Sensors for Future VR Applications

Chiao Liu, Michael Hall, Renzo De Nardi, Nicholas Trail, Richard Newcombe
Oculus Research, Facebook Inc.
1 Hacker Way, Menlo Park, 94025
Chiao.liu@oculus.com
Phone: 1-425-628-5304

Abstract:

Future generations of virtual reality (VR) systems will incorporate multiple sensors for various capture, tracking, mapping, reconstruction, and other machine perception functions. In this paper, we provide examples of some tracking and mapping functions that illustrate the critical requirements and performance metrics. The sensor performance, form factor, power, and data bandwidth are the main challenges in a battery powered, always on VR devices. We further propose a new figure of merit that incorporates both sensor power consumption and SNR into a single parameter. To overcome bandwidth, compute, and power challenges, on-sensor signal processing and early information extraction are necessary. We expect stacking technology will be the key enabler of new sensor architectures, and innovations on both sensing and processing layers will deliver intelligent machine perception sensors for the future generations of VR devices.

Introduction

Today's virtual reality (VR) systems are the beginning of the next generation computing platform. This journey, over the span of 10-15 years, will eventually reach billions of users. Future generations of VR systems will incorporate multiple sensors for various capture, tracking, mapping, scene reconstruction, and other machine perception functions. Unique requirements for these sensors include global shutter operation, very short exposure time to minimize motion blur, adequate performance in low light conditions and over a wide dynamic range, and ultra-low power consumption to support always-on functionality in a wearable battery powered device. At the system level, a large quantity of data is generated and must be processed in real time to support flawless interaction between the physical and virtual world, and to create a true "presence" experience for the user. The processing power and latency constraints pose significant challenges to current state-of-the-art technologies.

The core sensing and tracking functions for VR

In virtual reality, presence is the perception of being physically immersed in a non-physical world. The perception is created by surrounding the user of the VR system in images, sound or other stimuli that provide an engrossing total environment. From the virtual perception point of view, this can be summarized as delivering the right photons to the right places on the user's retina at the right time. Sensors are an integral part of this engineering challenge. We need sensors to capture the photons in the physical world so we can map out and reconstruct the physical surroundings, and mix and manipulate the physical world with virtual objects. We need sensors to measure the user's head pose and gaze direction. Moreover, all of these have to be done in real time with minimal latency. Figure 1 summarizes the core sensing and tracking functions of VR devices.

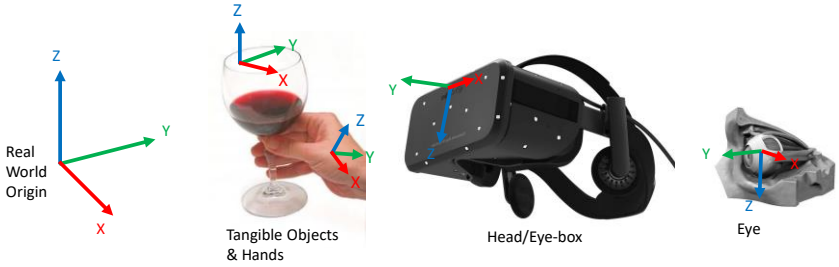


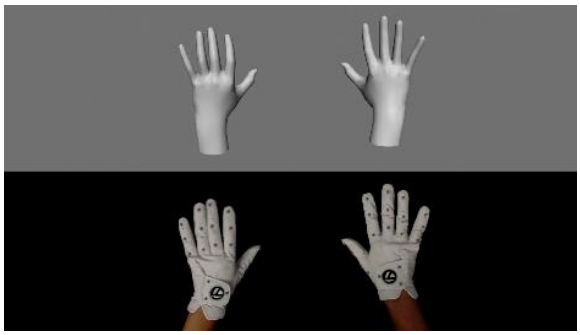
Figure 1 sensing and tracking functions of VR devices

Sensing and reconstruction of the real world require not only capturing the color and texture of your surrounding as in traditional photography, but also capturing precise geometric data and mapping out the spatial relationship among objects in 3D space. This is required for augmenting the physical world, and for accurately rendering occlusion when mixing physical and virtual objects. Lighting conditions vary widely, ranging from dark corners of the room to bright direct sunlight, so sensors must have a wide dynamic range. For depth sensors that use active NIR illumination, ambient light sources and the wide variety of object surface textures pose significant performance challenges [1].

Head mounted devices (HMD) use pose tracking to deliver the perception of presence. Tracking a user's head orientation in real time allows for dynamically adjusting images displayed on screen in front of each eye as to mimic the stereoscopic view that would be observed in reality. Head tracking requires high accuracy in six degrees of freedom (6DoF) and low latency to deliver true presence. Studies have shown that the Just Noticeable Difference (JND) shall be below 17 ms from head movement to updated display [2], and image capture and transmission must be a fraction of that. Simultaneous localization and mapping (SLAM) is the enabling computer vision technology [3, 4, 5].

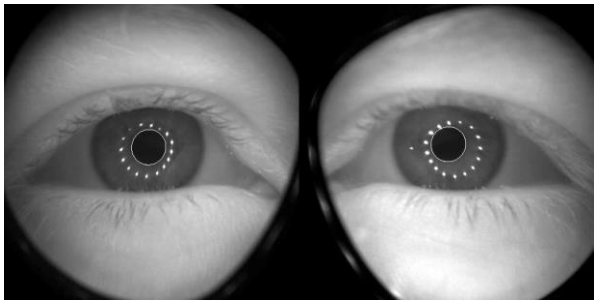
Eye tracking requires another 6 degrees of freedom per eye [6]. These sensors should be operated at a higher frame rate, around 240fps or higher, in order to capture rapid eye movements such as saccades. Eye tracking sensors are typically operated at near IR wavelengths, and stringent eye safety considerations requires excellent NIR sensor sensitivity over very short exposure times. The physical size of the camera module is also a challenge, as all of these light sources and cameras should fit comfortably on a light and stylish frame. Miniature camera modules hold great promise here, but the optical quality cannot be compromised.

No device is better than a user's own hands for interacting with the virtual world. Today, near-perfect hand tracking is possible [7], but it requires retroreflector-studded gloves and lots of rack mounted cameras. The goal is to replace those cameras with low power, small form factor hand tracking sensors. The hand tracking cameras could use either 2D passive sensors or a 3D depth sensor. They require high-speed operation at close to 100 frames per second, and must have a very wide field of view, approximately 140 x 120 degrees, to cover the whole range of hand and arm positions. The minimal detectable object has to be smaller than a finger.



(a) Hand tracking

(b) Facial tracking



(c) Gaze tracking

Figure 2. Example of various tracking functions

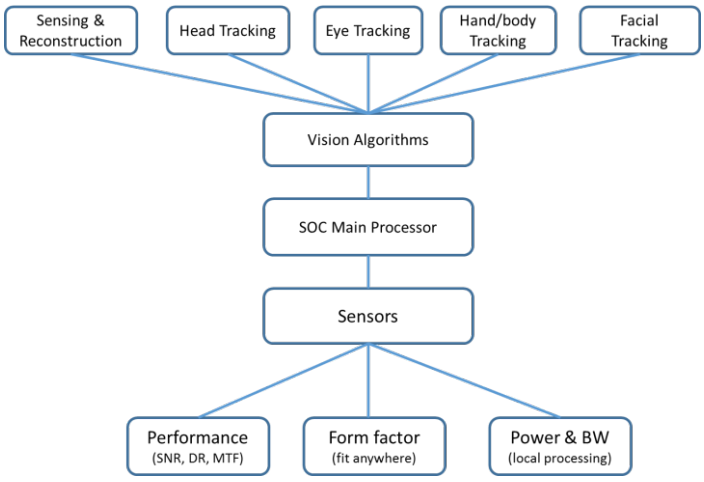


Figure 3 System stack diagram

Another element needed to create a virtual human in VR is facial tracking. Sensors used for facial expression capture operate in both visible and NIR light spectrum. The sensor needs to operate at 90-100 frame per second. Good image quality is important in capturing all the subtleties

Figure 2 provides examples of hand, facial and eye tracking functions. Figure 3 shows the system stack diagram. All the tracking functions are enabled by sophisticated computer vision algorithms. The algorithms have to be accurate and fast enough to deliver the best experience, but they also need to be power and memory efficient in order to fit into the computation resource budget that a mobile SOC processor can support. Sensors feed the data into an SOC. The sensor requirements are categorized by three pillars: performance, form factor, and power/bandwidth.

A New Figure-of-Merit

Previously in [5], a figure-of-merit was proposed to evaluate and compare different sensors in terms of power efficiency:

$$FOM = \frac{Energy}{bit} = \frac{Power}{\#of\ pixels \times FrameRate \times 2^N} \qquad (1)$$

Here N is the pixel bit depth.

For the applications described in this article, we are interested in more than raw data throughput. What is most important is the true information within a captured image. Therefore, we propose a modification to the figure-of-merit proposed in [5] in an effort to measure the energy per true information bit:

$$FOM = \frac{Energy}{SNR\_equivalent\_bit} = \frac{Power}{\#of\ pixels \times FrameRate \times 2^S} \qquad (2)$$

Where S is calculated as:

$$S = \frac{SNR_{dB}}{6dB} \qquad (3)$$

With this new definition, we normalize the sensor power consumption by its performance in term of SNR. When comparing different sensors' FOM, the SNR will be measured at the same condition that is close to use cases, for example, under 20 lux of a 3000K light source with a F/2 optics and 5 ms integration time. This new FOM enables us to benchmark different sensor products.

In Figure 4, we provide an example of how one type of measurement critical for machine perception depends on SNR at the sensor. Here, the detection probability is the fraction of recorded frames that result in an

unambiguous measurement. For low SNR, the sensor data is of poor quality and data received from many frames must be rejected. At high SNR, the data quality is high and measurements are reliable for all frames.

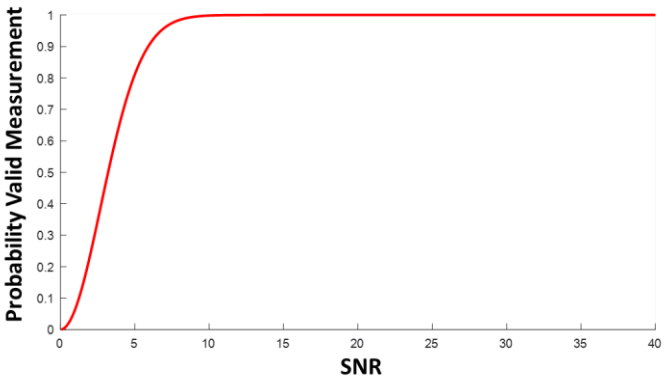


Figure 4. Probability of making a valid measurement as a function of sensor SNR.

Finally, to significantly reduce bandwidth and power, and to reduce the burden on the main SOC processor, we need on-sensor compute. There are various level of system integration from simple feature extraction and compression, to deep convolution neural network engine [9], to neural-inspired neuromorphic computing with high power and memory efficiency [10].

Conclusion

VR is the next generation computing platform. As the application examples described in this paper, multiple sensors are required for various capture, tracking, mapping, scene reconstruction, and other machine perception functions. Sensor requirements can be categorized into three areas: performance, form factor, and power/bandwidth. We propose a new figure-of-merit that normalizes the sensor power consumption to its performance under certain operation condition, which enables us to benchmark different sensor products. We believe stacked sensor technology will bring the next wave of innovations on both sensing and processing layers, and will deliver the truly intelligent machine perception sensors for the future generations of VR devices.

References:

[1] A.A. Dorrington, J.P. Godbaz, M.J. Cree, A.D. Payne, and L.V. Streeter, “Separating true range measurements from multi-path and scattering interference in commercial range cameras”. In IS&T/SPIE Electronic Imaging, pg. 786404–786404, 2011

[2] B. D. Adelstein, T. G. Lee, and S. R. Ellis, “Head tracking latency in virtual environments: Psychophysics and a model,” Human Factors and Ergonomics Society Annual Meeting Proceedings, 47, pp. 2083–2087, 2003

[3] G. Klein and D. Murry, “Parallel tracking and mapping for small AR workspaces”, ISMAR, 2007

[4] R. Newcombe, S. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time”, ICCV 2011.

[5] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry”, arXiv preprint arXiv:1607.02565, 2016.

[6] D. Hansen, and Q. Ji, “In the eye of the beholder: A survey of models for eyes and gaze”, IEEE Trans. on PAMI, 32(3):478–500.

[7] Robert Y. Wang and Jovan Popovi´c. “Real-time hand tacking with a color glove”. ACM Trans. on Graphics, 28(3):1–8, 2009

[8] S. Masoodian, A. Rao, J. Ma, K. Odame, and E. Fossum, “A 2.5pJ/b readout circuit for 1000fps single-bit Quanta Image Sensors”, IISW 2015

[9] G Desoli, et al., “A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems”, ISSCC 2017

[10] James, Conrad D., James B. Aimone, Nadine E. Miner, Craig M. Vineyard, Fredrick H. Rothganger, Kristofor D. Carlson, Samuel A. Mulder et al. "A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications." Biologically Inspired Cognitive Architectures (2017).