Vision: Cloud-Powered Sight for All

Showing the Cloud What You See

Paramvir Bahl Matthai Philipose Lin Zhong

Microsoft Research
Redmond, WA 98052

{bahl,matthaip,t-lizhon}@microsoft.com

ABSTRACT

We argue that for computers to do more for us, we need to show the cloud what we see and embrace cloud-powered sight for mobile users. We present sample applications that will be empowered by this vision, discuss why the timing is right to tackle it, and offer our initial thoughts on some of the important research challenges.

Categories and Subject Descriptors

A.1 [**General Literature**]: introductory and survey

General Terms

Algorithms, Design, Human Factors, Languages, Performance, Security

Keywords

Camera, cloud, computer vision, mobile computing, wearable computing

1. INTRODUCTION

What computing can do for us is fundamentally limited by what data we give to our computer. Today we give *instructions* in the form of touch, speech, gestures, key press, and button click. While we accomplish a lot with these interactions, the rate and quantity of information that we provide to the computer limits the useful things it can do for us. Think about how many words we can speak per minute and how many words we can type per minute and compare this to how fast your computer can process data.

In addition to explicit command-and-compute types of interactions, we also provide data to our computer implicitly, for example contacts, credit card transactions, web browsing digital photos, and videos. Using the tremendous power of the cloud, this data not only enables computing to be highly personalized but also empowers it to exploit human users as sensors, e.g., [AWB+11, ZZW+11]. Again, the rate of information transfer and the quantity of data provided to the computer limits the possibilities.

Researchers have been working on addressing these limits and we have seen examples where mobile computers directly acquire information from the physical world, including the human user, and share it with the cloud. Inspiring applications and services have been demonstrated that analyze continuously collected accelerometer and microphone data, e.g., [LPL+09], and occasionally data from phone camera, e.g., [ACR09]. The rate of information transfer is greater than the traditional forms but still limited.

In this paper, we assert that today's cloud powered computers should be configured to analyze a lot more data, which is provided to them at a much higher rate. If this is done, we will be able to unleash the creativity of developers who will write powerful applications that make use of this data and make us even more efficient. Today, we let the computer hear what we hear, and know where and how we move, but we have to do more. *We have to let our computer see what we see.*

Our rationale is simple: visual information is the richest form of human sensory input. More often than not, it also requires the highest information transfer rate. The human nervous system is centered on the visual experience. We learn, reason, and act based on what we see. Our interactions and collaborations with fellow humans are often triggered by our visual senses and many of our social experiences are influenced by it. By showing our computer what we see, we can not only create the next generation of personalized service but also overcome some the fundamental limitations of being human.

In the rest of the paper, we motivate our *show the cloud what you see* research agenda and describe some of the challenges. Specifically in Section 2, we offer application scenarios of cloud-powered sight. In Section 3 we discuss why this is a great time to work on this vision. And in Section 4 we present some initial thoughts on some of the important research challenges. We conclude by discussing related work and visions in Section 5.

2. MOTIVATIONAL APPLICATIONS

Predicting compelling applications is hard. However, some dimensions of the design space of mobile personal vision based applications seem clear. Figure 1 illustrates four dimensions along which the user experience of such applications may vary.

First, the degree to which raw pixels are interpreted before presentation to the user may vary from none at all (e.g., remotely viewed video) to some interpretation (e.g., reconstructed 3-D views) to extensive interpretation as symbolic data (e.g., via object or activity recognition). Second, the latency between sensing and presentation to the user may vary from faster than frame rate (for predictive applications), to frame rate (for interactive applications such as gaming) to multi-second response times (e.g., for reminders) to hour- or day-long delays in retrospective applications (e.g., vacation footage reconstruction). Third, the amount of user attention required to act based on the visual footage may range from none

MCS'12, June 25, 2012, Low Wood Bay, Lake District, UK.

at all if the system is set up to monitor continuously and act proactively, to point-triggers from the user when hits of information are required (e.g., "identify that person in front of me") to immersive experiences that demand full, extended attention (e.g. in augmented- or reconstructed-reality settings). Finally, user may expect action based on footage from individuals (often, their own), federated across groups affiliated with them (e.g., in social settings), or across people from a wide area whom they may not even know personally (e.g., to obtain city- or even world-wide perspectives).

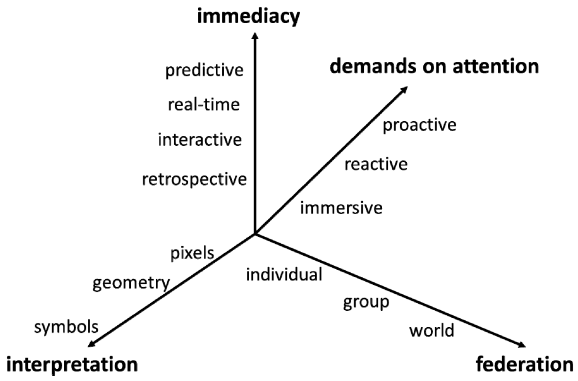

Figure 1: Design space for applications of cloud-powered sight

Many interesting applications fall within this space:

Augmented perception: Although the visual experience is rich, we are limited in what we can pay attention to at a time. This is highlighted by the invisible gorilla video [INV]. While this limitation has helped *Homo sapiens* survive in focusing on important signals and suppressing noise, it can fail us when multiple important signals are present simultaneously. With the cloud watching out for each individual human user, the chance that one misses a social clue or gets run over by a car will be significantly reduced.

Augmented cognition: Even if we perceive an object, we are limited in our cognitive power in analyzing the object and relating it to what we already know. That is, we have *limited rationality* as characterized by Herbert Simon. Occasionally we would have difficulty recalling the names of people we already know when we run into them on the street. Recognizing a person of interest from a crowd would be a challenge task even for an experienced detective, who can definitely use automatic face recognition as exemplified by the recent movie *Mission Impossible 4*.

Crowd-sourced reality: Today, when a driver uses Google Map's Navigate feature, his/her location information is collected and used along with that from many others to estimate the traffic and the estimation benefits the driver. Users get a better sense of the physical world by sharing data with the cloud. In the future, when users share what they see with the cloud, a spectator in the stadium of Super Bowl will be able to see the game from any angle and replay the most thrilling touch from the best perspective.

Crowd-sourced lifelog: People who show what they see to the cloud naturally get their lifelog when the cloud keeps a copy of what they see. If a sufficient number of people around a person show what they see to the cloud, the cloud can extract pieces that have the person captured and aggregate them to produce a comprehensive, though incomplete, lifelog for that person. This application is unique in that it brings benefit to people who do not show what they see to the cloud, which is important for the societal adoption of our vision.

3. WHY THE TIMING IS RIGHT

We are motivated by both evolutionary and revolutionary breakthroughs in multiple research areas. There are several interrelated reasons why we do not yet show the cloud what we see. (*i*) High cost of acquisition, analysis and storage of visual information is high; (*ii*) Short battery lifetime of the acquisition system; (*iii*) Low usability of the acquisition system, which is supposed to be wearable and portable; and (*iv*) Lack of applications and services built on top of the acquired visual information.

3.1 A Buck for Billion

Moore's Law is expected to survive for at least another decade. Looking at our history, it is very likely that the cost of a billion transistors will drop by many folds, getting closer to a US dollar, or "A Buck for Billion" as coined by Gene Frantz [Fra09]. Both the mobile and cloud will become more powerful by orders of magnitude and the same integrated circuit will become cheaper by orders of magnitude.

With this in mind, we expect to see mobile devices by the end of the next decade to be significantly more powerful, in-fact as powerful as high-end desktop computers today. Specifically, we expect to see a mobile systems-on-chip (SoC) with many billions of transistors (for example, Tegra 3 already has about one billion transistors today). Although not all these transistors will be active most of the time due to peak power and thermal constraints, with careful power management, such mobile SoC will deliver performance that is equivalent to today's state-of-the-art desktop microprocessor.

Similarly we expect to see microprocessors with many tens of billions of transistors, and today's microprocessors cost a tiny fraction of their current price. This trend has two implications: (*i*) the cloud will be more powerful by orders of magnitudes; and (*ii*) the cloud will be any place with sufficient energy supply, not necessarily sitting in the datacenters. It is possible a car and a residential house will have built-in the computational power of today's datacenter.

3.2 Battery and Thermal Management

While the research community has been pessimistic about battery and thermal management technologies for mobile systems, we still expect significant improvement in them in the next decade.

History shows that battery technology improves about 5-10% annually in terms of energy and power density [Pow95]. Although it is much slower than Moore's Law, 10% annual improvement will lead to 2.6x improvement in a decade. Furthermore, recent development in applying silicon nanowires and carbon nanotubes to energy storage is promising bigger breakthroughs, e.g., [CPL+08, RSG+09]...

The same integrated circuit will consume less energy as it is implemented using a newer process technology. Before 90nm technology, the new implementation consumes only half of the energy of the previous one [Fra08]. Although the increasing ratio of leakage power consumption has reduced the efficiency benefit of using a newer process technology, the general trend holds; and within a decade, the same computing, implemented using the newest

semiconductor process technology, will likely consume 10 times less energy than what it does today.

Finally, there has is an emerging trend of using highly specialized integrated circuits, instead of general-purpose processor, in common computation. This specialization will allow the same computation to be accomplished with lower energy consumption by many folds [HQW+10]. For example, face detection is realized with a dedicated integrated circuit on OMAP4 SoC from Texas instruments. Another example is that of the modern browser rendering engines, which employ the graphics accelerator inside a mobile SoC through OpenGL.

The implication is that we will be able to do greater than 30 times more work with a battery of the same size and weight by the end of the coming decade, and this brings a tremendous opportunity for mobile devices to do the heavy-lifting when analyzing visual scenes. While the cloud will still be used for delivering even more formidable computation, we expect to see it to be used for relieving the mobile from thermal management problem and to be used as a trusted agent to aggregate data from many mobile users.

3.3 Break-through in Sensors

We are motivated by the breakthroughs in building inexpensive and miniature sensors, driven by MEMS and nanomaterial research. From motion to sound to trace chemicals, an integrated circuit will be able to not only measure almost all physical properties we care but also extract meanings out of the measurement with low energy consumption. With these sensors, the cyber world can be tightly coupled with the physical world in an energy-efficient and cost-effective way.

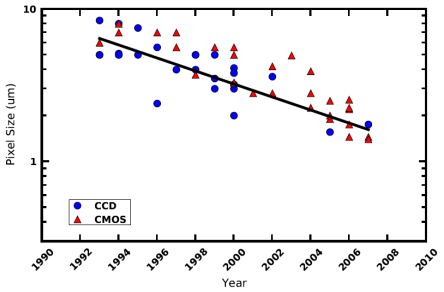

Figure 2: Pixel size halves every five years, not only allowing more pixels but also processing within a pixel. (Source: Keith G. Fife [Fif09]; included with the author's permission)

We are particularly excited by the development in camera technologies enabled by Moore's Law. (*i*) First, the diminishing size and cost of transistors allow not only smaller pixels, as shown in Figure 2 [Fif08], but also more intelligent pixels, i.e., pixels with processing built-in. For example, the digital pixel sensor technology from Pixim adds an analog-to-digital converter (ADC) alongside each pixel to extend the dynamic range of the pixel reading [Pix08]. Research prototypes have been reported that allow digital processing inside the image sensor, e.g., NAG09]], and that employ new architectures such as frameless address-event representation (AER) image sensors, e.g., [LSL11]. (*ii*) Second, the increasingly available computational power discussed in Section 2.1 also helps overcome the limitation in the image sensor and reduce the cost of visual information acquisition systems. In particular, compressed sensing, a compute-intensive method, has been exploited to simplify the hardware complexity and cost of acquisition systems [WMN11], e.g., the single-pixel camera [DDT+08]. (*iii*) Furthermore, refined semiconductor process technology allows novel optical structures for image sensors. For example, layers of gratings can be added to make pixels sensitive to the angle of light [WM12]. Array of small lens can be added to make multi-aperture image sensors for 3D imaging [FEW08]. (*iv*) Finally, combining the innovation in novel image sensor designs and the computational power allows novel ways of capturing visual experience. Digital light field photography [Ng09] leverages both multi-aperture image sensors and compute-intensive post-processing to eliminate the "focusing" step of photo capturing. Lytro, the first commercial light field camera based on the technology, allows a user to capture the image without focusing and to focus computationally afterwards. This technology apparently reduces the need of human engagement in capturing quality photos.

3.4 Computer Vision

Finally we believe computer vision research has reached a point that the coming of age for its applications is just around the corner as attested by numerous deployed applications for smartphones, e.g., Google Goggle and SnapTell, automobiles, e.g., automatic parking and driving, and face recognition in photo management. Our belief is based on three related developments in the recent five years, in addition to the increasing computational power.

First, novel algorithms and methodologies have been developed that exploit advancement in statistical machine learning, e.g., boosting and bagging, and exploit the increasingly available image datasets and computational power. Recent development along this line not only enable features of various abstract levels to be extracted, objects to be recognized, but also allows the same object captured from various intrinsic, e.g., camera, and external, e.g., perspective and illumination, conditions to be matched and aligned. Ongoing research from the computer vision community extensive investigates the extraction and use of high-level meanings of images, e.g., scene understanding, category recognition, and even context analysis. For an excellent comprehensive treatment, see [Sze10].

Second, large-scale image datasets are increasingly available so that sophisticated machine learning techniques are now applicable to computer vision problems. These datasets become possible because (*i*) digital cameras and camera phones have made photo capturing easy, leading to a massive number of digital photos being captured around the globe every day; (*ii*) High-speed Internet and the popularity of Web 2.0 services such as Facebook and Flickr have made a significant portion of digital photos captured public; and (*iii*) the computer vision community has made a conscious endeavor to create image databases that are not only massive in scale but also relational, rich in semantic and ontological information, e.g., ImageNet [DDS+09].

Finally, new imaging hardware allows previously impossible information to be captured. Many of these emerging camera technologies have already been discussed in Section 3.3. For example, depth cameras provide depth information of a scene and empower much more accurate object recognition and tracking, as exemplified by Microsoft Kinect; the digital pixel sensor technology from Pixim enables a much wider dynamic range in captured images; and light-field cameras like Lytro capture the entire light field of the scene and therefore more information by many folds than conventional cameras. Finally, sensors of other modalities are increasingly available on digital cameras and camera phones. They allow visual experience to be captured with a rich set of context information such as location and social settings. These

new capabilities will not only significantly improve the performance of existing computer vision solutions but also enable the exploration of novel ones such as context-aware object recognition.

4. RESEARCH CHALLENGES

We are excited by the vision of showing the cloud what we see because if realized it will fundamentally transform the entire ecosystem of computing, communication, and human-machine interaction.

4.1 Always-on Wearable Camera

The very first challenge is to build a visual experience capturing system that meets the following requirements: It must

- Capture application-specific information
- Last for a day without recharging;
- Have a wearable form factor;
- Be able to ship the captured information to the cloud in real-time, with delay under several seconds.

The real challenge is battery lifetime as many wearable camera prototypes have been built and commercially sold that meet the other three requirements. They are either fitted into a pair of glasses, attached to a helmet, mounted on top of the ear, or neckworn like a pendulum. Although none of them are able to directly ship the captured information to the cloud in real-time yet, it is not fundamentally difficult since smartphones are already able to stream video through the cloud in real-time with applications like Apple's FaceTime.

In Section 3.2, we discussed the issue of battery lifetime at length and claimed that over the next decade or so great strides will be made towards solving this problem. Yet we believe that this problem can be potentially addressed even today with the following related principles:

First, the system should capture application-specific information only, instead of compressed video. Useless data should be discarded as early as possible. To accomplish this, we should consider emerging image sensors that allow in-sensor processing while simultaneously working on designing novel image sensors that are geared toward *information capturing*, instead of *video capturing* – this is the design goal of most today's CMOS image sensor arrays used in wearable cameras.

Second, the system should be heterogeneous. Today's cameras including the wearable ones are largely homogenous. That is, a single CMOS image sensor array is employed to capture the video, which is then compressed with dedicated hardware. They operate either as on or off, which is usually dictated by the user through a button or remote control. Instead, we believe the system should leverage a collection of sensors of various modalities and a collection of image sensors with various specialties to capture the required information with low average power consumption. Recent research prototypes of wearable cameras, e.g., SenseCam [HWB+06], have already employ sensors of various modalities. In particular, sensors of lower power consumption, such as accelerometer and microphone are used to trigger the camera operation. As specialization has become a major approach for improving the energy efficiency of computing, we argue it should be applied to visual information capturing too: image sensors specialized for capturing different aspects of visual information should be combined to achieve high operational efficiency.

Finally, the system should leverage other personal computing resources, including computing, storage, energy, and connectivity. Today such resources are available in the smartphones we carry. In a decade, we will very much likely still carry a personal computing device in our pocket, though it may be called a different name. With a smaller form factor, this device will most likely have greater energy and more sensory and computational power than today's wearable systems by orders of magnitude.

4.2 Low-Latency, High Throughput Network

The first big challenge in *showing the cloud what you see* is to increase the network capacity. The applications we want to enable inherently create a tremendous burden on wireless networks. Even with the best video codecs, a single compressed video stream requires bandwidth of about a mega-bit or two per second and the overall demand on the network goes up linearly with the number of people being served. Today's mobile operators (MOs) are already scrambling to meet the exponential growth of data brought about by the popularity of smartphones and tablets. One of the strategies they are employing to curb this demand is tiered pricing. Still, even when users opt for a plan with several GB of data usage, the authors of [II03] calculated that a 24-hour 160 by 120 video stream at 10 frames per second, encoded in MPEG4 consumes more than 1.5 GB of bandwidth. Recently Cisco claimed that mobile video will grow at a CAGR of 90 percent between 2011 and 2016. Of the 10.8 Exabyte per month crossing the mobile network by 2016, 7.6 Exabyte will be due to video [Cisc11]. Note that this report is based on current end-user behavior but not on the vision we are describing in this paper.

In addition to needing greater capacity, many applications which take advantage of *showing the cloud what you see* require low latency and low jitter in uploading visual information, e.g., crowdsourced reality as outlined in Section 2. Measurement studies have shown that existing and emerging cellular networks have high latency and delay, incurred by the combination of last-hop wireless and sub-optimal Internet connectivity to the cloud. [HXT+10, HQG+12].

Solutions to these problems can be resolved by taking a multi-prong approach that may include: (*i*) lobbying the regulators and policy makers around the world to open up unused spectrum for both licensed and unlicensed use, (*ii*) encouraging MOs to increase spatial reuse by embracing smaller cell size networks e.g. via femtocells, (*iii*) helping Wi-Fi operators and MOs embrace traffic offloading between licensed and unlicensed networks, e.g. between 4G and Super Wi-Fi networks, and (*iv*) continuing to improve spectrum efficiency by at least an order of magnitude to meet the growing capacity demand.

The research and legal communities are working together like never before to find solutions to the looming spectrum crisis. Whitespace networks and more importantly database driven opportunistic networks are a recent example of forward thinking technologist and regulators teaming up to open additional spectrum for consumers who are demanding greater connectivity and bandwidth [BCM+09, MCT+10]. We must design and develop opportunistic networks that use both distributed spectrum databases and cognitive radios with distributed sensing capabilities. Opportunistic networks allow primary and secondary users to coexist, and therefore spectrum use is optimized.

To reduce the network latency while bringing the power of the cloud to the user, we advocate the use of *Cloudlets*, which we define as a resource rich infrastructure computing devices that

have high-speed Internet connectivity to the cloud. Mobile devices such as smartphones and tablets can use Cloudlets to augment their capabilities and enable applications that were previously not possible [SBC+09].

Access to cloudlets may be via Wi-Fi, Super Wi-Fi, or any of the emerging cellular technologies. But there is another possibility, the untapped 7 GHz of unlicensed spectrum available in the 60 GHz range. Recent advances in CMOS technology have reduced the cost of 60 GHz devices significantly; however the nature of 60 GHz radio waves leads to significant challenges for operating high rate links. All other factors being equal, a 60 GHz link is roughly 55 dB (a factor of 300,000× worse) than a 2.4 GHz link in terms of the signal-to-noise ratio (SNR) that determines packet delivery. This is due to three factors: first, the free-space path loss is higher due to the extremely small 5mm wavelength; second, the channels are 100 times wider and thus 20 dB noisier, and third most commercial equipment uses only 10mW transmit power (compared to 802.11's typical 50mW) in order to meet regulations and energy budgets. In addition to large bandwidth, directionality represents another novel aspect of 60 GHz technology and a good network design that connects the user's wearable device to the cloudlet can take advantage of this. Directional antenna effectiveness is inversely proportional to the square of the radio wavelength so the short wavelength of 60 GHz leads itself to compact antennas. With directional antennas, 60 GHz links can support multi-Gbps rates over distances of several meters, thus providing convenient access to cloudlets.

Finally, to improve spectrum efficiency, we advocate at least three hardware-driven technologies: First, as suggested in [YZS+11], mobile clients should have multiple antennas that exploit various forms of beamforming for the uplink; Second, base stations should employ a large number of antennas to communicate with many clients simultaneously through multi-user beamforming, and third, instead of competing for spectrum resources, mobile clients help each other through cooperative communication schemes.

There is a lot of research work we still have to do to enable high-capacity, low latency, low jitter networks, and as outlined above the solutions will come in many different forms.

4.3 Capturing and Representing Visual Information

There are three interesting opportunities to reduce the aggregated network capacity requirement from showing the cloud what users see. First, as discussed above, if only application-specific information, instead of compressed video, is shipped to the cloud, the capacity requirement will be significantly reduced. This opportunity will be naturally realized if the information capturing system discussed in Section 3.1 becomes information capturing, instead of video capturing.

Second, there is a lot of redundancy in a user's daily visual experience. For example, a user sees her car many times a day, but the car does not change very much in its appearance so often. Existing and emerging video encoding standards, e.g., H.264/MPEG AVC and HEVC, unfortunately, do not effectively exploit such redundancy. They do employ motion estimation to identify redundancy, e.g., the same car, in consecutive frames, but they do not do it for video captured at different times.

Similarly, there is a lot of redundancy in the visual experiences of multiple users. For example, users walking down the same streets will share much of their visual experience, without various temporal, angular shifts. What they see can be jointly encoded to achieve much higher compression rate and significantly reduce the aggregated capacity requirement.

The last two opportunities require new research into video analysis, representation, and compression. Some primitive steps have already been taken though for different goals. For example, there is significant work about video summarization to identify the most informative frames of the video sequences for both compression and retrieval. Similar methods have been applied to compress real-time video streaming for face-to-face videoconference [WC05]. We believe the following map-plus-dictionary approach holds a great promise.

Much of our visual experience is about things that do not move or change in a short time, e.g., buildings, hallways, and major pieces of home furniture and appliance. Given our location and the direction of our view, much of the visual experience can be derived from a map built beforehand. The visual information capturing system only needs to capture and share the difference between the real view and the one derived from the map. The direction of view can be efficiently estimated with kinetic sensors and the location of a user is increasingly available from a combination of various technologies. Interestingly, the visual information capturing and location estimation can be performed in an iterative manner, given the map. The map, on the other hand, can be constructed and maintained through a combination of war-driving, similar to how Google StreetView is created today, and crowd-sourced data collection, similar to how traffic map is generated today.

In addition to maps of relatively stable visual objects, a user will benefit from a personal dictionary of visual objects that she usually sees. For example, if the user owns a particular smartphone, it will appear in her visual experience very often and can be included in her dictionary for compressing her visual experience. We note that dictionary-based methods have been widely studied in the data compression community [LM99]. However, there is a key difference between our proposal and existing solutions. Existing dictionary-based methods all employ very low-level dictionary entries, usually computationally derived without understanding the semantics of the visual experience, e.g., spectrum and DCT components. In contrast, the dictionary in our envisioned approach consists of entries that are of much higher level of abstractions and are semantically meaningful, e.g., smartphone, dog, and car. The increase in abstraction of dictionary entries may require fundamentally different methods of deriving dictionary entries and decomposing visual experience into dictionary entries.

Finally, the map-plus-dictionary approach will need a rethinking of the representation of visual experience. Currently visual experience is captured as a sequence of frames, each of which is an array of pixels. However, the map-plus-dictionary approach will be most effective if the visual experience can be captured semantically or it should be represented based on how a human user would interpret it if given enough time. For example, when a user sees her house at different time of the day and from different perspectives, instead of recording the visual experience by pixels, the system can potentially describe and represent the visual experience with a dictionary entry of the house, the viewing angle, and the illumination setting. We expect recent development about feature-based alignment and deriving structure from motion in computer vision to provide useful tools for this purpose. For example, computer vision researchers are able to reconstruct multiple views of tourist sites using photos collected from keyword-based image search queries [SSS08]. And they are currently aiming at generating 3D models of these sites in this way. This new,

semantic representation captures a visual experience by analyzing it and apparently requires significantly more computation at the time of capturing, which bodes well with the increasing availability of computational power at a mobile or wearable device as analyzed in Section 2.

4.4 From Datacenters to Ubiquitous Cloud

As the cost of computing video decreases, we envision the current cloud model with highly centralized mega datacenters will be enhanced with smaller datacenters distributed in residences, communities, and businesses, thus making the cloud truly ubiquitous. Beyond the cloudlet vision outlined in [SBC+09], this distributed cloud model invites researchers to think about a multi-tier cloud computing paradigm, its management, and optimized use. There are multiple rationales behind this speculation. First, the computing power of a residential house will be comparable to that of today's small to mid-tier datacenters. Consequently, much of the processing will take place as close as possible to where the data is captured therefore significantly reducing the stress on the network and outside infrastructure.

Second, as the latency of the last-hop wireless link decreases significantly the latency due to the rest of the network infrastructure will become the bottleneck for instant network transactions, e.g., fast web loading and crowd-sourced reality. When this happens, we expect the base stations will evolve into mini datacenters serving its mobile users with low latency. The business implication of this is apparent.

Third, because visual experience is highly location-dependent and there is a huge redundancy in visual experience captured by co-located users, the distributed cloud model will be best positioned to leverage this redundancy.

Moreover a locally hosted, owned datacenter may provide the real-time power to co-located users while ensuring what you see does not go outside that property. This is highly desirable for corporations and private homes, which are particularly sensitive about sending any data, visual or otherwise to an outside cloud.

Finally, showing the cloud what we see will pose a grand challenge to the design of databases as user-generated data increases by many orders of magnitude. Moreover, the data from visual experience will be highly structured and have extremely rich relationships among entries. That is, the visual experience from a user may compose of a number of objects, which may in turn have subcomponents; at the same time, the visual experience from the same user but from different time may be highly similar or related; those of multiple co-located users may also have a lot of commonality and therefore related. The complicated structure and relationships will probably require a rethinking of the database design.

4.5 Programming Support

The key to the vision of showing the cloud what we see is the applications that can be easily built on top of the visual experience captured. One may argue that the availability of development kits for computer visions such as OpenCV have already achieved this goal. Yet we believe the abstraction level provided by computer vision development kits today is still too low. That is, developers still operate on pixel arrays; the development kits only provide APIs for developers to extract visual information, e.g., detecting and recognizing a face. In contrast, in the envisioned the system, the step of abstraction should have already been extracted during the capturing of the visual experience, as discussed in Section 4.3. The developers will mostly have a semantic representation of the visual experience available. We need programming support that enables them to easily work with this high abstraction level, instead of raw pixel arrays. Such abstractions include the properties of visual objects including semantic, physical and temporal properties, relationships between visual objects, and operations that can be applied to them. For example, we expect programmers to write code similar to the following

if (**any** car **is** approach**ing**) alert();

4.6 Legal and Privacy Challenges

Intille and Intille provide an excellent review of the legal issues related to wearable cameras in [II03]. Much of their speculations are applicable to our vision of *showing the cloud what you see*. In addition, we foresee several new privacy issues. First, there are three levels of privacy related to capturing our visual experience: analyzing the captured video data in real-time, storing it, and sharing / disclosing it, each with increasing legal liability.

Second, the legal liability depends on how the visual experience is represented. We note that some form of abstraction might reduce the privacy concern and legal liability. For example, while a friend we speak with may object to being videotaped during a conversation, he or she may be accommodating to being recorded as the person we spoke with at a particular time and location.

Third, the core problem in capturing other people's activities is that the person being captured has no control on how the video data is analyzed in real-time or years later or is shared with others. This is an important challenge to many social protocols we are used to today. We believe it is similar to the copyright management problem in the music industry. Without copyright management, when a song is digitalized and made available to a user, the creator loses all her control in how the song will be played, edited, and shared. Since this problem is considered reasonably solved today, we speculate that the control problem in capturing the visual experience will also be solved using an approach that involves robust human identification and a trusted information capturing system [LSW+12].

The adoption of our vision by the society will likely depend on how people who do not show what they see to the cloud react to being visually captured by early adopters. It is important that we think about what tangible and immediate benefits these people will receive. When the benefits overweigh the privacy concerns, the social barrier to adopting these technologies will be lowered.

5. PRIOR ART AND RELATED VISIONS

Our vision builds on multiple existing research visions.

MyLifeBits from Microsoft employs a wearable camera, SenseCam, to capture photos of meaningful moments of a user's life and envisions a lifelong database of captured photos that can be easily searched. The vision of MyLifeBits is to enhance users' long-term memory, which is known to have long latencies for writes, i.e., memorizing, and reads, i.e., retrieval. Our vision goes beyond just enhancing the long-term memory to include perception and cognition. Furthermore, our vision also brings the possibility of sharing with a much tighter coupling between the cloud and the mobile.

Realizing "the user attention is the most precious computing resource," Project Aura from CMU aims at reducing interrupts,

automating tedious work so that users can focus on more important ones. Driven by the same human limitation, our vision is, however, to extend the human capability, instead of maximizing its use. Envisioned about a decade before our vision, Project Aura emphasizes the use of audio input from the user and environment. In contrast, our vision sets the visual experience at the center stage.

Our vision also encompasses a closely related vision: augmented reality. While augmented reality can have a broad definition, e.g., including virtual markers on ESPN sports games, what is related to our vision is what captured by the Terminator Vision from *the Terminator* (1984): enhancing what we see on the go by overlaying digital information on top of a user's visual experience. It has three parts: (*i*) acquire information about what a user may see; (*ii*) figure out what information to add; and (*iii*) output that information to the user. Part (*i*) may employ a camera, e.g., Google Goggle, or may not, e.g., Google's Project Glass [Goo12], which relies on location information to infer about what a user may see most of the time. Part (*ii*) may or may not involve the cloud. And Part (*iii*) usually employs a head-mounted display (HMD). The video see-through type of HMDs actually employs a camera to capture the user's visual field; the captured visual field is then displayed along with augmented information. Augmented reality is related to our vision in two ways. First, we consider augmented reality as an application that can be built on top of showing the cloud what you see and, more importantly, it will be significantly better that way by closely coupling the cloud and leveraging sharing. Moreover, we consider technologies from augmented reality instrumental for addressing some of the challenges outlined above, in particular the user interface challenges.

The sensor network community has investigated wireless visual sensor networks of nodes with image sensors [SH09]. The used image sensors include both commercial-off-the-shelf ones, e.g., [KGS+05, RBI+05], and research prototypes, e.g., [GMG+11]. However, a focus on long-term operation leads to extremely tight power constraint and therefore much reduced capabilities in capturing and processing the visual experience; and such visual sensor networks are limited to applications that are usually concerned with surveillance and monitoring.

ACKNOWLEDGEMENTS

The authors are grateful for the useful input from David Chu, Stefan Saroiu, and Cha Zhang from Microsoft Research.

REFERENCES

[ACR09] Martin Azizyan, Ionut Constandache, and Romit Roy Choudhury. 2009. SurroundSense: mobile phone localization via ambience fingerprinting. In *Proc. ACM Int. Conf. Mobile Computing and Networking* (*MobiCom*), September 2009.

[Aura] Project Aura: distraction-free ubiquitous computing http://www.cs.cmu.edu/~aura/

[AWB+11] Ardalan Amiri Sani, Wolfgang Richter, Xuan Bao, Trevor Narayan, Mahadev Satyanarayanan, Lin Zhong, Romit Roy Choudhury, "Opportunistic Content Search of Smartphone Photos," *Technical Report TR0627-2011*, Rice University, June 2011.

[BCM+09] Paramvir Bahl, Ranveer Chandra, Thomas Moscibroda, Rohan Murty, Matt Welsh, "White Space Networking with Wi-Fi like Connectivity," in Proc. ACM SIGCOMM, 2009.

[Cis11] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010–2015", February 2011 http://newsroom.cisco.com/ekits/Cisco_VNI_Global_Mobile_Data_Traffic_Forecast_2010_2015.pdf

[CPL+] C. K. Chan, H. Peng, G. Liu, K. McIlwrath, X. F. Zhang, R. A. Huggins, and Y. Cui, "High-performance lithium battery anodes using silicon nanowires," in *Nature Nanotechnology*, vol. 3, No. 1, pp. 31--35, 2007.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," *IEEE Computer Vision and Pattern Recognition (CVPR),* 2009.

[DDT+08] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," in *IEEE Signal Processing Magazine,* 25(2), 83–91, 2008.

[FEW08] K. Fife, A. El Gamal and H.-S. P. Wong, "A Multi-Aperture Image Sensor with 0.7um Pixels in 0.11um CMOS Technology," in *IEEE Journal of Solid-State Circuits*, pp. 2990-3005, December 2008.

[Fif09] Keith G. Fife, Devices for integrated multi-aperture imaging, *Ph.D. Dissertation*, *Stanford University*, June 2009.

[Fra08] Gene A. Franz, "SoC in the new paradigm of IC technology," http://www.dallasces.org/talks/IEEEConsumerElectMtg-Dallas-August2008.pdf

[Fra09] Gene Frantz, 2020 Vision: Transistors a buck a billion: http://www.eetimes.com/electronics-news/4196903/2020-Vision-Transistors-a-buck-a-billion

[GMG+11] L. Gasparini, R. Manduchi, M. Gottardi, D. Petri, "An ultralow-power wireless camera node: development and performance analysis," in *IEEE Transactions on Instrumentation and Measurement*, vol.60, no.12, pp.3824-3832, Dec. 2011

[Goo12] Google, Project Glass: Thoughts, designs, and stories: http://g.co/projectglass

[HQW+10] Rehan Hameed, Wajahat Qadeer, Megan Wachs, Omid Azizi, Alex Solomatnikov, Benjamin C. Lee, Stephen Richardson, Christos Kozyrakis, and Mark Horowitz, "Understanding sources of inefficiency in general-purpose chips," in *SIGARCH Comput. Archit. News* 38, 3, June 2010, 37-47.

[HQG+12] Junxian Huang, Feng Qian, Alexandre Gerber, Z. Morley Mao, Subhabrata Sen, and Oliver Spatscheck, "A Close Examination of Performance and Power Characteristics of 4G LTE Networks" in *Proc. ACM Int. Conf. Mobile Systems, Applications, and Services* (*Mobisys*), June 2012.

[HWB+06] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood, "SenseCam: a retrospective memory aid," in *Proc. Int. Conf. Ubiquitous Computing* (*UbiComp*) 2006.

[HXT+10] Junxian Huang, Q. Xu, B. Tiwana, Z. Morley Mao, Ming Zhang, and Paramvir Bahl, "Anatomizing Application Performance Differences on Smartphones," in Proc. ACM Int. Conf. Mobile Systems, Applications, *and Services* (*Mobisys*), June 2010

[II03] Stephen S. Intille and Amy M. Intille, “New challenges for privacy law: wearable computers that create electronic digital diaries,” MIT House_n Technical Report, September 2003.

[INV] C. Chabris and D. Simons, the invisible gorilla: http://www.theinvisiblegorilla.com/videos.html

[KGS+05] Purushottam Kulkarni, Deepak Ganesan, Prashant Shenoy, and Qifeng Lu, “*SensEye*: a multi-tier camera sensor network,” in *Proc ACM Int. Conf. Multimedia* (MULTIMEDIA), 2005.

[KP10] D. W. F. van Krevelen and R. Poelman, “A survey of augmented reality: technologies, applications and limitations,” in *The International Journal of Virtual Reality*, 2010, 9(2):1-20.

[LM99] N. Jesper Larsson and Alistair Moffat, “Offline Dictionary-Based Compression,” in *Proc. IEEE Conf. Data Compression* (DCC), 1999.

[LPL+09] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell, “SoundSense: scalable sound sensing for people-centric applications on mobile phones,” in *Proc. ACM Int. Conf. Mobile Systems, Applications, and Services* (*MobiSys*), June 2009.

[LSL11] J. A. Leñero-Bardallo, T. Serrano-Gotarredona, and B. Linares-Barranco, “A 3.6 μs Latency Asynchronous Frame-Free Event-Driven Dynamic-Vision-Sensor,” in *IEEE Journal of Solid-State Circuits*, June 2011.

[LSW+12] He Liu, Stefan Saroiu, Alec Wolman, and Himanshu Raj, “Software Abstractions for Trusted Sensors,” in *Proc. ACM Int. Conf. Mobile Systems, Applications, and Services* (*Mobisys*), June 2012.

[MCT-10] Rohan Murty, Ranveer Chandra, Thomas Moscibroda, Paramvir Bahl, “SenseLess: A Database-Driven White Spaces Network,” *IEEE DySpan* 2011

[NAG09] A. Nilchi, J. Aziz, and R. Genov, "Focal-Plane Algorithmically-Multiplying CMOS Computational Image Sensor," in *IEEE Journal of Solid-State Circuits*, vol.44, no.6, pp.1829-1839, June 2009

[NG09] Ren Ng, Digital light field photography, Ph.D. dissertation, Stanford University, 2009.

[Pix08] Pixim Inc., White Paper: Digital pixel system technology, May 2008: http://www.pixim.com/assets/files/product_and_tech/Digital_Pixel_System_Technology_White_Paper_June_2_08.pdf

[Pow95] R. A. Powers, “Batteries for low power electronics,” in *Proceedings of the IEEE*, vol. 83, No. 4, pp. 687-693, 1995.

[RBI+05] Mohammad Rahimi, Rick Baer, Obimdinachi I. Iroezi, Juan C. Garcia, Jay Warrior, Deborah Estrin, and Mani Srivastava, “Cyclops: in situ image sensing and interpretation in wireless sensor networks,” In Proc. *ACM Int. Conf.* Embedded *N*etworked *S*ensor *S*ystems (*SenSys*), 2005.

[RSG+09] A. L. M. Reddy, M. M. Shaijumon, S. R. Gowda, and P. M. Ajayan, “Coaxial MnO2/carbon nanotube array electrodes for high-performance lithium batteries,” in *Nano Letters*, vol. 9, No. 3, pp. 1002--1006, 2009.

[SH09] Stanislava Soro and Wendi Heinzelman, “A survey of visual sensor networks,” in *Advances in Multimedia*, 2009.

[SBC+09] Mahadev Satyanarayanan, Paramvir Bahl, Ramon Caceres, and Nigel Davies, “The Case for VM-Based Cloudlets in Mobile Computing,” in *IEEE Pervasive Computing* 8, 4, 14-23, 2009.

[SSS08] Noah Snavely, Steven M. Seitz, and Richard Szeliski, “Modeling the world from Internet photo collections,” in *International Journal of Computer Vision*, 80:189-210, 2008.

[Sze10] Richard Szeliski, *Computer Vision: Algorithms and Applications*, Springer. Electronic version available from http://szeliski.org/Book/, the latest version September 2010

[WC05] Jue Wang and Michael F. Cohen, “Very Low Frame-Rate Video Streaming For Face-to-Face Teleconference,” in *Proc. IEEE Conf. Data Compression*, pp. 309-318, 2005.

[WM12] A. Wang and A. Molnar, “A Light-Field Image Sensor in 180 nm CMOS”, in *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, Jan. 2012.

[WMN11] Rebecca M. Willett, Roummel F. Marcia and Jonathan M. Nichols, "Compressed sensing for practical optical imaging systems: a tutorial", in *Opt. Eng.*, Jul, 2011.

[ZZW+11] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya and Venu Vasudevan, "Human as real-time sensors of social and physical events: a case study of Twitter and sports games," *Technical Report TR0620-2011*, Rice University and Motorola Mobility, June 2011 (http://www.sportsense.org).