, the contents of our inbox, and our daily calendar. Computers will be tasked to observe and interpret human social interactions to know what advice to give, and when and how to interject information. For example, in a recent trip to Korea, I found myself wishing to experience a meal at a local night market. However, my inability to speak Korean as well as my unfamiliarity with the market's social customs made for a challenging experience in the bustling atmosphere. Imagine the utility of a system that, given a similar view of the world as I, could not only identify the foods in front of me, but suggest how to assimilate into the crowd in front of a vendor (be assertive, attempt to form a line?), instruct me if it was acceptable to sit in a rare open seat near a family occupying half a table (yes, it would be okay to join them), and detect and inform me of socially awkward actions I might be taking as a visitor (you are annoying the local patrons because you are violating this social norm!). All of these tasks require mobile computing devices to constantly observe and interpret complex environments and complex human social interactions. Cameras located on the body, seeing the world continuously as we do, are an attractive sensing modality for these tasks.

Continuous capture of urban environments. Last, it is now clear that cameras will continue to be increasingly pervasive in urban environments. (It is estimated that about 280M security cameras exist in the world today, with cities such as London, Beijing, and Chicago featuring thousands of cameras in public spaces [3].) While today's deployments are largely motivated by security concerns, the ability to sense and understand the flow of urban life both in public and private spaces provides unique opportunities to better manage modern urban challenges such optimizing urban energy consumption, monitoring infrastructure and environmental health, and informing urban planning.

Putting it all together: one quintillion pixels. In 2030 there will be 8.5 billion people in the world [4], two billion cars, and (extrapolating recent trends [3]) at least 1.1 billion security/web cameras. Conservatively assigning one camera to each human and eight views of the road to each car, and assuming 8K stereo video stream per source (2 x 33 megapixels), there will be nearly one quintillion pixels across the world continuously sensing visual information. The engineering challenge of ingesting and interpreting this information stream is immense. For example, using today's state-of-the-art machine learning methods to detect objects in this worldwide video stream would consume nearly 10^{13} Watts of computing power [5], even if executed on today's most efficient parallel processors. This is approximately the same amount of power used by humans across the world today [6]. Clearly, advances both in image analysis algorithms and the design of energy-efficient visual data processing platforms are needed to realize ubiquitous visual sensing. Addressing this challenge will be a major focus of research spanning multiple areas of engineering and computer science in the coming years (machine vision, machine learning, AI, compilation techniques, and computer architecture). Success developing these fundamental computing technologies will provide us new, valuable technology tools to tackle some of the world's most important future challenges.

[1] *Two Billion Cars: Driving Toward Sustainability*, D. Sperling and D. Gordon, Oxford Academic Press 2010.

[2] Microsoft Hololens Website. https://www.microsoft.com/microsoft-hololens/

[3] *Video Surveillance Camera Installed Base Report*, HIS Technology, 2015

[4] United Nations Estimate, 2015.

[5] *GPU-Based Deep Learning Inference: A Performance and Power Analysis.* NVIDIA Corporation, 2015

[6] *2015 Key World Energy Statistics*. International Energy Agency, 2015