Authors

David Moloney, Movidius Ltd., O'Connell Bridge House, Dublin 2, Ireland

Oscar Deniz Suarez, VISILAB, Av. Camilo José Cela, s/n 13071 · Ciudad Real · España

A Vision for the future

For the past 40 years, computer scientists and engineers have been building technology that has allowed machine vision to be used in high value applications from factory automation to Mars rovers. However, until now the availability of computational power has limited the application of these technologies to niches with a strong enough need to overcome the cost and power hurdles. This is changing rapidly as the computational means have now become available to bring computer vision to mass market applications in mobile phones, tablets, wearables, drones and robots enabling brand new user-experiences within the cost, power and volumetric constraints of mobile platforms.

"Live in the future, then build what's missing" – Paul Graham

Computational imaging represents an historic transition from the 150-year old paradigm of taking photos on silver halide photo stock, chemically developing and printing them, to computing (making) pictures. According to Hayes [1] a digital camera is no longer a passive recording device, and is an image creation rather than a simple recording device. In existing digital still cameras, the focus is on making digital images identical to their chemical forebears. However, once the camera contains sufficient image-processing horsepower, a computational camera can move beyond the reality captured by conventional digital cameras. The sensor array in such a camera plays the role film used to, but it's the beginning rather the end of the image creation process.

In his book describing the work of photography pioneer Ansel Adams in the 1920s, 30s and beyond, William Turnage [2] claims Adams spent up to a day per print effectively doing manually what today would be called High Dynamic Range (HDR) photography: "[Adams] always said that the negative is the equivalent of the composer's score and the print is the equivalent of the conductor's performance". Similarly the concept of multi-aperture (array) and lightfield (plenoptic) cameras which allow depth to be recovered from images and enable a range of Depth-of-field (DoF) and other effects to be implemented computationally, have been around since the turn of the 20th century since Lippmann demonstrated his 3x4 array camera in 1911 [3], essentially waiting 100 years for the computational means to catch up with his vision.

Current computational photography architectures [4][5][6] focus very much on image capture and post-processing in the point and shoot camera paradigm within existing smartphone architectures. In this model image capture is controlled using APIs such as fCam, Android Camera 2.0 and the forthcoming Camera 3.0 API (Application Programming Interface) for Android (derived from fCam). These APIs are aimed at opening up the previously closed world of the camera and to some extent the camera Image Signal

Processing (ISP) internals to the application developer.

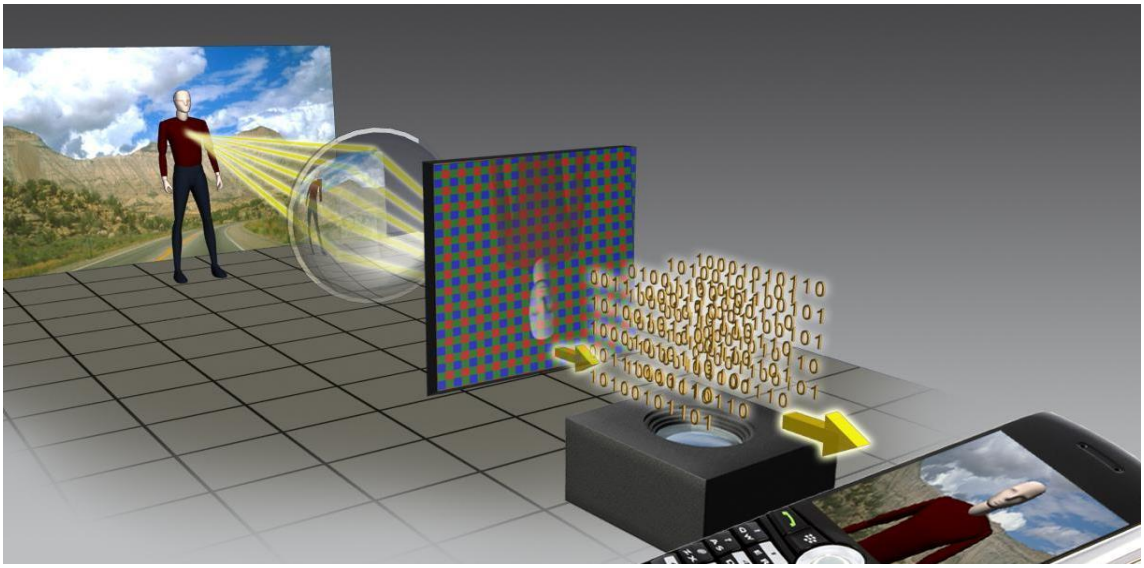

Figure 1 Imaging and Vision Capture and Processing Chain

This fine control of the capture process enables applications based on focal stacks (rapidly captured bursts of images with different camera settings) in smartphones such as High Dynamic Range (HDR) image and video capture, synthetic aperture photography where depth of field (DoF) can be varied to focus on different parts of the image, removal of unwanted persons or objects within images, best shots of groups (everybody smiling, facing the camera and not blinking etc.) and sophisticated ISP functions such as alignment of multiple raw images for super-resolution etc.

The Android Camera 3.0 API makes extensive use of burst mode, capturing sequences of high-resolution images, and associated depth information, at high rates in smartphones, leveraging existing GPU cores using computational imaging algorithms using Renderscript to deliver the best images in a platform-independent manner. Similar efforts to standardise camera APIs, the capture and fusion of sensor metadata and hardware acceleration for computer vision are ongoing within the Khronos industry driven standards organisation, as well as more domain-specific initiatives such as Advanced Driver Assistance Systems (ADAS) within the automotive community driven by EU NCAP and US safercar safety standards.

The APIs and programming models proposed in [4][5][6] aim to leverage the existing hardware processing resources by layering software to marshal the heterogeneous computational resources in Application Processors (APs) to do a new job they were never designed to do. Generally the hardware architectures underlying these APIs, and currently used to deliver these advanced camera functions, are multicore CPUs, DSPs and GPUs already present in existing or derivative Application Processors (APs) which are widely used in smartphones and tablets. The result is a compromise that constrains the computational capability of the system, delivers poor user experiences and very poor battery life.

Moving Beyond Taking Pictures

One of the defining characteristics of computer vision is that it turns into a means of making measurements and inferences about the world and taking action based on those measurements rather than simply capturing a scene by measuring the light incident at each pixel in an array. On a very basic level many of us have unwittingly been using massive numbers of cameras to measure the world for the past 15 years since Microsoft introduced the first camera mouse in 1999. These optical mice use a combination of an LED mounted at a glancing angle and simple camera to extract the underlying texture in any surface the mouse is moved across, with post-processing to extract a motion vector which is used to control the cursor on a PC screen. In the past few years new human interaction devices such as Microsoft's Kinect, Leap Motion's camera-based gesture device and Tobii's eye-tracker have begun to revolutionise the way we interact with digital content on PCs and gaming consoles.

Figure 2 HP Sprout workstation

More recently devices like HP's Sprout use a combination of a physical LCD touchscreen, and downward facing video projector coupled with a computer-vision enabled horizontal touch surface (TouchMat) to build a user interface that extends outside the box, allowing the user to interact with the real world in new and exciting ways, capturing and manipulating 2D and 3D digital content in a highly intuitive manner. These capabilities have also begun to appear in mobile devices such as Amazon's FirePhone [26] which boasts four dedicated computer-vision (global-shutter) cameras mounted in the corners of the device along with associated IR LEDs in addition to the conventional front and rear-facing cameras. The four cameras

allow novel UI features such as viewpoint dependent 3D rendering which adapts to the users head-position, HDR image capture, cloud-backed image search, etc.

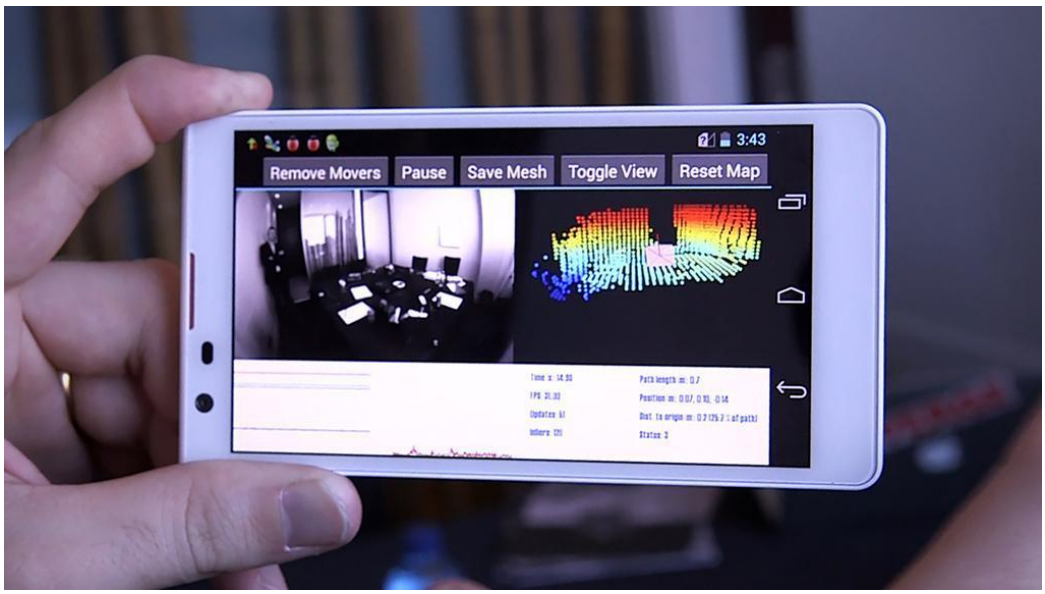

Figure 3 Project Tango Phone

The next step is of course to move beyond the limitations of our personal devices entirely using telepresence and autonomous drones and robots as well as embedding vision in a broader range of products. A key enabler for these use-cases is Simultaneous Localization and Mapping (SLAM), widely used by autonomous robots operating in unknown environments and developed in the early 1990s for Mars rovers [7][8]. In such systems, SLAM software turns a conventional RGB camera into a 6 DoF (Degrees of Freedom) transducer and the autonomous mobile platforms enabled by it must be able to reliably measure ego-motion as otherwise long-term navigation is impossible, and conventional means of location determination versus a known reference such as GPS are unavailable. Stereo odometry determines the ego-motion of a stereo camera in the 6 degrees of freedom (DoF) that are possible in the 3D world (3 for translation and 3 for rotation) and compensates for wheel slippage which can otherwise cause problems with odometry. Furthermore sensor fusion and tracking are also integral components of many autonomous vehicles and robots.

Until very recently SLAM algorithms have been academic in nature and have not been optimised for embedded platforms. An approach to implementing SLAM on embedded platforms leveraging SIMD coprocessors, DSP and multi-core CPUs is outlined in [9] and rapid progress in embedding such technology is being made through initiatives like Google's Project Tango [19] and SLAMbench [10]. A good example of a consumer device incorporating SLAM is the Dyson 360 Eye robotic vacuum cleaner [25] and we can expect that such systems will rapidly fall in price and increase in capabilities in the coming years driven by advanced semiconductors.

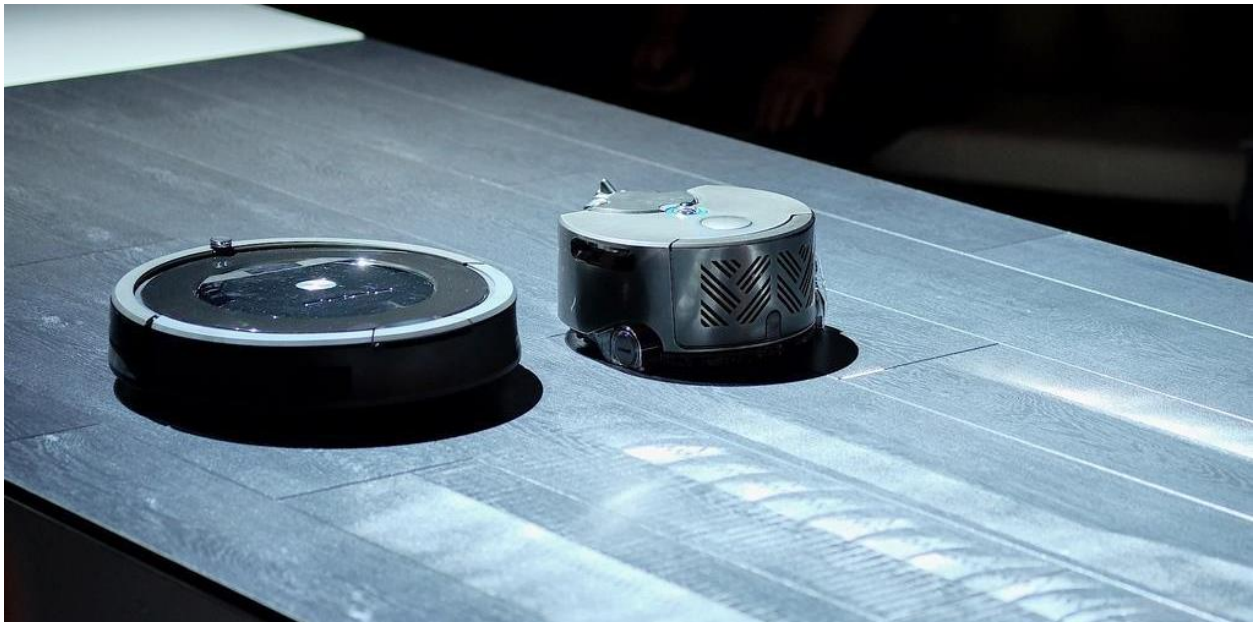

Figure 4 Dyson 360 Eye

Returning to human vision, the idea of Virtual Reality has been around for a long time but the experience never delivered on the promise with many of the devices either not working or inducing nausea in the unfortunate wearer. The key issue according to Abrash [11] is latency or in other words the delay between head motion and the corresponding virtual world update reaching the eyes. Too much latency results in images drawn in the right place, but at the wrong time, creating anomalies which are amplified by head motion; the faster the head moves, the greater the anomaly, compounded further when the head change direction. Moving beyond the confines of traditional VR devices like Oculus Rift companies like Magic Leap [24] who recently raised over $542M from Google and other investors hint at an intriguing world of seamless mixed reality blending rendered graphics with reality using advanced displays and enabled by computer vision.

Figure 6 Rendering of Mixed Reality by Magic Leap

Latency is even more demanding in applications like active headlights in cars, such as the CMU SmartHeadlight programme directed by Prof Takeo Kanade [12] where the total round-trip latency from snowflake or raindrop reflection, through sensor, computer-vision hardware and closing the loop by modulating an LED projector array in the headlight is on the order of 1500ms.

Figure 7 ProxDynamics PD-100 Drone

Another important application for computer vision is in giving an out-of-body-experience by mounting a camera on a drone and having it track the user for instance to make sports and activity videos. These devices have just begun to appear on the market and it can be expected that such devices will proliferate the consumer electronics space in the coming years as prices fall and fly-time increases. Weight of both battery, electronics and airframe as well as the sensor array are key concerns with the electronics often being the limiting part of programs like sFly [22], which achieved a 10min autonomous fly time and required about 1W for a 100g payload. The military, law-enforcement and the emergency services are being targeted by dedicated price-is-no-object devices such as the ProxDynamics Black Hornet PD-100 drone helicopter [23] which weighs in at a miniscule 16g (equivalent to 3 sheets of A4 paper) and boasts 3 cameras and a wireless link that operates to over 1km range.

Beyond UI, smart cameras can also be used to measure other phenomena for instance Eulerian Video Magnification [13] can be used to amplify natural movement, luminance or colour changes and make it visible to the human eye much in the same way as time-lapse photography allows us to see plants grow or clouds move across the sky. We can expect such features to appear in medical equipment and baby-monitors in the not too distant future. Eulerian magnification can also be used to reconstruct sound [14] that causes a plant's leaves to move in a sealed room when fed into an appropriate inverse acoustical model, however the civilian applications for this technology are less clear.

Computer vision and computational photography are undoubtedly in a rapid growth phase with new vision-based applications appearing on a regular basis. One prominent example is Placemeter, which uses public video feeds (from old unused smartphones attached to windows) and computer vision algorithms to create the first ever, real time layer of data about places, streets and neighbourhoods. Placemeter [21] collects and serves up-to-the-minute information like how crowded a place is, how long the wait is, and whether it will get more or less crowded in the next hour. There are many more examples: image search, panoramas, face detection in smartphones' cameras, face recognition biometrics to unlock devices, video stabilization in YouTube, Facebook's facial recognition for photo tagging etc.

Eyes Everywhere

It is clear that computer vision is on an explosive growth phase transitioning from traditional automation in factories to the world in which people work, live, and play. While computer vision is a mature research field from a theoretical point of view, practical ubiquitous vision has not progressed to the same extent as systems have been confined to labs and factories. The underlying difficulties are the primarily the computational and cost requirements imposed by consumer devices. Human vision is immediate, we open our eyes and can immediately recognize and categorize objects and the structures of scenes. What we do not realize is the vast computational resources that our brain brings to bear unconsciously to make all of this happen. When our eyes are open, vision accounts for two-thirds of the electrical activity of the brain, and the sensors, our eyes, are located right there in the same casing as our brains. That our brain manages to delivery all of this functionality, which we can only dream of duplicating in a machine, it is even more amazing when we think that our brain consumes a mere 20W and is fuelled by environmentally-friendly renewable sugars. This being said even our crude approximations of human vision are now starting to yield useful results in practical power and cost envelopes.

With the advent of cloud computing, it is tempting to think that the cloud alone can bear the computational load associated with vision processing. However, crunching the enormous volume of visual data is currently beyond the reach of all but a few very large companies, and network bandwidths cannot cope with the massive use of cameras. Power efficiency is a major issue and wirelessly transmitting data for remote computation can cost up to a million times more energy per operation compared to processing locally in a device. A further reason for pushing processing and even limited decision making to the network edge in applications like virtual or augmented reality, autonomous vehicles and robots is that data-centres have major issues with latency. Finally, privacy in the cloud is a concern, especially when we consider that more often than not, the subject of our monitoring will be humans. Our own brains are co-located with our eyes, processing visual data close to the point of origin and our reflexes deal with the latency of transmitting stimuli to our brain and back to the muscles using local neural circuits to save time and get us out of “harm’s way”.

Placing the computational resources close to optical and other sensors in Cyber-physical Systems clearly solves many of the same issues. Going beyond this perhaps we require not only “intelligence everywhere” but also “eyes everywhere” for many applications. In the

scientific and technical literature, the closest systems to this 'eyes everywhere' paradigm are to be found under the terms 'embedded vision' and 'vision sensor networks'. On the one hand, embedded vision refers to vision systems that are integrated into more complex devices such as automotive or medical equipment. Those systems are a natural evolution of fixed industrial vision systems based on PCs and smart cameras. While such systems are increasingly used, power and size requirements are not so stringent and development is typically application-specific. Vision sensor networks, on the contrary, aim at smaller power-efficient vision sensor nodes that can operate standalone, however the proliferation of standards and lack or interoperability has frustrated widespread adoption and commercialization, causing some researchers [15] to observe: "Given the fair number of proposed designs, it is somewhat surprising that a general-purpose embedded smart camera, based on an open architecture is difficult to find, and even more difficult to buy."

Current heterogeneous platforms such as mobile phone application processors can be used to prototype algorithms but from published results they appear limited to VGA 30fps resolution and around 5W power dissipation with technologists like John Carmack, CTO of Occulus [17] saying that Application Processor power dissipation limits realistic VR experiences to around 15 minutes. Even then implementing relatively simple pipelines on these platforms involves a high level of complexity involved in the coordination of multiple ARM processors, NEON SIMD extensions and GPUs all of which must share access to data-structures in memory through shared access or copying data. In the case where data is copied to and from the GPU it means that a certain level of granularity in terms of GPU tasks is required to make the effort of moving data to and fro, worthwhile. This complexity must either be managed explicitly by the programmer with a mixture of C/C++ code for the ARM, SIMD assembler for NEON and OpenGL ES or OpenCL shaders or via vendor supplied libraries and APIs.

Other options such as miniature IP cameras and smart cameras do not get close to the required size, energy consumption, cost and processing power. On the related level, in 2014 Freescale released the Wearable Reference Platform (WaRP) [16]. WaRP is the first attempt to provide a reference platform for future development of wearable devices. While it is open, small and can be connected to cameras, WaRP has not been designed for mobile embedded vision, which is the most challenging capability in terms of required processing power and energy consumption.

In this context, innovative vision applications are typically based on (little effective) DIY kits, smartphones or else are supported by large companies that can fund the specific hardware and software developments needed. Research and academia do not seem to have a versatile platform for deploying innovative vision applications and for rapidly designing new products based on the latest advances in the field.

A Myriad of new possibilities

In the aforementioned context, Movidius is focused on bringing human vision and scene understanding to mobile devices allowing low power always-on vision capabilities in devices with the very low latencies required for interactive services, self-driving cars and robots etc. Movidius is enabling a swathe of new computer vision applications to be brought to the

mass market for the very first time in embedded devices such as mobile phones, tablets and cameras. The first generation of the Movidius Myriad Vision Processing Unit (Myriad1 VPU) powers the computer vision subsystem in Google's Project Tango (see Figure 3) where it handles all of the high performance ISP, feature tracking and tracking tasks in 10x less power than any other solutions available on the market today [19]. In August 2014 Movidius introduced its second generation Myriad2 which was designed to achieve 20-30x the processing per watt of the previous generation Myriad1. Myriad2 is a System-on-Chip (SoC) that embeds a software controlled multi-core, multi-ported memory subsystem and caches which can be configured to allow a large range of workloads to be handled, providing exceptionally high sustainable on-chip data and instruction bandwidth to support the twelve processors, 2 RISC processors and high-performance video hardware accelerator filters. Supporting a sustained throughput of 600Mpixels/sec means that Myriad2 can deliver 1080p120 with less than 20% of the available pixel bandwidth. As a result of the highly power-efficient architecture of Myriad2 an OpenCV compatible multi-scale Haar Cascade consisting of 20 stages, computed using twelve SHAVEs and one of the HW accelerators in Myriad2 can calculate 50,000 multi-scale classifications for each 1080p resolution frame, in less than 7 msec (a key requirement for immersive VR [11]).

Because of this performance and the focus on embedded vision systems Myriad2 has been recently selected to power the "Eyes of Things" (EoT) platform [18], an Innovation Action funded by the European Union's Horizon 2020 Framework Programme for Research and Innovation. The objective is to build an embeddable always-on computer vision system which can be used as a generic platform for any applications requiring mobile/embedded vision. The hardware and software infrastructure developed in EoT will be available to OEMs and the public in general, and applications for EoT nodes are myriad including: wearable, UAVs, robotics, surveillance, etc. It is estimated that the use of the EoT platform will save up to 41% in time-to-market for advanced vision-based applications.

Conclusions

It is clear that a wide range of computational photography and computer vision techniques can greatly enhance user experiences and while these techniques are well understood they are in a considerable state of flux which precludes the use of fixed-function hardware as there is no "one-size-fits-all" device or algorithm for many or all of these applications. In this context a software programmable device which is optimised for vision workloads clearly offers an optimal balance of power and performance. It is our belief that Vision Processing Units (VPUs) will become the industry standard way of handling such workloads using open APIs like Khronos OpenVX [20] in the same way that GPUs are used to process computer graphics workloads and Myriad is simply the first example of this class of device to make it to market. As we see it the availability of capable, flexible and power efficient homogeneous processor architectures architected for computer vision will allow users to begin to expect to "Live in the future" where always-on computer vision applications are possible at HD resolution, dissipating only 100s of mW. These platforms will be utilised in a range of use-cases including standalone AR displays, wearable devices, UAVs, mobile robots, surveillance cameras and also tablets and phones where the computer vision processor offloads demanding tasks from the AP. For the first time the device designers will focus on

designing world-beating products with revolutionary CV functionality in record timescales and with exceptionally low power requirements.

References

[1] B. Hayes, "Computational Photography", American Scientist, March-April 2008, Volume 96, Number 2, pp.94-99.

[2] W.A. Turnage, “Ansel Adams: Our National Parks”, Little, Brown and Company; 1st edition (May 21, 1992).

[3] “Integral Photography: A New Discovery by Professor Lippmann”, Scientific American, 19 Aug 1911, pp.164.

[4] S.H. Wang, U. Timofei, “Computational Photography. A new way to look at image capture”, SAMSUNG Whitepaper, 2013.

[5] NVIDIA, “Chimera™: The NVIDIA Computational Photography Architecture”, http://www.nvidia.com/object/white-papers.html

[6] Qualcomm, “Qualcomm Snapdragon 800 product brief”, https://www.qualcomm.com/media/documents/files/snapdragon-800-processor-product-brief.pdf

[7] R. Smith, M. Self, and P. Cheeseman. 1990. “Estimating uncertain spatial relationships in robotics”. In Autonomous robot vehicles, Ingemar J. Cox and Gordon T. Wilfong (Eds.). Springer -Verlag New York, Inc., New York, NY, USA 167-193.

[8] J.J. Leonard, HF Durrant--Whyte, “Mobile robot localization by tracking geometric beacons”, Robotics and Automation, IEEE Transactions on 7 (3), 376-382, 1991.

[9] B. Vincke, A. Elouardi and A. Lambert, “Real time simultaneous localization and mapping: towards low- cost multiprocessor embedded systems”, EURASIP Journal on Embedded Systems 2012, 2012:5.

[10] L. Nardi, B. Bodin, M. Z. Zia, J. Mawer, A. Nisbet, P. H. J. Kelly, A. J. Davison, M. Luján, M. F. P. O'Boyle, G. Riley, N. Topham, S. Furber, “Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM”, 8 Oct 2014, http://arxiv.org/abs/1410.2167v1

[11] M. Abrash, “Latency – the sine qua non of AR and VR”, Posted December 2012, http://blogs.valvesoftware.com/abrash/latency-the-sine-qua-non-of-ar-and-vr/

[12] R. Tamburo, E. Nurvitadhi, A. Chugh, M. Chen, A. Rowe, T. Kanade and S. G. Narasimhan. “"Programmable Automotive Headlights”, European Conference of Computer Vision (ECCV), 2014. Lecture Notes in Computer Science. Volume 8692, 2014, pp 750-765.

[13] HY Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand and W. T. Freeman, “Eulerian Video Magnification for Revealing Subtle Changes in the World”, ACM Transactions on Graphics (Proc. SIGGRAPH 2012), vol. 31, No. 4.

[14] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand and W.T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video", ACM Transactions on Graphics (Proc. SIGGRAPH, 2014, vol. 33, no. 4, pp.79:1-.

[15] B. Murovec, J. Pers, R. Mandeljc, V. Kenk and S. Kovacic, "Towards commoditized smart-camera design," Journal of Systems Architecture, vol. 59, no. 10, p. 847–858, 2013

[16] http://www.warpboard.org/

[17] http://fortune.com/2014/12/29/oculus-vr-john-carmack-extended/

[18] http://eyesofthings.eu

[19] http://www.extremetech.com/extreme/187232-movidius-the-chip-maker-behindgoogles-tango-wants-to-be-the-king-of-computational-photography

[20] https://www.khronos.org/openvx/

[21] http://www.placemeter.com/

[22] http://www.sfly.org/

[23] http://gizmodo.com/5981975/black-hornet-the-195000-spy-plane-that-fits-in-the-palm-of-your-hand

[24] http://uk.businessinsider.com/rony-abovitz-magic-leap-2014-12?r=US

[25] https://www.dyson360eye.com/

[26] http://www.amazon.com/Fire_Phone_13MP-Camera_32GB/dp/B00EOE0WKQ