https://doi.org/10.2352/ISSN.2470-1173.2018.17.AVM-105

Fundamental Imaging System Analysis for Autonomous Vehicles

Robin Jenkin and Paul Kane; ON Semiconductor; San Jose, California, USA

Abstract

This paper explores the use of existing methods found in image science literature to perform 'first-pass' specification and modeling of imaging systems intended for use in autonomous vehicles. The use of the Johnson Criteria [1] and suggestions for its adaptation to modern systems comprising neural nets or other machine vision techniques is discussed to enable initial selection of field of view, pixel size and sensor format.

More sophisticated Modulation Transfer Function (MTF) modeling is detailed to estimate the frequency response of the system, including lower bounds due to phase effects between the sampling grid and scene [2]. A signal model is then presented accounting for illumination spectra, geometry and light level, scene reflectance, lens geometry and transmission, and sensor quantum efficiency to yield electrons per lux second per pixel in the plane of the sensor. A basic noise model is outlined and an information theory based approach to camera ranking presented. Thoughts on progressing the above to look at color differences between objects are mentioned.

The results from the models are used in examples to demonstrate preliminary ranking of differently specified systems in various imaging conditions.

Introduction

The application of deep learning to self-driving vehicles has become a tractable challenge in recent years. As a result, the number of cameras on new vehicles is expected to increase significantly and questions around the optimum quality and configuration of these systems have started to arise.

Quite clearly the ultimate needs of Advanced Driver Assistance Systems (ADAS) are different from those intended for the human visual system (HVS). Information is required to be extracted from the images for an ADAS system in a timely manner that is intended to control a vehicle, whereas images for human consumption are generally optimized for aesthetic image quality, or to permit a human observer to extract information. This immediately leads to differences in system configurations, such as resolution and pixel size. Any resolution beyond the needs for the driving task burdens ADAS systems with unnecessary computation and the need for fast response times and low motion blur decreases exposure times, increases frame rates required and ultimately the sensitivity needs of the pixel. Cameras that produce images for human consumption generally cover a wide range of artistic intent and rendering, from portraiture to landscape, and phone screens to large prints. Pixel counts tend to be higher and time constraints on processing are not nearly as prescriptive.

These differences further lead to alternative color filter array choices to be made and practical differences in the in the tuning of image signal processing pipes (ISPs) designed for each. RCCC and RCCB versus the typical RGGB arrangements are common for ADAS systems as well as adjustments in demosaic, noise, sharpening and color processing blocks.

Until we fully understand what is required by neural networks to maximize performance we should treat them as alien observers. Quality requirements for human viewing may be a good starting point and perhaps where fundamental features are specified in early neural layers we can get some clues about what would maximize output from neurons by examining them closely. In networks where no low-level features are defined, however, we do not get these clues. In short, the industry still has a long way to go to arrive at neural network calibrated image quality metrics in a similar manner to psychovisually calibrated image quality metrics from the likes of Keelan et al. [3]. Compounding this at a system level, the way in which objects are memorized, tracked and their trajectory anticipated in the human visual system is highly advanced. Neural networks largely identify objects on a frame by frame basis after which they are fed into tracking and motion algorithms.

A number of the fundamentals that we take for granted in the field of pictorial imaging should be reexamined. For example, there is no standard "Macbeth" chart for automotive scenes. Road sign and marking colors, asphalt, concrete, and the mean color for a car would be useful to have access to and standardize for the comparison of measurements. Many of the assumptions around reflectivity of objects need to be updated as retroreflective materials are often used in signage and far from Lambertian. Standard spectra need be agreed upon for analysis involving the traffic signals, headlamps, tail lamps, nighttime rural and city skies etc. in a similar manner to those already derived for daylight, tungsten and fluorescent spectra by organizations such as the Commission Internationale de l'Eclairage (CIE).

In this instance imaging science is catching up with the rapid progress of practical application the field, but there is a wealth of existing knowledge, especially in medical and defense imaging, that may be drawn upon. We should be using this work more heavily to analyze and guide the design of automotive cameras and not reinvent the wheel. The most useful analysis will ultimately rate the ability of these systems to perform the tasks they are designed for and this intent should be kept at the forefront of analysis efforts.

Johnson Criteria

Johnson was perhaps the first person to objectively link task performance to imaging system parameters in a meaningful way in the 1950's [1]. Richardson et al. have an accessible description of the approach [4]. Through experimentation, Johnson determined the number of cycles, or line pairs, that were required for personnel to perform various tasks, such as detection, recognition or identification of a target. Despite the large variety in targets, ranging from people to tanks, he found that the number of cycles needed to perform each task correlated well with the smallest object dimension. His tasks were, detection, orientation, recognition and identification. Detection is merely confirming "something" is there, without knowledge of what it is. Orientation, which way "it" is pointed. Recognition allows the determination of the class of object, person, tank, plane etc. and identification the type, e.g. T72 Tank. The number of cycles derived by Johnson were 1±0.25, 1.4±0.35, 4±0.8 and 6.4±1.5 respectively. These are usually simplified to 1, 2,

4 and 8 cycles. Figure 1, shows the appearance of an object rendered using each of the criteria above.

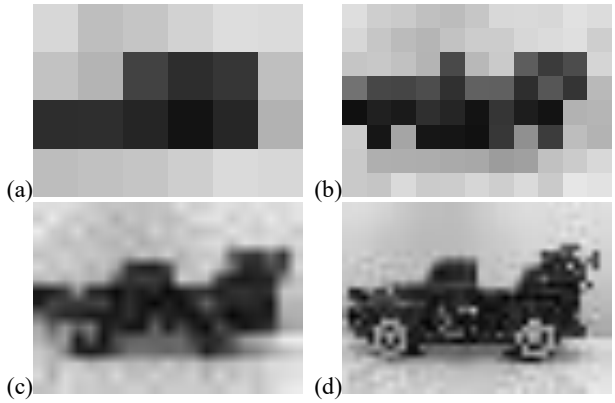

Figure 1. Rendering of an object using the Johnson criteria. (a) One cycle or detection, (b) two cycles or orientation, (c) four cycles or identification and (d) eight cycles or recognition.

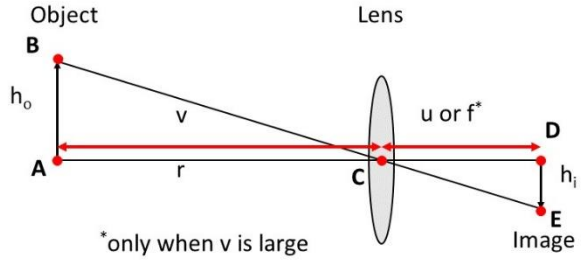

Figure 2. The geometry of the thin lens equation.

The power of using the Johnson criteria is that combining this with the thin lens equation [5], Eq.1, Figure 2, we can quickly estimate the relative performance of an imaging system to conduct a task. The thin lens equation states:

$$\frac{1}{f} = \frac{1}{u} + \frac{1}{v} \tag{1}$$

where f is focal length, v is object distance and u the distance from the lens to the image. When object distance, v, is large, $\frac{1}{v} \to 0$, and $\frac{1}{f} \approx \frac{1}{u}$, so $f \approx u$. We will now refer to v as range, r. From Figure 2, we can see similar triangles are formed between points ABC and CDE and therefore:

$$\frac{h_i}{f} = \frac{h_o}{r} \tag{2}$$

where h_i is image height, h_o, object height, f, focal length and r, range. We know our pixel size, p, and we know the number of cycles that we require to perform the task according to the Johnson criteria, n. Therefore, the image height needed to complete that task is simply:

$$h_i = 2np \tag{3}$$

Substituting Eq.(3) into Eq.(2) we find a pixel size needed to complete a given task at a given range for a specified object size, Eq.(3).

$$p = \frac{h_o f}{2nr} \tag{4}$$

It is worth reiterating here that it is the minimum object dimension that is used. i.e. for a 1.8×0.5m person we would use 0.5m. Equation 4 is not in a convenient form to use in practice. More commonly we start with sensor dimensions and have an idea of the field of view for a specified lens. Given the pixel size as before, the number of active pixels in the horizontal direction, A_{HOR}, and the horizontal field of view, FOV_H, the focal length of the lens is given by:

$$f = \frac{pA_{HOR}}{2\tan\left(\frac{FOV_H}{2}\right)} \tag{5}$$

Rearranging Eq.(4) we can find the number of cycles, hence the capability to perform a task, at a particular range for a given object size:

$$n = \frac{h_o f}{2pr} \tag{6}$$

The equation is intuitive, and a number of basic observations can be made. Object size and focal length are in the numerator, so as object size and focal length increase the number of cycles increases and we can perform more sophisticated tasks with our image at longer ranges. Double either the object size or focal length, we double the distance at which we can perform that task. Pixel size and range are in the denominator and as pixel size and range increase, the number of cycles decreases, and we can do less with our image. We may also see that, to a first order, pixel size and focal length will drive system size for a constant aperture. Figure 3 shows curves for the number of cycles on Euro NCAP, child, adult and bicyclists [6] for a sensor with 2000, 3um square pixels in the horizontal direction with a 60-degree field of view lens.

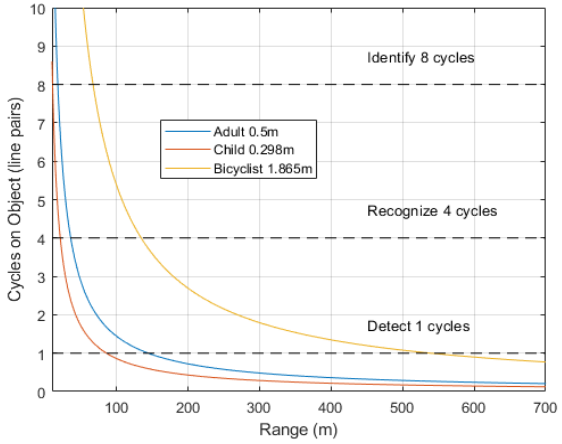

Figure 3. Johnson Criteria for Euro NCAP objects calculated using an imaginary camera system with a 60-degree HFOV with a horizontal pixel count of 2000 and pixel size of 3.0um. Euro NCAP Adult is 1.8x0.5m, Child 1.154x0.298m and Bicyclist 1.865x1.89m [7].

The Johnson criteria has several shortcomings and really only provides first order rank of camera systems. It was developed for human observers looking through image intensifiers. We should not expect it to work without modification for automotive imaging. We can readily update the number of cycles, however, needed to perform a particular task, such as reading the text on a STOP sign, or identifying a traffic light using a convolutional neural network (CNN) or machine vision algorithm. Johnson does not account for f-number, exposure, atmospheric conditions, or noise and we should expect the performance in low light or fog to be completely different. Even the addition of these factors may not be sufficient. For example, atmospheric scattering is generally exponential with distance. Additionally, the Johnson criteria is very poor at predicting the visibility of self-luminous objects at sub-pixel sizes, namely lights. Johnson is a resolution metric, rather than a signal metric and as such it is unable to account for shear power transmitted by point sources into the point spread function. What it does give us however is a very quick way to rank systems in terms of their likely performance in reasonable imaging conditions based on imaging geometry. Figures 4 and 5 demonstrate Johnson as applied to a commercially available camera with a f/2.4, 3.3mm, 56-degree horizontal field of view lens, and 3264×2448 1.12um pixel sensor. Figure 4 shows number of cycles versus distance for 20.3cm (8 in) lettering on a 1.70m wide sign and Figure 5 the corresponding image sequence at various distances for that sign. Despite the simplicity of the approach, it does a credible job of predicting a measure of task capability for the images for human viewing. Updating the number of cycles needed for network performance could be a relatively simple task. Data transport, computational burden and thermal dissipation are directly related to pixel count and having a tool to optimize these at an early stage aides system design enormously.

Modulation Transfer Function and Point Spread Function

Not all pixels are created equal and a more sophisticated approach to modeling resolution through a system is afforded by the transfer function, *M(ω),* which is the Fourier Transform of the point spread function. The modulus of *M(ω)* is the Modulation Transfer Function (MTF). Signal transfer of any number of components, such as lens, pixel, demosaic and crosstalk may be estimated, and the system response computed by cascading the transfer functions of the individual components, as shown in Eq.7.

$$M_{SYS}(\omega) = M_{LENS}(\omega) \times M_{PIX}(\omega) \times M_{DEM}(\omega) \times M_{CROSS}(\omega) \quad (7)$$

Performing calculations in this manner, the effect of aperture on system resolution may be included. For a diffraction limited lens, given wavelength λ and the f-number, $M_{LENS}(\omega)$ is calculated using [7]:

$$M_{LENS}(\omega) = \frac{2}{\pi}\left[\cos^{-1}\frac{\omega}{\omega_0} - \frac{\omega}{\omega_0}\sqrt{1-\left(\frac{\omega}{\omega_0}\right)^2}\right] \quad (8)$$

where

$$\omega_0 = \frac{1}{\lambda(f-number)} \quad (9)$$

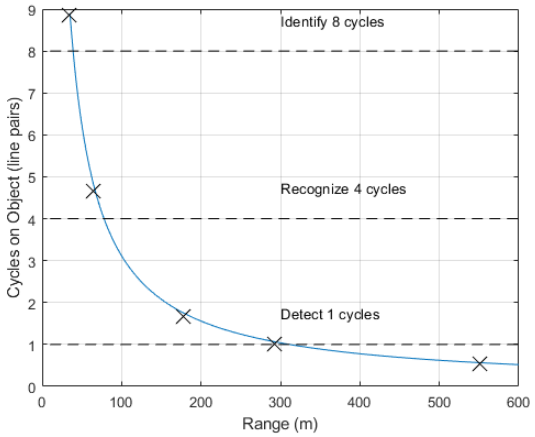

Figure 4. Johnson curve for a commercially available f2.4 3264x2448 1.12um pixel camera with 56-degree horizontal field of view imaging 20.3cm (8 in) text. Markers represent distances at which images were captured in Figure 5.

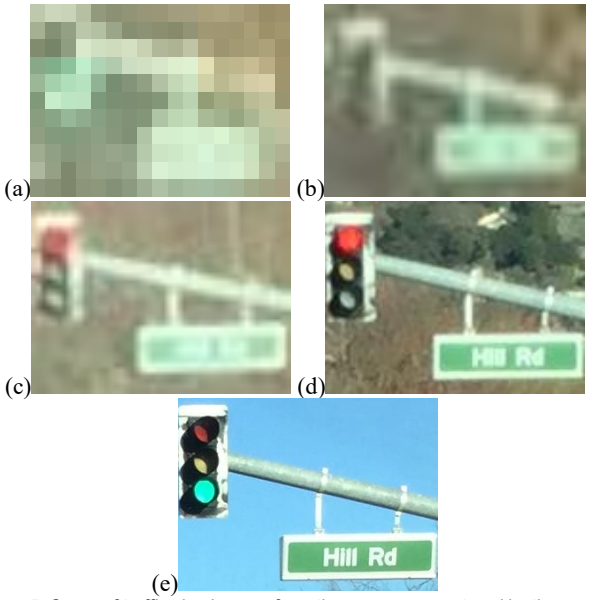

Figure 5. Crops of traffic sign images from the sequence captured by the system described in Figure 4. The Johnson curve in Figure 4 indicates that the writing should be 'undetectable' in (a) 0.54 cycles, 'detectable' in image (b) 1.02 cycles, it's orientation largely 'discernible' in (c) with 1.67 cycles, recognized as writing in (d) 4.66 cycles and identifiable (completely legible) in (e) with 8.85 cycles.

For more representative results, $M_{LENS}(\omega)$ can be computed for each wavelength under consideration and an average calculated, and weighted according to the system quantum efficiency at each of those wavelengths. Further, the f-number used in the calculation may be de-rated as suggested by Keelen [8] to simulate typical lens performance. Pixel response, $M_{PIX}(\omega)$, may be calculated using:

$$M_{PIX}(\omega) = sinc(\pi p \omega) \quad (10)$$

where p is the pixel size and ω, spatial frequency, as previously. It should be noted, however, that this result gives the optimum spatial frequency response for the pixel, when the signal is precisely in phase with the sampling grid. Jenkin [9] and others [10]

have shown that when the sampling grid is out of phase with the signal, the frequency response degrades. It may be shown that when considering all possible phase differences between the signal and the sampling grid, the minimum possible response is [9]:

$$M_{PIX_{MIN}}(\omega) = \left(\frac{\cos(\pi\omega s)\sin(\pi\omega p)}{\pi\omega p}\right) \quad (11)$$

where s is the sampling pitch. The average response is then given by [9]:

$$M_{PIX_{AVE}}(\omega) = \left(\frac{\cos^2\left(\frac{\pi\omega s}{2}\right)\sin(\pi\omega p)}{\pi\omega p}\right) \quad (12)$$

where all terms are defined as previously. It is argued that for safety critical systems, the average or phase de-rated transfer functions are more appropriate estimates of pixel response. The line spread functions (LSFs) calculated from the above result corresponding to the minimum response is shown to be equivalent to two neighboring pixels [2]. Figure 6 shows the maximum, minimum and average transfer functions calculated for a 3µm pixel.

It is possible to calculate LSFs from the $M_{SYS}(\omega)$ by taking the modulus of the inverse Fourier transform. To increase sample points in the resultant LSF and to avoid aliasing and false high frequencies being introduced, mirroring $M_{SYS}(\omega)$ around the DC (zero frequency) value and padding with an equal number of zero points to yield a curve that is four times the size of the original MTF is recommended. Figure 7 shows $M_{SYS}(\omega)$ modeled for an imaginary 3.75µm pixel f/2.4 system, for the sampling grid in and out of phase, Eqs 10 and 11. Crosstalk and demosaic have been omitted for simplification. Figure 8 shows the corresponding LSFs. Determining the width of the LSF using an appropriate criterion such as the full-width-half-max(FWHM) or 80% encircled energy yields a consistent manner to calculate an *effective* pixel size (cell size) to use in a Johnson criteria type calculation. This in turn allows the relative performance of systems with different pixel sizes and apertures to be predicted and compared.

Signal Modeling

Resolution modeling does not yield a complete picture of the performance of any imaging system as sensitivity to light will also determine whether signals are recorded at a threshold where they can be detected. There are many approaches recorded in the literature. Richardson details one of the most accessible that can readily be modified for automotive imaging [11,12].

Ambient light level and its color temperature are first specified in lux and kelvin, L_{AMB}, and CT_{AMB} respectively. Photometric units are used as they are familiar and easily understood. Scaling for the radiometric properties of the illumination falling outside of the sensitivity of the human visual system is included in the calculation as follows. For convenience a blackbody curve is generated at the specified color temperature, CT_{AMB}, to model the spectrum of the illumination. In practice this could be any illumination spectra specified in $Wm^{-2}\mu m^{-1}$ [4].

$$W(\lambda) = \frac{C1}{\lambda^5\left[e^{\frac{C2}{\lambda CT_{AMB}}}-1\right]} \quad (13)$$

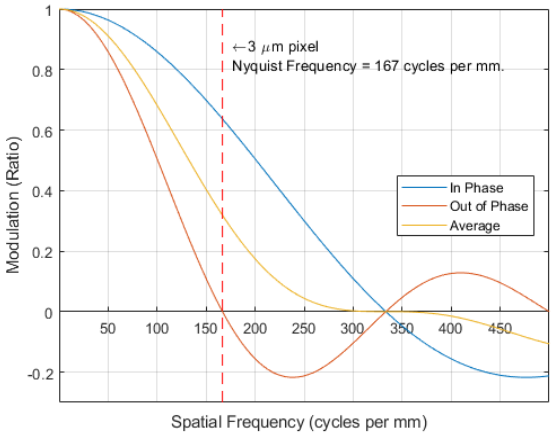

Figure 6. The maximum (in phase), minimum (out of phase) and average transfer functions for a 3.0um pixel [2]. Note that the transfer function below zero represents a phase reversal of the signal.

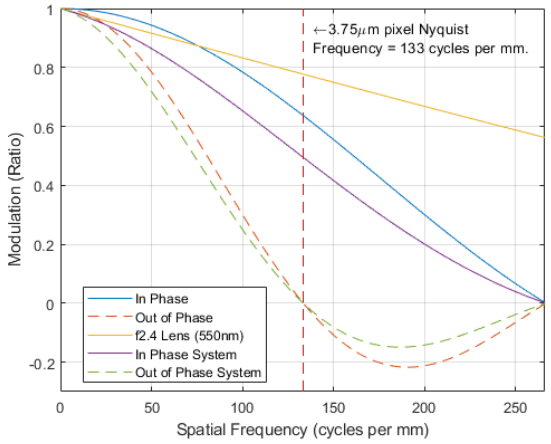

Figure 7. Maximum and minimum transfer functions calculated for an imaginary 3.75µm pixel, f2.4 system with the sampling grid in and out of phase with the signal. Also shown is the f2.4 lens transfer function calculated using a wavelength of 550nm.

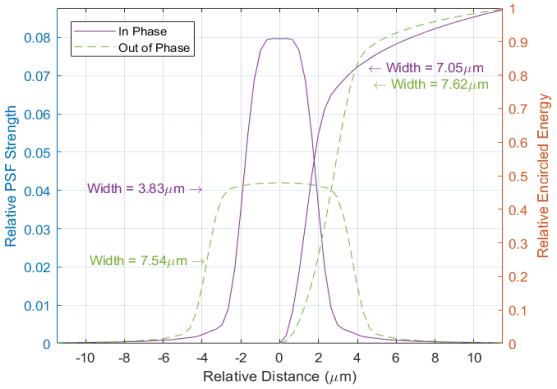

Figure 8, Line Spread Functions(LSFs) and corresponding encircled energies for the system transfer functions shown in Figure 7. Widths of LSFs are calculated at FWHM and 80% encircled energy.

where $C1$ is the first radiation constant, 3.74x10^8 Wm^{-2}μm^{-4}, and $C2$ the second radiation constant, 1.44x10^4 μmK^{-1} [12]. The relative spectral luminous efficiency curve, $V(\lambda)$, of the CIE is scaled by the peak luminous efficacy of human vision (683 lumens per watt at 555 nm) [13] and multiplied by the blackbody curve above then integrated to yield the total lux, L_{SOURCE}, represented by the illumination curve generated:

$$L_{SOURCE} = 683.\int_{\lambda_{MIN}}^{\lambda_{MAX}} W(\lambda).V(\lambda)d\lambda \quad (14)$$

where λ_{MAX} and λ_{MIN} are the maximum and minimum wavelengths of interest.

The object is considered a Lambertian reflector with reflection, R_{OBJ}, and thus the light scattered, L_{REF}, by the object in units of lux m^{-2}str^{-1} is [12]:

$$L_{REF} = \frac{R_{OBJ}.L_{AMB}}{\pi} \quad (15)$$

A factor, L_{SCALE}, by which to multiply $W(\lambda)$ may then be calculated, Eq.16, to yield the blackbody curve correctly scaled to the wattage required to yield the lux reflected from the object. We then multiply by the absolute quantum efficiency curve of the sensor, $Q(\lambda)$, and absolute transmission of an infrared filter, $I(\lambda)$, to yield the spectrum of light available to the sensor in Wnm^{-1}m^{-2}str^{-1} before lens and pixel geometry are considered, $P(\lambda)$.

$$L_{SCALE} = \frac{L_{REF}}{L_{SOURCE}} \quad (16)$$

and

$$P(\lambda) = L_{SCALE}.W(\lambda).I(\lambda).Q(\lambda) \quad (17)$$

The solid angle, Ω, of the lens collecting the signal reflected from the projected pixel area is [12]:

$$\Omega = \frac{\pi D_{OPTICS}^2}{4r^2} \quad (18)$$

where, D_{OPTICS}, is the effective diameter of the lens and r is range as previously. Multiplying by the solid angle and transmission of the lens, T_{OPTICS}, yields the power per nm per square meter, P_{SENSOR}, captured by the sensor:

$$P_{SENSOR}(\lambda) = L_{SCALE}.W(\lambda).I(\lambda).Q(\lambda).\Omega.T_{OPTICS} \quad (19)$$

Multiplying by the area of the pixel, A_{PIXEL}, yields the power per nm per pixel.

$$P_{PIXEL}(\lambda) = L_{SCALE}.W(\lambda).I(\lambda).Q(\lambda).\Omega.T_{OPTICS}.A_{PIXEL} \quad (20)$$

The energy per photon, $E(\lambda)$, is calculated using:

$$E(\lambda) = \frac{hc}{\lambda} \quad (21)$$

where h is Plank's constant, 6.62x10-34 m^2 kg s^{-1}, and c is the speed of light, 299792458 ms^{-1}. Dividing $P_{PIXEL}(\lambda)$ by $E(\lambda)$, multiplying by the integration time, T_{INT}, and integrating yields the total number of photoelectrons captured by the pixel, PH_{PIXEL}:

$$PH_{PIXEL} = \int_{\lambda_{MIN}}^{\lambda_{MAX}} \frac{T_{INT}.P_{PIXEL}(\lambda)}{E(\lambda)} d\lambda \quad (22)$$

Finally, if the total number of photons calculated as being detected by the pixel exceeds the linear full well for the pixel, the total number of photons is clipped at that value.

Headlamps

A rudimentary model of headlamps may be created by specifying a color temperature, C_{LAMP}, and luminous flux, LF_{LAMP}, that is emitted into an elliptical beam of horizontal angle, α_H, and vertical angle α_V, Figure 9. The horizontal, S_H, and vertical, S_V, semi-radii, at range, r, are then calculated:

$$S_H = r\tan\left(\frac{\alpha_H}{2}\right) \quad (23)$$

$$S_V = r\tan\left(\frac{\alpha_V}{2}\right) \quad (24)$$

The cross-sectional area of the headlamp beam, A_{LAMP}, is simply:

$$A_{LAMP} = \pi S_H S_V \quad (25)$$

and thus the number of lux per square meter falling on the object provided by the headlamp, L_{LAMP}, is the total luminous flux, LF_{LAMP}, divided by the area, A_{LAMP}, provided the beam size is larger than the object considered:

$$L_{LAMP} = \frac{LF_{LAMP}}{A_{LAMP}} \quad (26)$$

Calculations may then proceed from the point of generating and scaling the illumination spectra generated for the color temperature, Eq.13 onwards, to yield the total number of photons generated per unit exposure for the headlamp. The ambient and headlamp components may then be combined to yield the total photoelectrons in the pixel. The ambient and headlamp components should be calculated separately as they will likely have significant color temperature differences. Further, as models are developed to incorporate retroreflective material behavior, the response from each is likely to be different based on geometry. It should also be noted that we ignore the area where light from both headlamps overlap in this calculation, and that we have assumed a uniform beam with elliptical symmetry, the area of which scales in proportion to the distance. In reality, the beam is not symmetric (to

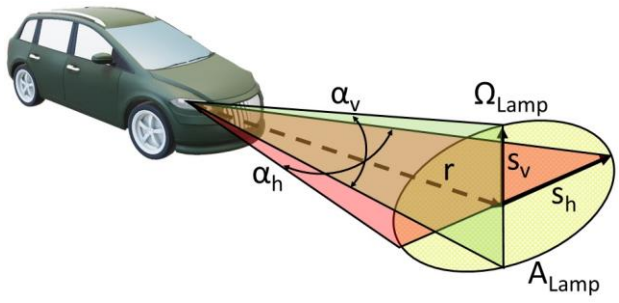

Figure 9. Headlamp model describing the spread of the luminous flux, LF_{LAMP} into area $A_{LAMP} = \pi S_v S_h$ at range, r. The vertical and horizontal angle of the beam is α_v and α_h respectively. The solid angle of the headlamp, $\Omega_{Lamp} = A_{Lamp} / r^2$.

avoid blinding the oncoming traffic) and the contribution of multiple headlamps must be considered. Figure 10, shows however, that within approximately 40 meters, the contribution of a typical 1500 lumen headlamp is a significant proportion of the total illumination against a 2 lux ambient light level.

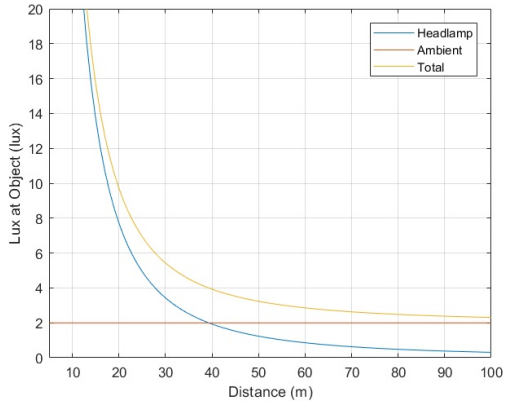

Figure 10. Example of model for 1500 lumen, 3200k headlamp casting light into a 60x30 degree elliptical beam. Also shown is a 2 lux ambient background light level to yield the total lux falling onto the object versus distance.

Noise Modeling

Noise degrades the detectability of signals, especially in low light conditions. There are many sources of noise in imaging systems and it would be tempting to generate a dozen or so terms to model this. Holst and Lomheim provide a good overview of these and we concentrate on the most significant terms here [14].

Generally, we may divide these into signal dependent and independent sources. Whilst signal dependent sources are troublesome, independent are more so as they remain at a fixed level as the signal diminishes and tend to determine the performance of the camera system at low light levels. The relative levels of dependent and independent noise can also yield a crossover point where in high light levels, System A may be preferred over B, whereas in low light conditions System B may well be the better choice.

Our first signal dependent noise source, photon shot noise, N_{SHOT}, is due to the quantum nature of light, a Poissonian process, and increases as the variance of the mean signal [15]. Therefore, an exposure of N quanta will yield shot noise of $\sqrt{N}$. Pixel-to-pixel variations in sensitivity, or pixel response non-uniformity (PRNU), N_{PRNU}, may simply be modelled as a percentage of the mean signal level. Read noise, N_{READ}, is signal independent and may be thought of as being generated by any process that reads the signal. This is usually expressed as a root mean squared fixed number of electrons. Dark current, a second signal independent source, is caused by thermal generation of carrier pairs in the bulk silicon. It is usually expressed as the number of carriers per ms at a nominal temperature, usually junction temperature.

Dark current adds its own shot noise to the signal, with a standard deviation of N_{DARK}. Independent noise sources add via quadrature and thus the total noise, N_{TOTAL}, in an exposure is given by [15]:

$$N_{TOTAL} = \sqrt{{N_{SHOT}}^2 + {N_{PRNU}}^2 + {N_{READ}}^2 + {N_{DARK}}^2} \quad (22)$$

Read noise and dark current generally vary according to pixel size. Elementary models of this variation will affect results at low light levels when attempting to optimize for pixel size.

The signal to noise ratio may now be calculated using the results from the above signal and noise modeling for various object, lighting and imaging properties. For objects with an image greater than one pixel in size, the signal should be above the noise floor to be detectable. Further, a threshold SNR may be set to determine if the object is detectable by an algorithm. Richardson details an approach to signal modeling for objects where the image is less than one pixel in size [12].

Information Theory Approach to Camera Ranking

A challenge that remains is combining resolution, signal and noise models into a method that allows the comparison of camera systems that have competing aims. For example, Camera A, may have high sensitivity due to a larger pixel, but low resolution, whereas Camera B may trade this sensitivity for resolution, by having smaller pixels. Compounding this may be different focal lengths, f-numbers of lenses used, as well as choices of color filter array and available quantum efficiency versus the spectra of light available for detection.

A great deal of performance analysis of neural networks relies on modification and use of information theory [16]. If neural network performance does increase as the information made available to it increases, then we might assume, that to a first order, a camera capable of providing more objective information describing a given scene or target would be desirable. Of course, a secondary concern is that, above a certain threshold, increase in network performance may well be asymptotic with a given increase in information leading to diminishing returns for increasing camera cost. Additionally, once network performance is sufficient for the task at hand, namely driving, any additional system cost is superfluous.

Information capacity, C, describes the ability of a system to store or transmit information in an objective manner and may be generally defined as [17]:

$$C = n \log_2 m \quad (23)$$

where m is the number of independent levels a symbol may transmit and n the total number of symbols. Using the above for example, the maximum information that may be carried by a 640×320 pixel, 256 level image would be:

$$C = 640 \times 320 \log_2(256) = 204800 \times 8 = 1638400\ bits = 204800\ bytes = 200\ Kb.$$

Jenkin has previously detailed the estimation of the information capacity of an imaging system [18], and we only superficially describe it here.

Earlier, it was shown that MTF modeling may be used to calculate an effective cell size for an imaging system via conversion to the LSF and subsequent determination of encircled energy or LSF width at a specific signal level. The total number of symbols, n, in the image is then:

$$n = \frac{A_{SENSOR}}{\pi\left(\frac{W_{LSF}}{2}\right)^2} \quad (24)$$

where A_{SENSOR} is the area of the active portion of the sensor and W_{LSF} the width of the LSF. The signal level and noise in the pixel may be determined using the techniques above for a target at a given range. As noise in the system increases, its ability to record independent distinguishable levels decreases as they require more separation. If the noise is ergodic, the number of recording levels in a single pixel is [19]:

$$m \approx \frac{PH_{PIXEL}}{2kN_{TOTAL}} + 1 \tag{25}$$

where k is a constant. Because our effective cell size is larger than a single pixel, a correction is required according to Selwyn's Law, detailed in [18] reducing the effective noise fluctuations, becoming:

$$m \approx \frac{PH_{PIXEL} \cdot W_{LSF}}{2k\, N_{TOTAL} \cdot p} + 1 \tag{26}$$

where p is pixel size as previously. We can now use numerous approaches to compare camera systems. Using a 100% Lambertian reflector, illumination conditions and exposure, it is possible to calculate the maximum number bits possible that may be recorded for a single frame.

While the above is useful for comparing camera systems of a similar field of view, it does not give us an idea of how a scene of interest is mapped to the field of view of the camera, namely, absolute performance with respect to the real world. We can achieve this by calculating the solid angle of the lens, and dividing the information capacity per frame to yield information per steradian per frame. As an example, if we compare an 8Mp 100-degree HFOV versus a 3.5Mp 50-degree HFOV camera, each with the same format sensor and f-number lens, such that they can both record 256 distinct levels, they would yield 7.6 Mb and 3.8 Mb *per frame* respectively. The 8Mp camera records more information per frame. The solid angle of the lenses, however, is 0.01 and 0.0039 steradians, yielding 761 Mb Str^{-1} and 868 Mb Str^{-1} respectively. The 3.5Mp system actually records more information per unit solid angle than the 8Mp. The coverage of the scene by each system is vastly different, which is also an important consideration, but the basic ability to identify an object within the field of view is now objectively compared. The above is a trivial example, as the number of recorded levels were contrived to be the same, but once the effects of f-number, quantum efficiency and sensor performance are included, this becomes a powerful technique with which to fairly evaluate the ability of various systems with hugely differing specifications.

A final approach is to move from generalized throughput for the camera system to that for specific targets. By multiplying Eq.19 by the reflectivity of the object in question, O_{REF}, and substituting the area of the image of the target, $A_{TARGET,}$ for the active area of the sensor in Eq.24 it is possible to calculate the number of bits yielded for a specific target and conditions. The area of the target may be calculated using Eq.2 to yield the width and height. Incorporating the target reflectivity and size as opposed to calculating maximum throughput is interesting as it forces the calculations into regimes where the noise floor of the sensor is more acute and may affect results in low light conditions.

Figure 11 shows maximum information capacity per frame per steradian for a fictitious 6.0 × 3.4 mm sensor with varying pixel sizes, f2.2 50-degree HFOV lens with transmission 0.9. Read noise is set at 3 electrons, PRNU at 0.5%. A typical monochrome quantum efficiency curve is used in conjunction with a 670 nm

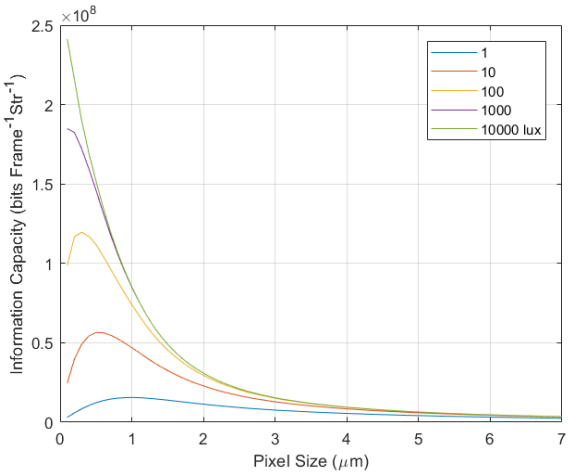

Figure 11. Information capacity per frame per steradian versus pixel size for constant sensor format 6.0x3.4mm, with 50-degree HFOV f2.2 lens. Other imaging conditions are described in the text.

infrared cut-off filter. Ambient illumination is modeled at a color temperature of 5500k and 1, 10, 100, 1000 and 10000 lux. The mean transfer function (Eq.12) is calculated and the aperture for the lens MTF is degraded 10%. The full-width-half-max of the LSF is used for cell size estimation. The integration time is 10ms. Wavelengths are modeled between 380 and 700nm. Parameters are chosen to demonstrate the capability of the modeling and do not represent a real system.

It may be seen that the information capacity continues to rise as pixel size is made smaller for all light levels when keeping the sensor size the same. This is intuitive as the number of pixels increases. The information capacity reaches a peak and starts to diminish, however, as the signal that is being divided by an increasing number of pixels gets smaller and is challenged by the noise floor of the sensor. The pixel size that produces the optimum

information capacity increases as the mean light level decreases as may be noted. The position of these peaks will shift as the noise sources in the sensor are modelled with increasing accuracy. There is little increase in information capacity above 100 lux for these exposure conditions aside from pixel sizes below 0.9μm. The light level modelled is saturating the pixel for the exposure time.

Figure 12, illustrates information capacity per frame of the 8Mp versus 5Mp sensors with the same sensor format, f/2.2 100-deg HFOV lens with ambient illumination of 5500k between 1 and 60 lux. The read noise has been artificially set to 6e- and 3e- in the 8Mp and 5Mp sensors respectively and all other imaging parameters are as before. It may be seen that the modeling correctly predicts that the 8Mp sensor will outperform the 5Mp at all light levels above approximately 2 lux. Below 2 lux the curves merge, the high read noise modeled in the 8Mp sensor is degrading the low light performance. This effect is exacerbated when we additionally model a 1.8×0.5 person in the scene at a distance of 25m with mean reflectance of 0.5%, Figure 13. The low reflectance of the person forces the signal well into an effective sub-lux range of 0.005-0.3 lux despite the ambient conditions of 1 to 60 lux. The crossing point of the two sensors, where the 5Mp device starts to outperform the 8Mp, is seen to be approximately 30 lux. Adding a headlamp of 2100 lumens, 3200k with a beam spread of 60 degrees horizontal and 30 degrees vertical, Figure 14,

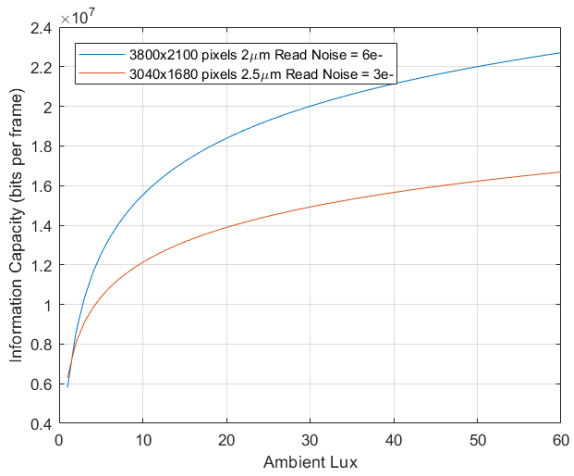

Figure 12. Information capacity per frame of 8Mp versus 5Mp sensors with the same sensor format, f/2.2 100-deg HFOV lens with ambient illumination of 5500k versus ambient light level. The read noise has been artificially set to 6e- and 3e- in the 8Mp and 5Mp and the exposure time is 10ms. Other parameters are specified in the text.

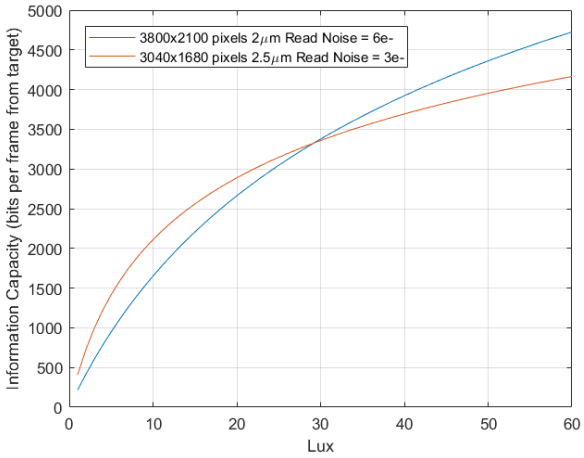

Figure 13. As for Figure 12, though imaging a person of 1.8x0.5m and mean reflectance of 0.5% at a distance of 25m. Note the increased performance of the 5Mp over the 8Mp below 43 lux.

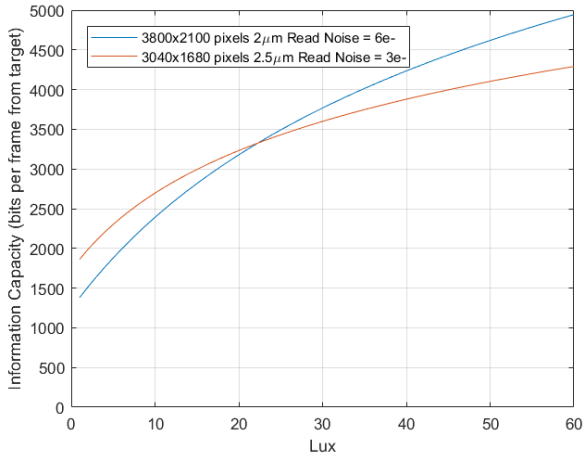

Figure 14. As for Figure 13 with a 2100 lumen headlamp added with color temperature of 3200k, horizontal and vertical beam spread of 60 and 30 degrees respectively. Note the increased information recorded over that in Figure 12, however, also that the crossing point remains as the object is only poorly lit by the headlamp at 25m distance.

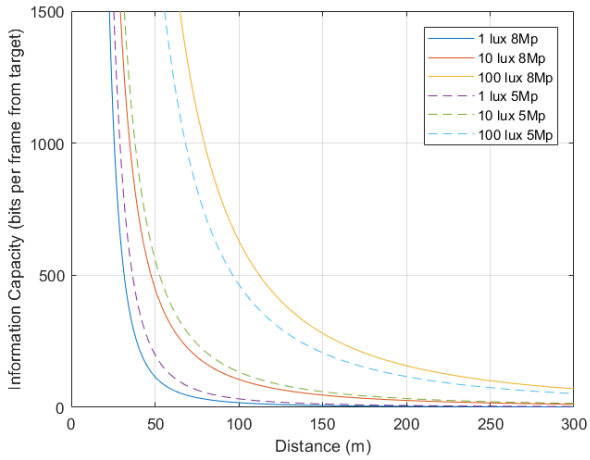

Figure 15. Information recorded from the target and imagers described in Figure 14 versus distance. At 100 lux the 8Mp imager outperforms the 5Mp. At 10 and 1 lux, the 5Mp system performs best for the distances shown.

improves the recorded information but does not eliminate the crossing point between the sensors because, at 25m the person is only dimly lit by the modelled headlamp.

Generating curves of information recorded from the target per frame versus distance for the above imaging conditions with the headlamp in place for 1, 10 and 100 lux again illustrates the ability of the modeling to predict the better performance of the 5Mp sensor in the low light conditions, Figure 15. As we examine distances where the object comes into range of the headlamp, however, we can see that the 8Mp again starts to outperform that of the 5Mp, below 10m, Figure 16. We would expect this crossing to occur at greater distances for materials that are retroreflective and can return more of the headlamp illumination, illustrating the need for improved and varied object modelling. It is worth reiterating that the sensor and imaging parameters have been set up to illustrate the capability of the modelling and do not represent real systems.

Figure 17 elucidates why the Johnson criteria approach works well in good lighting conditions by comparing information capacity curves with Johnson curves for a 1.8×0.5m person with 1000 lux ambient lighting using the above 5Mp and 8Mp parameters. When imaging is not limited by poor signal, geometry completely dominates the information capacity calculations. The curves are a similar shape, aside from a scaling error, which is determined by the method chosen to evaluate the LSF width and

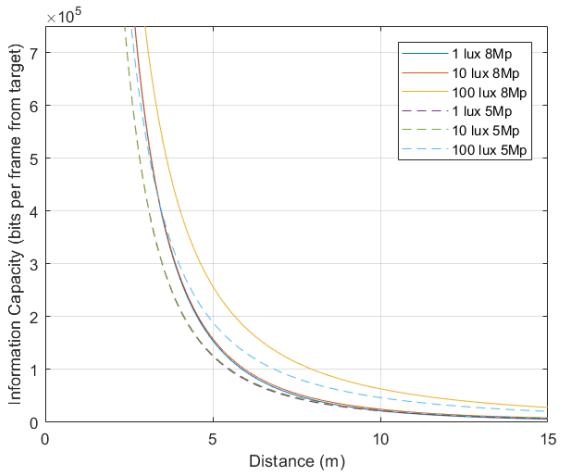

Figure 16. As for Figure 15, though rescaled to focus on the 0 to 15m range. Notice that as the distance between the object and the headlamps is reduced the 8Mp sensor again starts to outperform the 5Mp due to the increased illumination reaching the object.

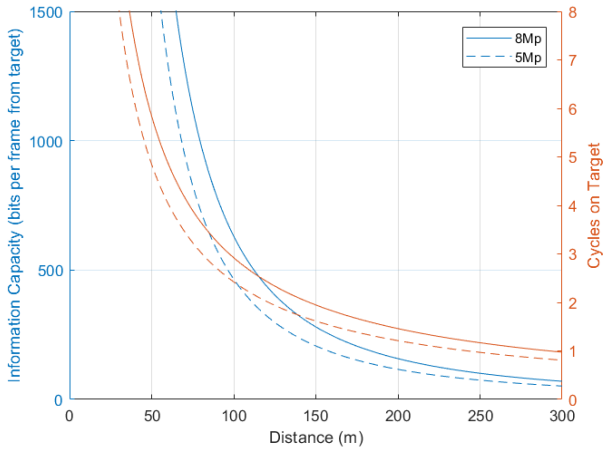

Figure 17. A 1.5x0.8m person, imaged at 1000 lux, 5500k ambient illumination with the 8Mp and 5Mp systems and f2.2, 100-deg HFOV lens versus distance, blue curves. Also shown is the Johnson curve for the sensors determined using the FWHM of the calculated LSF for each, orange curve.

hence the cell size in the information capacity calculation, as opposed to the pixel size in the Johnson calculation.

Monochromatic systems have been modeled in this paper, however, it is perfectly feasible to model color systems by using quantum efficiency curves for each channel to arrive at the signal level for each. Color crosstalk may then be modelled by mixing a proportion of each channel determined by the CFA arrangement. Instead of assuming a neutral reflector, a term representing the spectral reflectance of the object may be added into Eq.19. The multiple channel results may then be input into color correction matrices etc. and, if a number of colored objects are modeled, color separation determined.

Further work is needed to model high dynamic range systems and it is anticipated that this may be achieved by repeating calculations for multiple exposure times and combining the results or by adding terms representing de-sensitization of channels.

Conclusion

An initial approach to modeling and analyzing camera systems intended for autonomous vehicles has been presented that has drawn almost entirely from the existing imaging science field. First order ranking is generated by the use of the Johnson criteria. Threshold detection may be estimated by signal and noise models. A more sophisticated approach is afforded by information capacity modeling that is able to predict change in rank of systems with respect to light level and distance. The models are able to compare camera systems with vastly differing imaging specifications, with reference to target parameters and scene coverage, to predict which system will provide a network with the most objective information. Further work is needed to extend this modeling to color and HDR systems which also account for retroreflective materials.

References

[1] J. Johnson, “Analysis of Image Forming Systems”, Image Intensifier Symposium, Fort Belvoir, Va., 1958. P.249

[2] R. Jenkin, et al, “Analytical MTF Bounds and Estimate for SFR in Discrete Imaging Arrays Due to Non-Stationary Effects”, J. Img. Sci. Tech., vol. 47, no. 3, pp. 200-208, 2003.

[3] Keelan, B. W., Handbook of Image Quality: Characterization and Prediction, Marcel Dekker, Inc., New York, (2002).

[4] M. Richardson et al., Surveillance and Target Acquisition Systems, Brassey’s Land Warfare Series, London, 1997.

[5] E. Allen and S. Triantaphillidou, Eds., 10th Ed., The Manual of Photography, Focal Press, London 2011, p 108.

[6] European New Car Assessment Programme (Euro NCAP), Test Protocol – AEB VRU systems, Version 2.0.2, November 2017.

[7] C. N. Proudfoot, Ed., Handbook of Photographic Science and Engineering, 2nd Ed., IS&T, 1997.

[8] B. Keelan, “Imaging Applications of Noise Equivalent Quanta”, Proc. Electronic Imaging, Image Quality and System Performance XIII, Imaging Sci. and Technol. 2016.

[9] R. Jenkin, PhD Thesis, University of Westminster, London, 2001.

[10] J. C. Feltz and M.A. Karim, "Modulation Transfer function of Charged Coupled Devices”, Appl. Opt., vol. 34, no. 4, pp. 746-751, 1990.

[11] M. A. Richardson, "Electro-Optical Systems Analysis Part 1" Jour. Battlefield Technol., vol. 5, no. 2, p. 24, 2002.

[12] M. A. Richardson, "Electro-Optical Systems Analysis Part 2" Jour. Battlefield Technol., vol. 5, no. 3, p. 21, 2002.

[13] R. W. G. Hunt, 2nd Ed., Measuring Color, Ellis Horwood, 1991.

[14] G. C. Holst and T. S. Lomheim, CMOS/CCD Sensors and Camera Systems, 2nd Ed., SPIE, 2011.

[15] J. C. Dainty and R. Shaw, Image Science: Principles, Analysis and Evaluation of Photographic-type Imaging Processes. London: Academic Press Ltd., 1974.

[16] R. Shwartz-Ziv, N. Tishby, “Opening the Black Box of Deep Neural Networks via Information”, arXiv:1703.00810 [cs.LG], 2017.

[17] G. C. Higgins, Jour. App. Photo. Eng., no. 3, pp. 53, 1977.

[18] R. B. Jenkin, S. Triantaphillidou and M. A. Richardson, "Effective Pictorial Information Capacity as an Image Quality Metric" Proc.

Electronic Imaging, Imaging Sci. and Technol./ SPIE, vol. 6494, 2007.

[19] J. H. Altman and H. J. Zweig, J. Phot. Sci., no. 7, pp. 173, 1963.

Author Biography

Robin Jenkin received, BSc(Hons) Photographic and Electronic Imaging Science (1995) and his PhD (2001) in the field of image science from University of Westminster. He also holds a M.Res Computer Vision and Image Processing from University College London (1996). Robin is a Fellow of The Royal Photographic Society, UK, and a board member of IS&T. Robin currently works at NVIDIA Corporation where he maintains an interest in modeling image quality. He is also a Visiting Professor at University of Westminster within the Computer Vision and Imaging Technology Research Group. This work was completed at ON Semiconductor in his role leading algorithm and prototype module development.

Paul Kane received the M.S. in Optics from the University of Rochester, NY. He was a scientist at the Kodak Research Laboratories for 28 years, working primarily in the areas of imaging science and optics. His projects there included system modeling and simulation, image processing for OLED displays, 3D imaging and modeling light scattering from microparticles. In 2015 he joined ON Semiconductor as an Algorithm Design Engineer, focusing on automotive and security applications. He holds 35 U.S. patents in the areas of optics and image science.

Free access to this paper is brought to you with the generous support of ON Semiconductor.

All research funding for this paper is referenced in the text; unless noted therein, no research funding was provided by ON.