Digital Foveation: An Energy-Aware Machine Vision Framework

Ekdeep Singh Lubana and Robert P. Dick, *Member, IEEE*

***Abstract*—In machine vision applications, imaging systems and analysis algorithms are generally interdependent and energy intensive. We describe a machine vision energy minimization framework in which imaging hardware and vision algorithms are co-designed and tightly integrated. Digital foveation is inspired by the human vision system, which uses a spatially varying sensing architecture to generate oculomotory feedback and capture a series of high-resolution images using the densely sampling fovea. A multiround process with bidirectional information flow between camera hardware and analysis software optimizes energy consumption while preserving accuracy. By using existing hardware mechanisms, namely, row / column skipping, random access via readout circuitry, and frame preservation, digital foveation adapts to the chosen analysis algorithm. It aims to transmit and process only the necessary parts of the scene under consideration. This framework is general across a wide range of embedded machine vision applications and enables large improvements in energy efficiency. When evaluated for an embedded license plate recognition vision application, it reduces system energy consumption by 81.3% with at most 0.65% reduction in accuracy.**

***Index Terms*—Fovea, machine vision, multiresolution processing, multiround analysis.**

I. Introduction

MACHINE vision has transformed numerous practical domains—including security, healthcare, banking, and transportation. Its applications are expected to have a market value of $15.46 billion by 2022 [1]. However, the high energy consumptions of most such systems wastes money and limits deployment scenarios. Thus, efficient image analysis is essential for energy-constrained machine vision.

Cameras have generally been treated as black boxes; opportunities to adapt dynamically to the needs of specific imaging tasks are generally overlooked. We argue for an adaptive framework that uses energy-efficient techniques to adaptively gather problem-specific information in a multiround process, allowing efficient analysis without degrading accuracy. At the heart of our approach is the concept of dynamically varying the regions and resolutions transmitted by the camera under guidance by multiround image analysis algorithms.

It has been shown that in the presence of scene clutter, classification algorithms using kernels with fine sampling at the center and coarse sampling at the periphery perform better than uniform kernels [2]. A similar structure is found in the human vision system. The retina uses a central, dense sensing region called the fovea for high-resolution capture of a small portion of the scene; while a sparse, peripheral sensing region captures the rest of the image at low resolution. The low-resolution data are used for detecting objects of interest and generating oculomotory feedback. This allows the fovea to be directed, sequentially, to regions of interest, while efficiently building scene understanding. By using broad, coarse sampling to detect objects of interest and narrow, high-resolution sampling at the fovea, the optical sensory system reduces throughput across the vision pipeline, thus enabling efficient analysis.

Inspired by the foveated, variable-resolution architecture of biological vision systems, we developed and evaluated an algorithmic framework, called *digital foveation*, for energy-efficient image sensor control and image analysis. The framework discards information irrelevant to the analysis algorithm, while preserving details, in an application-oriented manner (see Fig. 1). In the general case, this system gathers images at varying resolutions. Under guidance by analysis algorithms, it determines corresponding locations for application-oriented transmission and processing. We experimentally evaluated digital foveation using low-resolution, uniformly sampled captures to enable identification of regions of interest. In subsequent rounds, the camera captures higher-resolution images in these regions. A key observation is that varying the resolutions of image regions to reduce camera and analysis energy consumption across the imaging pipeline requires minimal or no changes to camera hardware. This enables multiresolution, multiround analysis analogous to many biological vision systems.

Using sparse sampling for detection of regions of interest can result in removal of important information. Thus, the subsampling routine used should be capable of determining an ideal resolution to optimize energy consumption under an accuracy constraint. To this end, Digital foveation may use object size as a metric to adaptively determine the ideal subsampling levels for a given input, thereby meeting accuracy

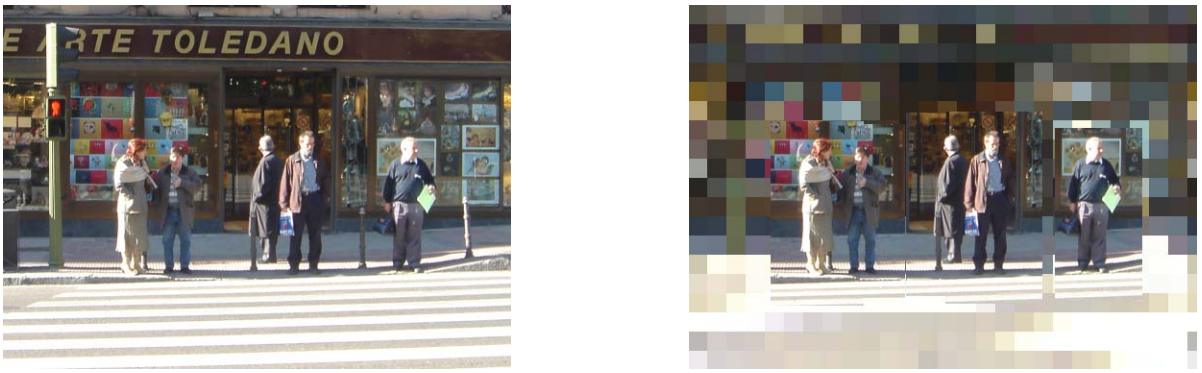
Manuscript received April 3, 2018; revised June 8, 2018; accepted July 2, 2018. Date of current version October 18, 2018. This article was presented in the International Conference on Hardware/Software Codesign and System Synthesis, 2018 and appears as part of the ESWEEK-TCAD special issue. *(Corresponding author: Ekdeep Singh Lubana.)*

E. S. Lubana is with the Department of Electronics and Communication Engineering, Indian Institute of Technology, Roorkee, Roorkee 247667, India (e-mail: ekdeeplubana@gmail.com).

R. P. Dick is with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48105 USA (e-mail: dickrp@umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2018.2858340



(a) (b)

Fig. 1. (a) Conventional uniform-resolution sampling and processing approach used in most machine vision applications. This is appropriate in image reproduction applications, where aesthetics are important, but not in energy-constrained machine vision applications. (b) By sampling irrelevant background information at low resolution and regions of interest at high resolution, energy consumption is dramatically reduced while preserving accuracy.

constraints set by the designer. For example, one might require the same number of pixels to accurately classify an object regardless of its original scaling. The approach is detailed in Section VII

This paper makes the following contributions. It describes a biologically inspired, multiround, variable-resolution framework for use in energy-efficient machine vision applications. Our approach uses existing hardware mechanisms, namely, row/column skipping, random access via readout circuitry, and frame preservation, to support adaptive control of the active sensing region (see Section IV). The framework is appropriate for a wide range of vision applications, i.e., any machine vision application for which it is possible to productively guide later sampling strategies by using information gathered in prior samples. The use of an analogous scheme in biological vision systems suggests that such applications are common. To the best of our knowledge, this is the first time such an approach has been modeled, described, and demonstrated.

Using a Sony IMX219 image sensor and Raspberry Pi 3, we evaluated digital foveation for license plate recognition, which accounts for roughly 100-billion image captures per year. We find that energy consumption reduces by 81.3% with at most 0.65% reduction in accuracy (see Section VII).

II. Related Work

This section summarizes related work on energy-aware machine vision. Most prior work has focused on narrow portions of the system, such as application-oriented circuits in the imaging pipeline or use of a time-efficient, software-only approach to multiresolution analysis.

Inspired by multiresolution biological vision systems and motivated by the need for energy efficiency, early researchers developed foveated, or retina-like, image sensors. This concept is distinct from the Digital foveation framework we propose. In foveated imaging, the resulting bio-mimetic sensors use custom hardware with spatial pixel distributions similar to those of retinas, containing a densely sampled central region (fovea) and sparsely sampled peripheral region. Conventional image processing algorithms cannot be directly used on the data acquired by such sensors. Algorithms designed to process nonuniform resolution images use analogous operations, e.g., chirp transform instead of discrete Fourier transform. However, due to their complexity and a lack of translational invariance, these algorithms have not been extensively studied [3]. An alternative approach is to use log-polar mapping [4] to approximate uniformly sampled images and use conventional processing algorithms; however, this is highly inefficient and thus defeats the purpose for using foveated image sensors. Further, the use of such sensors requires mechanical gimbals that impose speed, cost, reliability, and energy penalties [5].

Redeye [6], a convolutional neural network accelerator, shifted early processing (i.e., convolution operations) to the analog domain and outputted processed features. While their simulations indicate 45% reduction in energy consumption, Redeye ignores practical constraints introduced by the preprocessing pipeline. Since the (neglected) image signal processor (ISP) is programmed to digitally preprocess images for demosaicing the Bayer filter pattern, removing digital noise, and providing local contrast enhancement, substantial energy-relevant changes would be necessary for real-world use. It would be necessary to either carry out these steps in the analog domain or to convert between digital and analog domains multiple times, thereby potentially reducing the reported benefits.

To the best of our knowledge, only LiKamWa *et al.* [7] have used existing hardware for reducing energy consumption in imaging systems. They devised a power model and found that optimizing camera clock frequency can reduce sensing power consumption by up to 50%. The power modeling portion of their work is the foundation for our sensing energy analysis.

Kulkarni *et al.* [8] demonstrated the energy implications of using scene captures of varying resolution for multicamera surveillance networks. This reduces energy consumption by 85% with respect to CMUcam: a widely used, high-resolution surveillance camera. Their approach, named SensEye, uses a network of cameras with different resolutions. In contrast, Digital foveation uses a single camera with images that are subsampled using existing readout mechanisms to tradeoff energy consumption, coverage, and resolution.

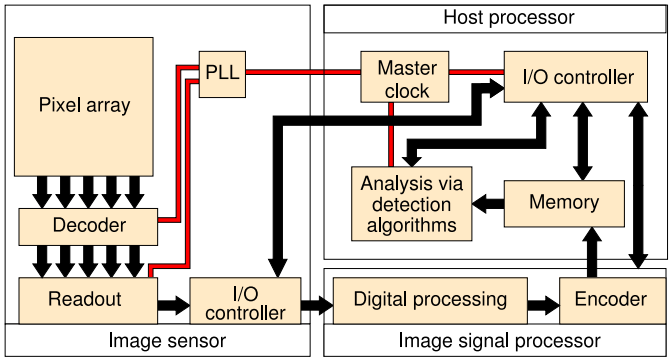


Fig. 2. Conventional image analysis pipeline. The sensor and ADC convert incident light to digital data. The ISP then denoises and demosaics the data and the host/applications processor performs image analysis.

Wang *et al.* [9] showed the implications of using software-only, multiround analysis to improve processing time and energy in computer vision applications. They forego use of existing hardware resources in the vision pipeline. As a result, their software-only approach reduces energy consumption by only 16.1%, when compared with the conventional, high- and uniform-resolution approach (see Section VI). Digital foveation reduces energy consumption by 76.3% for the chosen application of license plate recognition by avoiding transmission of superfluous data to the applications processor.

III. Conventional Image Analysis Framework

This section explains the framework used in conventional machine vision imaging systems to establish a basis for comparison with Digital foveation. It also enumerates the elements involved in the conventional imaging pipeline, shown in Fig. 2. Our focus is on electronic components; electro-mechanical components, which are used for focusing light on the image sensor plane, will not be discussed. Although one might reap even greater benefits from digital foveation by modifying the use of focusing machines, this paper demonstrates that large reductions in energy consumption are possible even without such changes.

A. Analog Signal Capture (Image Sensor)

The imaging pipeline starts at the image sensor: a 2-D array of pixels for sensing incoming light. A shutter controls exposure duration, which can be adjusted by the developer to improve the signal-to-noise ratio.

Access circuitry is used to acquire pixel values and perform analog black level calibration. Digital data are ultimately transferred to the host processor (see Fig. 2). The analog signal chain and readout circuitry are the most power-intensive components in the sensing stage, consuming 70%–80% of the power [10]. Energy consumed in the readout chain is related to the readout rates of sensors. Readout chain energy consumption is proportional to time.

B. Internal Communication Among Units

The mobile industry processor interface (MIPI) is used by most camera manufacturers for internal communication within components in the imaging pipeline. It is energy efficient [11] due to a low power consumption of 40.7 mW and transfer rate of 4 Gb/s.

C. Digital Processing (Image Signal Processor)

The sensor communicates with an ISP for digital processing. The image, which has a Bayer-pattern morphology, is demosaiced, producing a "RAW-RGB," "RAW-YUV," or other image format. It is then encoded into a standard, compressed format, e.g., JPEG, via an encoder pipeline in the ISP.

D. Machine Vision Focused Processing (Host/Applications Processor)

After digital processing, the compressed image is stored in local or remote memory of the programmable host (application) processor that performs machine vision tasks on the captured frame. It uses an I/O controller to sense interrupts, configure registers, and control the pipeline during frame capture.

Digital signal processing at the ISP and image analysis at the host processor account for 90%–95% of the total energy. Therefore, reducing data per analysis task can dramatically reduce energy consumption; this observation is of critical importance.

IV. Digital Foveation: Energy-Aware Framework for Imaging Systems

Digital Foveation, a framework inspired by the multiround, spatially varying-resolution imaging approach used in human vision and many other biological systems, adapts resolutions and sensed areas under control of image analysis algorithms. For example, it might use low-resolution images for detecting objects of interest and high-resolution images to examine those objects in detail.

Digital foveation is illustrated in Fig. 3. Each round in the process consists of image sensing under control of an application-specific analysis algorithm. The resolution and bounding box(s) of the image(s) are specified by the algorithm, generally producing much less data than a high-resolution, full-scene image. The analysis algorithm then determines whether enough information is available to complete the assigned task with adequate accuracy. If not, it guides the next round based on the information gathered in previous rounds.

We now describe the pipeline of digital foveation when used in a two-round image analysis process. The pipeline, shown in Fig. 4, begins with sensing and uses existing subsampling mechanisms to produce low-resolution images, which are used to determine locations of regions of interest. We refer to the vectors bounding the areas of interest as *foveal coordinates*; while the bounding box is analogous to the fovea. The foveal coordinates are provided as feedback to the sensor, which outputs higher-resolution captures of those regions. Unlike conventional foveal imaging, our approach permits fully electronic changes to the position, size, and resolution of the digital fovea, without using mechanical components, such as gimbals. There is no reliance on custom (and scarce) image processing algorithms designed for foveated sensors.

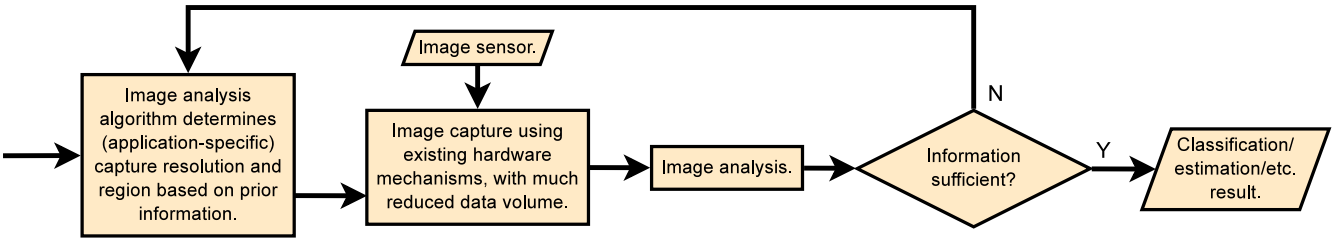


Fig. 3. Digital foveation: a multiround energy-efficient machine vision framework.

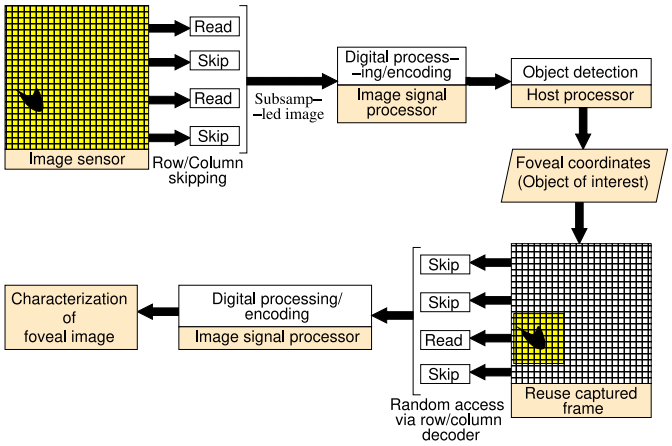


Fig. 4. Region of interest is located using low-resolution subsampled captures and analyzed using high-resolution captures. Existing camera hardware mechanisms are used to optimize energy consumption under an accuracy constraint.

Using a multiround detection approach that avoids wasteful pixel sensing, data transfer, and analysis, digital foveation enables large reductions in data transfer and processing times (70%–75%, typically) and energy consumptions (75%–85%, typically). Prior work has shown that multitiered approaches can improve characterization accuracy but reduce detection accuracy [9], while greatly reducing energy consumption.

Digital foveation uses several existing sensing mechanisms to tightly integrate the multiround algorithmic framework with hardware components.

A. Foveal Coordinates and Image Subsampling

An image sensor of a given resolution can produce lower resolution images by subsampling, i.e., row/column skipping and pixel binning. Pixel binning uses additional capacitive circuitry to average pixel values within a rectangular region, producing a single red/green/blue tuple. It reduces aliasing and improves signal-to-noise ratio by using additional averaging circuitry, at the cost of some increase in sensor power consumption [12].

Row/column skipping ignores selected rows/columns and requires no additional hardware. Modern CMOS image sensors further disable parts of their readout circuitry (such as, row/column decoders and ADCs) to reduce energy consumption at the subsampling stage [12]. Most of our discussion will assume pixel skipping. However, sensing energy consumption is small compared with that of the ISP and host processor, making pixel binning an option.

B. Foveal Capture via Random Access to the Pixel Array

Modern CMOS image sensors allow random access to pixel arrays by using parallel readout and row/column decoders [13]. Sensors, such as ON Semiconductor's NOII4SM6600A-D [14], set readout registers that output specific rectangular windows using this feature. The row/column skipping mechanism for image subsampling is implemented using this image sensor feature, too. We use the rectangular window capture mechanism to extract a high-resolution image of the objects of interest.

C. Reusing Captured Frames

Image sensors store captured frames in a pixel array, discarding them only when the pipeline is reinitiated via another capture event [13]. This permits rapid wakeup and readout of the same image signal; significantly benefitting our multiround analysis framework. The sensed data can be read again instead of capturing another frame.

V. Energy Characterization

This section explains the power consumption characteristics of imaging pipeline components and indicates their dependencies on throughput and time. SanMiguel and Cavallaro [15] described a power modeling approach for smart camera networks that accounts for current activity states of components and the corresponding activation durations. Zhang *et al.* [16] described a parametric power consumption model for mobile embedded systems that accounts for hardware component activity and power management states. They found that component power consumptions can generally be treated as independent, provided that systemic effects that change activities and power management states are accounted for. Based on this prior work, we developed a parallel, parametric model for the energy consumed by an imaging system per frame capture. It estimates energy consumption based on the activity state dependant power consumptions and durations. In Section VI, we use the model to determine the implications of our proposed framework on system energy consumption and latency.

A. Image Sensing

During exposure (T_{exp}), the image sensor is idle, i.e., it is not processing captured data via the analog signal chain or outputting it using the readout circuitry. It becomes active only when readout begins. After outputting the image data, sensors typically enter standby state, thereby reducing energy

consumption. The energy consumption per captured frame is the sum of the state-dependent component power–time products

$$E_{\text{sensor}} = \left(P_{\text{ele,idle}} + P_{\text{ana,idle}}\right)T_{\text{exp}} + \left(P_{\text{ele,active}} + P_{\text{ana,active}}\right)T_{\text{active}} \quad (1)$$

where $P_{\text{ele,idle}}$ and $P_{\text{ele,active}}$ are average power consumptions for an image sensor's power-intensive elements, excluding the analog signal chain, in idle and active states. Similarly, $P_{\text{ana,idle}}$ and $P_{\text{ana,active}}$ correspond to the analog signal chain's power consumption.

Digital logic element, PLL, and I/O controller power consumptions are roughly linear in clock frequency [17]. The analog signal chain's idle power consumption is linear in clock frequency (f), but also depends on image size (R, in pixels) [7]. This power reduction results from disabling row/column parallel ADCs and decoders when reading subsampled images [12]. Thus, the analog signal chain's active power consumption follows:

$$P_{\text{ana,active}} = a_1(R) + a_2 \quad (2)$$

where a_1 and a_2 are constants with units mW/megapixel and mW, which depend on physical characteristics and external clock frequency, f.

Typically, the sensor processes and outputs one pixel per clock period. Active duration is therefore a function of image resolution, i.e.,

$$T_{\text{active}} \approx R/f. \quad (3)$$

Using the random access capabilities of image sensors, a subimage of resolution R_2 can be produced from a sensor of resolution R_1 ($R_2 \leq R_1$). The energy required per frame capture is shown in (1) and (2)

$$E_{\text{frame}} = a\frac{R_1 \cdot R_2}{f} + b\frac{R_2}{f} + c \cdot T_{\text{exp}} \quad (4)$$

where, $a = a_1$, $b = (P_{\text{ele,active}} + a_2)$, and $c = (P_{\text{ele,idle}} + P_{\text{ana,idle}})$.

In the conventional framework, sensors transfer the entire pixel array. However, in digital foveation, only the region of interest is transferred. Typical machine vision applications concentrate on a region of interest varying from 1%–10% of the image size. Thus, our feedback approach reduces throughput across the vision pipeline by processing only relevant sections of the scene.

B. Digital Processing at the Image Signal Processor

The ISP is idle during sensing and active when processing the image. The last step of processing is encoding, after which the result is written to memory. The host processor then initiates the applications pipeline and the ISP becomes idle. As a result, the energy consumed at this stage is a function of time required for processing, T_{ISP}, which is linear in image size (see Section VI).

Assuming T_{ISP} is the time required for processing and T_{app} is the time for host processor image analysis

$$E_{\text{process}} = P_{\text{ISP,idle}}\left(T_{\text{exp}} + R/f + T_{\text{app}}\right) + P_{\text{ISP,active}}(T_{\text{ISP}}). \quad (5)$$

C. Host Processor

The applications pipeline involves the host processor, which remains idle during image sensing and digital processing, but activates when processing the image. The I/O controller and other peripherals used by the host processor are required for configuring and controlling the camera before, during, and after image transfer. These remain active during both image acquisition and analysis. The time required for the application pipeline (T_{app}) is a function of the image size, leading to the following host processor power consumptions:

$$P_{\text{host,idle}} = P_{\text{comp}} + P_{\text{app,idle}} \quad (6)$$

and

$$P_{\text{host,active}} = P_{\text{comp}} + P_{\text{app,active}}. \quad (7)$$

The energy consumption of these components, E_{host}, follows:

$$E_{\text{host}} = P_{\text{host,idle}}\left(T_{\text{exp}} + R/f + T_{\text{ISP}}\right) + P_{\text{host,active}}\left(T_{\text{app}}\right). \quad (8)$$

D. Communication-Dependent Latency

Intercomponent communication latency depends on the total amount of data transferred, which includes the 20%–50% overhead resulting from transmitting configuration and control parameters. An image of resolution R at p bits per pixel has the following latency:

$$\text{latency} = (1+h) \cdot \left(\frac{p \cdot R}{\text{BR}} + \frac{24R}{\text{BR}}\right) \quad (9)$$

where h is the overhead proportion, BR is the bit rate, and $(24R/\text{BR})$ indicates communication of a 24-bit RGB image. The MIPI interface power consumption (called P_{comm} in this paper) can be multiplied with the communication latency to calculate the communication interface energy consumption.

Although digital foveation requires more pipeline steps than conventional frameworks, the dramatic reduction in total transferred and analyzed data reduces communication latency and time (see Section VI).

E. Net Energy Consumed

Table I enumerates the energy consumed by image analysis components and provides energy model parameters for digital foveation and the conventional imaging framework. R is the image sensor resolution. The two frameworks are assumed to use the same ISP and host processor. Digital Foveation uses a subsampled image of resolution R_d and a high-resolution, foveated capture of size R_{fovea}. T_{ISP} differs between the frameworks (see Fig. 5) due to the multiround nature of digital foveation.

TABLE I
ENERGY CONSUMED BY CONVENTIONAL AND DIGITAL FOVEATION FRAMEWORKS

Vision framework	Energy model
Conventional	$a\frac{R^2}{f} + b\frac{R}{f} + cT_{exp} + P_{ISP,idle}\left(T_{exp} + \frac{R}{f} + T_{app}\right) + P_{ISP,active}\left(T_{ISP}\right) + P_{host,idle}\left(T_{exp} + \frac{R}{f} + T_{ISP}\right) + P_{host,active}\left(T_{app}\right) + P_{comm} \cdot (1+h)\left(\frac{p.R}{BR} + \frac{24R}{BR}\right)$
Digital foveation	$a\left(\frac{R_d^2}{f} + \frac{R \cdot R_{fovea}}{f}\right) + b\left(\frac{R_d}{f} + \frac{R_{fovea}}{f}\right) + cT_{exp} + P_{ISP,idle}\left(T_{exp} + \frac{R_d}{f} + \frac{R_{fovea}}{f} + T_{app}\right) + P_{ISP,active}\left(T_{ISP}\right) + P_{host,idle}\left(T_{exp} + \frac{R_d}{f} + \frac{R_{fovea}}{f} + T_{ISP}\right) + P_{host,active}\left(T_{app}\right) + P_{comm} \cdot (1+h)\left(\frac{p(R_d+R_{fovea})}{BR} + \frac{24(R_d+R_{fovea})}{BR}\right)$

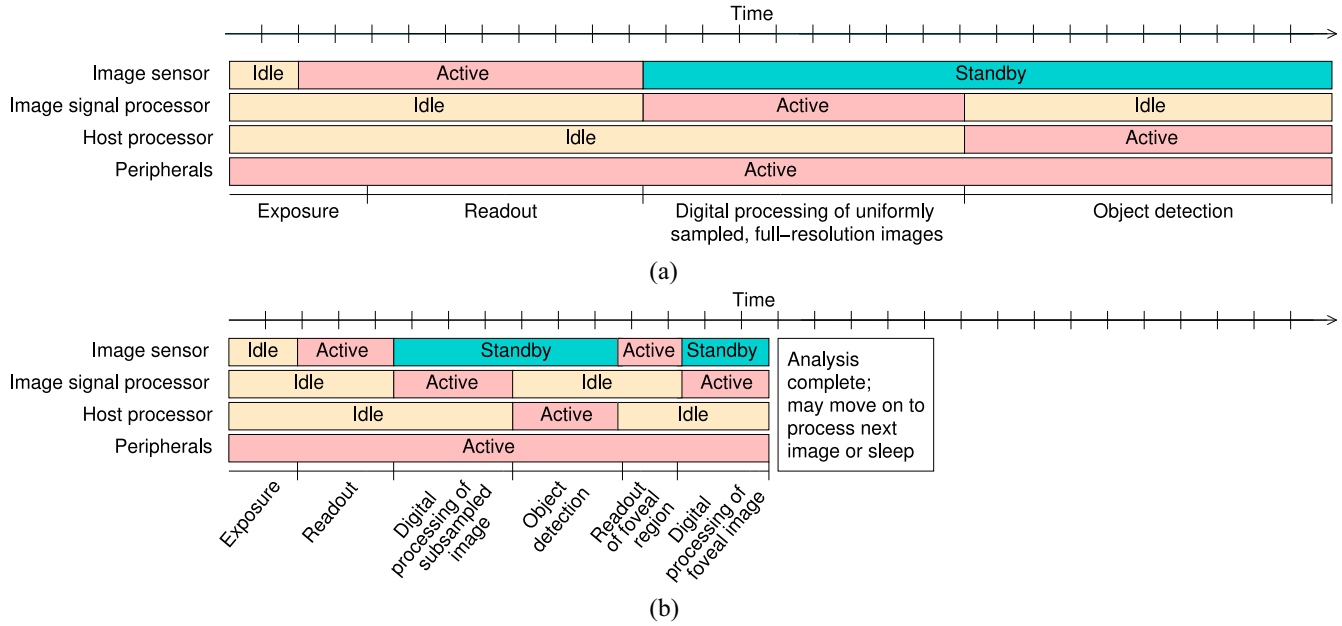


Fig. 5. Timing diagram for the (a) imaging pipelines for conventional and (b) digital foveation vision frameworks. Both show the times spent by components in different activity / power management states. Digital Foveation requires less time and energy.

VI. DIGITAL FOVEATION EVALUATION

We evaluated digital foveation on an imaging system composed of a Sony IMX219 image sensor and a Raspberry Pi 3, which is commonly used in low-budget commercial machine vision applications. Our evaluation focuses on the energy consumption, accuracy, and latency of digital foveation relative to the conventional machine vision framework. The license plate recognition application is considered, but the concepts in this paper are general across a wide range of machine vision problems, as shown in Section VII. A morphological image processing algorithm is used [18] for license plate segmentation.

Digital foveation can be used in systems containing GPUs and field-programmable gate arrays, as well as CPUs, and our preliminary analysis suggests that similar relative energy savings are possible. However, describing these experiments and analysis in detail is beyond the scope of this paper.

We use the power models described in Section V to determine the energy consumption implications of differing design decisions. We characterize our test imaging system's components to find the required coefficients and durations for images of varying resolutions.

A. Power Consumed by Image Sensor

The Sony IMX219 has a maximum resolution of 3280×2464 pixels, i.e., 8.08 M-pixels. Its datasheet [19] reports power consumptions for 3280×2464 and 3280×1844 resolutions at a 12 MHz clock frequency. For constant clock frequencies and activity states, analog component power consumption is roughly constant; however, the analog signal chain power consumption depends linearly on output resolution. We can therefore use the two resolution–power points and (2) to determine the following relationship between image size and power:

$$P_{\text{ana,active}} = 8.27\,\text{mw/M} - \text{pixels} \cdot R + 17.364\,\text{mW}. \quad (10)$$

We indicate the characteristics of Sony IMX219 in Table II.

TABLE II
12 MHz SONY IMX219 POWER CONSUMPTIONS

Component/State	Power (mW)
$P_{ana,idle}$	58.3
$P_{ana,active}$	8.27R + 17.364
$P_{ele,idle}$	83.5
$P_{ele,active}$	113.03

TABLE III
HOST PROCESSOR AND GPU POWER

State	Power (mW)
$P_{host,idle} + P_{GPU,idle} + P_{comp}$	0.95
$P_{host,active} + P_{GPU,idle}$	1.70
$P_{host,idle} + P_{GPU,active}$	1.35

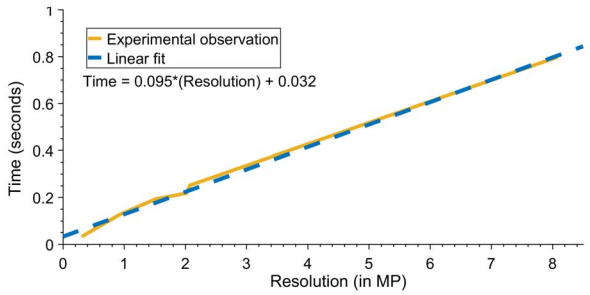


Fig. 6. Time required by the Raspberry Pi ISP pipeline is a nearly linear function of resolution.

B. Power Consumed by Image Signal Processor and Host Processor

The Raspberry Pi 3 has a dedicated image signal processing pipeline embedded in its GPU [20]. Thus, P_{ISP} is approximated by the GPU's power consumption. The license plate segmentation algorithm is implemented on an ARM Cortex A53.

We used an ammeter to measure the values P_{CPU} and P_{GPU}. Specifically, we characterize $P_{\text{CPU,idle}} + P_{\text{GPU,idle}}$, and then run CPU-intensive tasks to determine $P_{\text{CPU,active}} + P_{\text{GPU,idle}}$ and GPU-intensive tasks (using OpenGL ES) to determine $P_{\text{CPU,idle}} + P_{\text{GPU,active}}$. Our measurements are shown in Table III.

C. Image Signal Processing Pipeline and Detection Models

We now describe our methods for determining image signal processing time and energy consumptions. We calculate the effects of resolution on time by using the PiCam and MMAL encoder libraries, which direct the ISP to process and encode the image to JPEG and other compressed formats. We use the MMAL encoder to resize the original image and measure the effect of throughput reduction on the time required by the image signal processing pipeline. Our measurements indicate a linear relation between the two (see Fig. 6).

The host processor determines license plate foveal coordinates, which are used for high-resolution capture. The procedure is illustrated in Fig. 7. We report the times required for processing unsubsampled, 2×2 subsampled, and 4×4 subsampled images in Table IV.

Fig. 7. Demonstration of digital foveation in license plate recognition. A high-resolution image is subsampled and used to identify the region of interest, which is used in high-resolution capture and analysis.

TABLE IV
TIME CONSUMED BY HOST PROCESSOR

Resolution	Sub-sampling	Time (ms)
3,280 × 2,464	None	322.9
1,640 × 1,232	2×2	93.5
820 × 616	4×4	31.7

D. Net Energy Reduction

Using a 4×4 subsampled image for detection and a foveal, high-resolution capture of size 328×246 ($\approx 1/10$ the maximum image resolution) for license plate recognition, we observe a net 76.5% reduction in energy, compared with the conventional framework. Image sensor energy is reduced by 76.7% and ISP plus host processor energy is reduced by 76.3%. Table V contains these results. The digital foveation communication latency [see (9)] is 6.7% of that for the conventional framework. This results in 93.3% reduction in communication-related energy.

Similar savings occur for other image sensors. For example, replacing the Sony IMX219 image sensor power model with that of the OmniVision's OV5620 security camera sensor [7], results in digital foveation reducing energy consumption by 80.1%. These results suggest that digital foveation is applicable in multiple imaging systems and applications.

E. Comparison With Software-Only Approach

The idea of exploiting subsampled images for reducing processing time has been considered in other research (see Section II). Decreased processing time reduces energy consumption even when carried out on the host processor, but other energy-hungry portions of the pipeline, such as digital preprocessing on the ISP remain unchanged. Although the direct contribution of the image sensor to energy consumption is minimal, adaptively controlling its sampling regions and resolutions enables a dramatic reduction in system-wide energy consumption, and this benefit cannot be achieved by downsampling on the host processor. Throughput optimization in digital foveation reduces image signal processing time from 796.0 to 76.9 ms, saving energy. 4×4 image subsampling in

TABLE V
COMPONENT ENERGY CONSUMPTIONS FOR CONVENTIONAL AND DIGITAL FOVEATION FRAMEWORKS

Component	Conventional (J)	Digital Foveation (J)	Reduction
Image sensor	0.142	0.033	76.7%
Image signal processor and host processor	2.257	0.533	76.3%
Communication interface	0.015	0.001	93.3%
Total	2.414	0.567	76.5%

digital foveation results in 76.3% energy consumption reduction. However, a software-only approach would have reduced energy consumption by only 16.1%.

VII. IMPACT ON ACCURACY: CHOOSING THE RIGHT SUBSAMPLING LEVEL

As shown in Section VI, digital foveation has the potential to dramatically reduce energy consumption in machine vision applications. Since state-of-the-art detection algorithms base themselves on scale-invariant features; they are inherently robust to image subsampling. However, aggressive subsampling may reduce accuracy if the sampled object is too small. We are thus faced with the problem of minimizing energy consumption under a constraint on accuracy. This general problem can be broken into two classes, depending on application scenario.

If positions and orientations are constrained such that objects of interest will occupy a similar number of pixels in the captured images, a single statically set subsampling level is adequate to optimize energy-accuracy properties.

In applications, where the object of interest will occupy dramatically different pixel counts in different images, multiple (dynamically selected) subsampling resolutions are required to minimize energy consumption under detection accuracy bounds.

Thus, we evaluate our proposed methodology on two applications–one with similarly sized objects of interest (license plate recognition) and one with variably sized objects of interest (face detection). We based our analysis on an imaging system using Aptina image sensors characterized in prior work [7] and a Raspberry Pi 3. Energy per frame is calculated using the power models described in Section V.

A. Similarly Sized Objects of Interest

We use 2×2 and 4×4 subsampled images of license plates from a public dataset [21] and process them using a morphological image processing algorithm [18]. An open-source license plate recognition platform [22] is used to characterize the plates. The number of correctly estimated characters is used as the metric for net detection-plus-characterization accuracy.

Fig. 8 shows that increased subsampling reduces energy consumption. For 4×4 subsampled images, detection accuracy decreases by 9.2%; however, and characterization accuracy increases by 9.7%. Similarly, for 2×2 subsampled images, detection accuracy decreases by 0.95%, and characterization accuracy increases by 5.6%. This somewhat counterintuitive

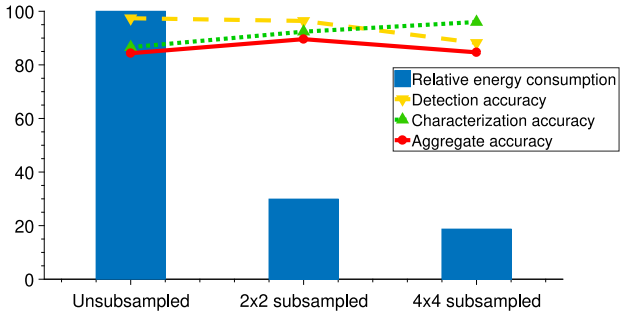


Fig. 8. Energy consumption and accuracy as functions of subsampling resolution. Bars indicate normalized energy consumptions. Lines indicate impact on detection accuracy (dashed yellow), characterization accuracy (dotted green), and aggregate accuracy (solid red) for license plate recognition.

increase was also observed by other researchers [9]. We attribute it to the removal of feature noise from the scene background, which affects the detected window in fine-grained, high-resolution images. Subsampled images focus on global features that, depending on size, can be best detected at a particular resolution. Thus, given accuracy and energy constraints, a developer can determine an ideal application-dependent subsampling level for similarly sized objects from the accuracy-subsampling curves in Fig. 8. For example, in the above analysis, 2×2 subsampling reduces energy consumption by 70.1% and increases aggregate accuracy by 4.3%. Thus, an accuracy- and energy-sensitive application can benefit from 2×2 subsampling. A 4×4 subsampling level reduces energy consumption by 81.3% and reduces aggregate accuracy by 0.65%. For applications in which subsampling generally decreases aggregate accuracy, the designer faces an accuracy-constrained energy minimization problem that the digital foveation framework makes explicit and solvable.

B. Variably Sized Object of Interest

To analyze the effects of digital foveation on detection accuracy for variably sized objects of interest, we use FDDB's benchmark dataset for face detection [23] and segregate data into three classes-based upon a parameter, s, that defines the ratio of the number of pixels occupied by the object of interest to the image resolution—a metric directly proportional to the object's size. The three classes follow: 1) $s = 0 - 0.09$; 2) $s = 0.09 - 0.17$; and 3) $s = 0.17 - 0.35$. The Viola–Jones face detection algorithm [24] is used for analyzing images. These images were subsampled until a significant drop ($>10\%$) was observed in detection accuracy, which occurred at a subsampling level of 8×8 (see Fig. 9).

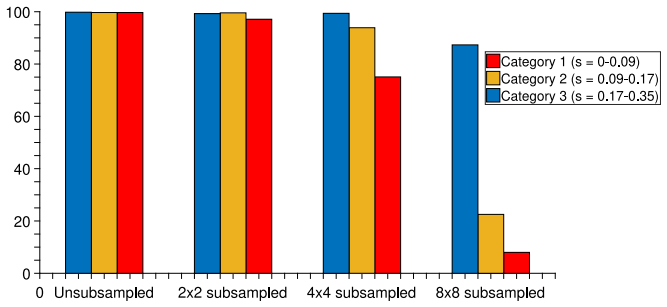


Fig. 9. Detection accuracy for the three categories of images as functions of subsampling level. One thousand six hundred images from each category were analyzed.

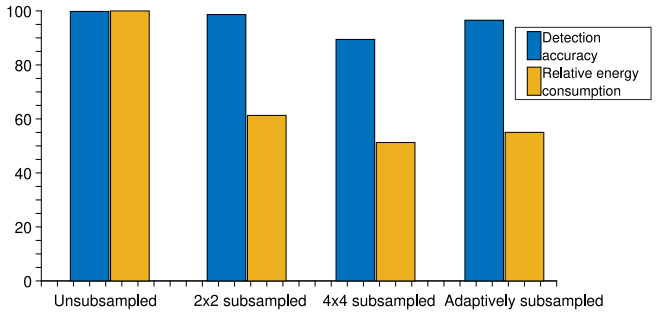


Fig. 10. Energy consumption and accuracy as a functions of subsampling resolution. For adaptive-subsampling, aggregate accuracy has been calculated using (11).

It is easy to claim that 4×4 subsampling is ideal for images in the third category ($s = 0.17-0.35$), for the detection accuracy drops by only 0.4% at that level; however, if the same level is used for images in the first category ($s = 0-0.08$), detection accuracy reduces by 25%. This implies that an adaptive approach is necessary for choosing a subsampling level that minimized energy consumption under accuracy constraints. To this end, we use highly subsampled images to determine the sizes of a objects of interest. Based upon object size, one can determine the subsampling level that minimizes energy under an accuracy bound as shown in Algorithm 1. We trained a convolutional neural network classifier, C, on images from categories $X_1, X_2, \ldots, X_n$, where $n = 3$ with 8×8 subsampling. The curves in Fig. 9 are used to determine the ideal subsampling level, L_i, for a given category, X_i. We used a 6% accuracy degradation constraint. The aggregate accuracy of the adaptive-subsampling pipeline follows:

$$\langle \text{accuracy} \rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} P\big(X_j \mid C = X_i\big)P_d\big(X_j L_i\big). \qquad (11)$$

Compared with the conventional approach, the net detection accuracy drops by 2.5% and energy consumption drops by 46%. Compared to digital foveation with 4×4 subsampling, which reduces energy consumption by 49.8%, object size adaptive subsampling achieves higher accuracy (96.6%, compared to 88.9%) with less reduction in energy consumption (46%, compared to 49.8%). Detailed results can be found in Fig. 10. We would like to note that the energy reduction is smaller than that for license plate detection because the images were smaller (400×300, on average). As a result, the exposure energy for the image sensor was comparable to the processing energy at the host processor and ISP. Energy savings will typically be higher in machine vision applications, where images are typically larger than 1.3 MP.

Algorithm 1 Calculate Ideal Subsampling Level(**input** G, D)

```
input training data, accuracy constraint, test sample;
function CATEGORIZE(training data)
    compute binning categorization thresholds based on object size
    compute detection accuracies for all categories
    return categorized data and detection accuracies
end function
L=2;

procedure LEVELCALC(categorized data)
    subsample all images to level L×L
    store detection accuracy at the given level
    if accuracy drop across all categories, w.r.t. unsubsampled images > threshold (e.g., 10%) then
        L = L + 1;
        LevelCalc(categorized data)
    else
        accuracy=100;
        while accuracy > accuracy constraint do
            train a classifier at level L subsampling
            compute accuracy using Equation 11
            L = L − 1;
        end while
    end if
end procedure

function DETECTOR(test sample, L)
    subsample test sample to level L×L
    compute test sample category using trained classifier
    pass the image through detection algorithm
end function
```

VIII. Multiframe Capture: Video-Based-Applications

Section VI shows that by reducing readout and processing time, digital foveation can enable large energy savings in machine vision applications that would conventionally use single-frame capture and analysis. Multiresolution methods for video-based applications have been introduced in the past, where one camera continuously captures low-resolution images until an object of interest is found, at which point a separate camera is activated to capture at higher resolution [8]. Digital foveation can be adapted to these applications by analyzing a low-resolution frame for object detection and thereafter estimating object velocities and future object locations using the incoming low-resolution buffer frames.

When a particular resolution is selected, the camera must generate buffers corresponding that resolution. The latency and

energy consumption for reconfiguration can be large or small, depending on the camera. For example, the MMAL encoder library used in this paper generates buffers every time an image capture is requested [25]. Since we base our analysis on the total time including configuration, capture, and processing, we account for configuration latency and energy consumption. However, it has been previously shown that some Android cameras can take significant time for pipeline reconfiguration [26]. Thus, pipeline reconfiguration latency may or may not constrain video frame rate, depending on camera.

IX. Conclusion

We described digital foveation, an adaptive framework to minimize energy consumption in machine vision applications. It is inspired by the use of variable-resolution sensing oculomotory feedback in the human vision system. Digital foveation exploits existing hardware mechanisms guided by image analysis management algorithms to control a multiround process that expends energy only on the most useful data from the image. In an example license plate recognition system, the approach reduces energy consumption by 81.3% with at most 0.65% reduction in aggregate accuracy.

References

[1] "Machine vision market research report—Global forecast 2022," Market Res. Future, Pune, India, Rep. MRFR/SEM/0981-CRR, May 2017.

[2] B. Cheung, E. Weiss, and B. A. Olshausen, "Emergence of foveal image sampling from learning to attend in visual scenes," in *Proc. Int. Conf. Learn. Represent.*, vol. abs/1611.09430, 2016.

[3] M. Yeasin and R. Sharma, *Foveated Vision Sensor and Image Processing—A Review*. Heidelberg, Germany: Springer, 2005, pp. 57–98.

[4] F. Jurie, "A new log-polar mapping for space variant imaging: Application to face detection and tracking," *Pattern Recognit.*, vol. 32, no. 5, pp. 865–875, 1999.

[5] S. Giulio and M. Giorgio, *Retina-Like Sensors: Motivations, Technology and Applications*. Vienna, Austria: Springer, 2003, pp. 251–262.

[6] R. LiKamWa, Y. Hou, J. Gao, M. Polansky, and L. Zhong, "RedEye: Analog ConvNet image sensor architecture for continuous mobile vision," in *Proc. IEEE Int. Symp. Comput. Archit.*, Seoul, South Korea, 2016, pp. 255–266.

[7] R. LiKamWa, B. Priyantha, M. Philipose, L. Zhong, and P. Bahl, "Energy characterization and optimization of image sensing toward continuous mobile vision," in *Proc. ACM Int. Conf. Mobile Syst. Appl. Services*, Taipei, Taiwan, 2013, pp. 69–82.

[8] P. Kulkarni, D. Ganesan, P. Shenoy, and Q. Lu, "SensEye: A multi-tier camera sensor network," in *Proc. Int. Conf. Multimedia*, Singapore, 2005, pp. 229–238.

[9] Z. Wang, Q. Hao, F. Zhang, Y. Hu, and J. Cao, "A variable resolution feedback improving the performances of object detection and recognition," *Inst. Mech. Eng. I J. Syst. Control Eng.*, vol. 232, no. 4, pp. 417–427, 2018.

[10] *1/2.5-Inch 5Mp CMOS Digital Image Sensor*, Data Sheet MT9P031, Aptina, San Jose, CA, USA, 2005.

[11] K. Lim, G. S. Kim, S. Kim, and K.-H. Baek, "A multi-lane MIPI CSI receiver for mobile camera applications," *IEEE Trans. Consum. Electron.*, vol. 56, no. 3, pp. 1185–1190, Aug. 2010.

[12] P. S. Gupta and G. S. Choi, "Image acquisition system using on sensor compressed sampling technique," *J. Elect. Imag.*, vol. 27, no. 1, pp. 013–019, 2018.

[13] J. Ohta, *Smart CMOS Image Sensors and Applications*. Boca Raton, FL, USA: CRC Press, 2007.

[14] *6.6 Megapixel CMOS Image Sensor*, Data Sheet NOII4SM6600A, ON Semicond., Phoenix, AZ, USA, Dec. 2016.

[15] J. C. SanMiguel and A. Cavallaro, "Energy consumption models for smart camera networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2661–2674, Dec. 2017.

[16] L. Zhang *et al.*, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones," in *Proc. ACM Int. Conf. Hardw. Softw. Codesign Syst. Synth.*, Scottsdale, AZ, USA, 2010, pp. 105–114.

[17] D. Duarte, N. Vijaykrishnan, and M. J. Irwin, "A complete phase-locked loop power consumption model," in *Proc. IEEE Design Autom. Test Europe Conf.*, Paris, France, Mar. 2002, pp. 1108–1109.

[18] F. Faradji, A. H. Rezaie, and M. Ziaratban, "A morphological-based license plate location," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1. San Antonio, TX, USA, Sep. 2007, pp. 57–60.

[19] *IMX219PQH5: Module Design Reference Manual, V2.2*, Sony, Tokyo, Japan, 2015.

[20] D. Jones. (2015). *Camera Hardware–Picamera 1.13 Documentation*. [Online]. Available: https://picamera.readthedocs.io/en/release-1.13/fov.html

[21] S. Riaric, "License plate detection, recognition, and automated storage," Univ. Zagreb, Zagreb, Croatia, Rep., 2003. [Online]. Available: http://www.zemris.fer.hr/projects/LicensePlates/

[22] "Openalpr documentation," OpenALPR Technol. Inc., Las Vegas, NV, USA, Rep. OpenALPR 2.5.103, 2015.

[23] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Dept. Comput. Sci., Univ. Massachusetts, Boston, MA, USA, Rep. UM-CS-2010-009, 2010.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

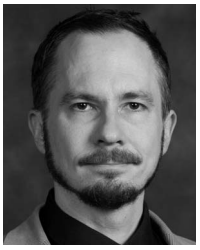
[25] D. Jones. (2016). *API-Mmalobj–Picamera 1.13 Documentation*. [Online]. Available: https://picamera.readthedocs.io/en/release-1.13/api_mmalobj.html

[26] J. Hu, J. Yang, V. Delhivala, and R. LiKamWa, "Characterizing the reconfiguration latency of image sensor resolution on android devices," in *Proc. ACM Int. Workshop Mobile Comput. Syst. Appl.*, Tempe, AZ, USA, 2018, pp. 81–86.

Ekdeep Singh Lubana is currently pursuing the undergraduation degree with the Department of Electronics and Communication Engineering, Indian Institute of Technology, Roorkee, Roorkee, India.

He was researching as a Visiting Scholar with the EECS Department, University of Michigan, Ann Arbor, MI, USA. He has previously researched in the field of machine learning, machine vision, and image sensors. His research on physiological stress sensing in plants, as a Visiting Scholar with the Indian Institute of Technology, Bombay, Mumbai, India, led to two patents and a publication. His current research interests include embedded systems, computer architecture, image sensors, energy efficient machine vision, and machine learning.

Mr. Lubana was a recipient of the Ericsson Innovation Awards and the Accenture Innovation Challenge, in 2017, in various international and national technical competitions.



Robert P. Dick (S'95–M'02) received the B.S. degree from Clarkson University, Potsdam, NY, USA, in 1996 and the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 2002.

He is an Associate Professor of electrical engineering and computer science with the University of Michigan, Ann Arbor, MI, USA. He researched as a Visiting Researcher with NEC Labs America, Princeton, NJ, USA, in 1999, and as a Visiting Professor with the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2002, and was on the Faculty of Northwestern University, Evanston, IL, USA from 2003 to 2008. He served as the CEO of the Stryd, Inc., Boulder, CO, USA, from 2015 to 2016, which produces wearable electronics for athletes.

Dr. Dick was a recipient of the NSF CAREER Award, the Department's Best Teacher of the Year Award in 2004, the Computerworld Horizon Award in 2007 for his technology, and the Best Paper Award at DATE in 2010 for his research. His paper was selected as one of the 30 in a special collection of DATE papers appearing during the past ten years. He served as the Technical Program Committee Co-Chair of the 2011 International Conference on Hardware/Software Codesign and System Synthesis, as an Associate Editor of the IEEE Transaction on Very Large Scale Integration (VLSI) Systems, and as a Guest Editor for *ACM Transaction on Embedded Computing Systems*.