On Sensor Bias in Experimental Methods for Comparing Interest-Point, Saliency, and Recognition Algorithms

Alexander Andreopoulos and John K. Tsotsos, *Senior Member*, *IEEE*

Abstract—Most current algorithm evaluation protocols use large image databases, but give little consideration to imaging characteristics used to create the data sets. This paper evaluates the effects of camera shutter speed and voltage gain under simultaneous changes in illumination and demonstrates significant differences in the sensitivities of popular vision algorithms under variable illumination, shutter speed, and gain. These results show that offline data sets used to evaluate vision algorithms typically suffer from a significant sensor specific bias which can make many of the experimental methodologies used to evaluate vision algorithms unable to provide results that generalize in less controlled environments. We show that for typical indoor scenes, the different saturation levels of the color filters are easily reached, leading to the occurrence of localized saturation which is not exclusively based on the scene radiance but on the spectral density of individual colors present in the scene. Even under constant illumination, foreshortening effects due to surface orientation can affect feature detection and saliency. Finally, we demonstrate that active and purposive control of the shutter speed and gain can lead to significantly more reliable feature detection under varying illumination and nonconstant viewpoints.

Index Terms—Active vision, attention, shutter speed, gain, feature detection, saliency, recognition.

1 Introduction

The concept of *active perception* was introduced by Bajcsy [1] as "a problem of intelligent control strategies applied to the data acquisition process." The use of the term *active vision* for describing the problem was first introduced by Aloimonos et al. [2], where it was shown that a number of problems that are ill-posed and nonlinear for a passive observer are significantly simplified for an active observer. The idea that a serial component is necessary in a vision system was further popularized by Ballard [3]. Tsotsos [4] formalized the problem and proposed that the active vision problem should be considered a special case of the attention problem, which is acknowledged to play a fundamental role in the human visual system (HVS) [5], [6]. See [7] for a relevant literature survey.

The human visual system uses a number of complex and highly active mechanisms in order to compensate for luminance changes. Nonlinear intensity adaptation and contrast gain control in the human retina [8], [9] are some of those mechanisms. The information sent by the projection of certain ganglion cells directly to the pretectum [10] is largely responsible for the pupillary light reflex, which also compensates for luminance changes. Efforts have been made to build functional simulations of the retina [11], but the problem is still far from solved or completely understood. Adaptation to luminance is a significant topic of research in physiology and psychophysics, but related research in the robot and active vision literature for real-time parameter control has been limited. In contrast to the example set by the HVS, most current computer vision systems are characterized by their passive and often somewhat monolithic approach to adaptation under changes in radiance (e.g., autogain/autoexposure) or changes in the physical sensor used to sense the scene. It is common knowledge in the vision community that for most vision algorithms, physically changing the sensor used by a visually guided agent can cause a significant change in the system performance and may lead to the need for retraining any algorithm used on images acquired with the new sensor. Typical autogain and autoexposure algorithms implemented on board vision-based sensors operate based on the average image intensity, while disregarding the foreground region of interest, making such algorithms susceptible to background bias. As is demonstrated in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, for example, increasing the background intensity (e.g., a black versus a white background) while keeping the foreground constant can significantly change the autoexposure/gain settings, affecting an algorithm's ability to process the foreground data. Such weaknesses can easily be exploited by a malicious adversary in certain vision systems (e.g., automated surveillance or security applications). Images which are first contrast/gain normalized globally suffer from the same background/foreground dilemma. As discussed in Section 2, localized image contrast normalization does not necessarily maximize the likelihood of reliable

• *The authors are with the Department of Computer Science and Engineering, Computer Science and Engineering Bldg., Centre for Vision Research, York University, 4700 Keele St., Toronto, ON M3J 1P3, Canada. E-mail: {alekos, tsotsos}@cse.yorku.ca.*

Manuscript received 3 July 2010; revised 22 Dec. 2010; accepted 3 Mar. 2011; published online 28 Apr. 2011.
Recommended for acceptance by S. Belongie.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-07-0502.
Digital Object Identifier no. 10.1109/TPAMI.2011.91.

feature extraction either, due to saturation, noise, and nonuniform signal quantization effects induced by variable gain and exposure.

Digital vision sensors are primarily built using radiometric CCD and CMOS technologies. In general, CCDs have a better image quality than CMOS sensors, which tend to perform poorly, especially in the presence of little visible light (low radiance), and suffer from relatively high gain and offset fixed pattern noise (FPN) [12]. However, CMOS sensors tend to be significantly smaller and lighter than CCDs and require less power to operate [13]. The low weight of CMOS cameras often makes them ideal for a number of devices such as laptops and cellphones. CMOS circuit functions also tend to require fewer solder joints, which significantly decreases the chance of circuit failures in harsh environments (e.g., in high radiation environments such as outer space).

The dynamic range of a pixel on a digital camera characterizes the measurable range of sensor irradiance and is typically defined in terms of the maximum electron charge and minimum electron charge measurable by the corresponding potential wells [14], [15]. In the case of a grayscale sensor, the quantization of these charges using analog-to-digital (A/D) conversion results in a maximum gray value of I_{max} and a minimum electron charge detectable corresponding to a gray value I_{min}. If we ignore the effects of any quantization errors, the dynamic range [16] is given by

$$\mathrm{DR} = 20 \log_{10}\left(\frac{I_{max}}{I_{min}}\right) \mathrm{dB}. \quad (1)$$

For a CCD/CMOS camera with 8-bit channels, $\mathrm{DR} \approx 48\ \mathrm{dB}$ for each channel. In contrast, the dynamic range in human vision is over $90\ \mathrm{dB}$. The sensor irradiance is controllable by adjusting the shutter speed, or the aperture, if allowed by the camera. A number of papers have dealt with the problem of capturing high dynamic range (HDR) images based on sequential changes to camera exposure levels or aperture control [17], [18], [19], [20], [21], [22], [23], [24]. Other approaches are based on fusing multiple image detectors with varying exposures, using multiple sensor elements for each sensor pixel, and adaptively controlling the pixel exposures [16]. Problems in building a high dynamic range image from low dynamic range images include reduced spatial resolution, the requirement for complex and expensive hardware, and the inability to deal with motion [23].

As pointed out by Wandell et al. [25], studies in human temporal integration of photoreceptor signals show significant variation with the mean intensity level, hinting at the existence of specialized processes in the human visual system that deals with multiple exposures of varying duration. It is quite likely that a form of dynamic range extension of photoreceptor signals takes place in the HVS through the use of spatial mosaics created over an estimated range of 10-100 ms. Within this context, we investigate the extent to which adaptive exposure can help deal with the sensitivity of vision systems under noncanonical scene luminance, and investigate the extent to which sensor shutter speed and gain constitute a confounding factor in the reliability and consistency of the results obtained with offline data sets.

While the effects of illumination variations, image rotations, scale changes, and viewpoint changes on interest-point detectors and (to a lesser extent) saliency algorithms have been previously investigated in the literature [26], [27], [28], [29], [30], [31], the effects of variable sensor shutter speed and voltage gain under simultaneous changes in the illumination conditions and the identification of these camera intrinsic parameters as a significant source of bias in offline image data sets and the performance of vision algorithms have been largely overlooked in the literature. Given the ubiquity of shutter speed control and voltage gain control in modern digital cameras, the lack of a significant literature on the effects of such latent variables on the reliability of vision algorithms is surprising. An indication of the severity of the problem is evident in that, for most of the published work, and most of the offline data sets used to evaluate vision algorithms, there is no indication as to whether autogain/autoexposure was used during image acquisition nor is there any serious discussion of the sensor parameter settings used. For example, in the investigation by Schmid et al. [26] on the effects of the sensor's noise rate, there is no discussion on the effects of gain and shutter speed on algorithmic performance. As we show, such intrinsic camera parameters form latent variables which can have a nontrivial effect on image noise, contrast, the offline evaluation, and online performance of vision algorithms. While there is a rich literature on robust interest-point detectors under variable illumination and contrast [26], [27], [28], [29], [30], [31], such purely software-based approaches do not address the problem of making a vision system camera independent since they still exhibit sensitivity, in particular in low illumination conditions where dark current noise effects become prominent and in high illumination conditions where saturation effects may become evident in some of the camera channels. Furthermore, computer vision data sets (e.g., Caltech-101) are often gathered under controlled illumination conditions and are often prescreened to exclude images with corrupting artifacts (e.g., noise and saturation effects). In contrast, an active vision system whose extrinsic coordinate frame changes drastically over time can acquire images of questionable quality due to the large number of candidate viewpoints from which it can sense the scene [32]. Thus, enabling an active vision system to intelligently control the intrinsic camera parameters becomes all the more pertinent.

Within this context, we evaluate the effects of camera shutter speed and voltage gain, under simultaneous changes in illumination, on a number of popular and robust interest-point and saliency algorithms. We use these results to argue that the problem of active intrinsic camera parameter control has not attracted the amount of research that is commensurate with its significance for the evaluation and online performance of vision systems. We argue that a reliable vision system that is not sensitive to physically changing the mounted sensor and whose performance is not dependent on the sensor on which the system was trained must actively and purposively control the related camera parameters. We show that offline data sets used to evaluate vision algorithms typically suffer from a significant sensor specific bias, which can make many of the

experimental methodologies used to evaluate vision algorithms unable to provide results that generalize in less controlled environments. Section 2 discusses the role of shutter speed and voltage gain in the image formation process. Section 3 outlines the interest-point and saliency algorithms that we are evaluating. Sections 4 and 5 present our experimental methodology and results. Section 6 concludes the paper.

2 The Image Formation Process

We first discuss the effect of shutter speed and gain on the image formation process [12], [14], [15], [16], [33]. The photoelectric effect is a linear process, and as a result, CCD and CMOS sensors exhibit a strong linear relationship between the photon induced sensor charge and the resulting image intensity. Sometimes however, this linear relationship does not hold due to artificial nonlinearities which are introduced for image and color enhancement or image compression. The linearity is also constrained by a number of sources of error, such as quantization noise, shot noise, dark current induced noise, and amplifier induced read noise.

Two types of noise usually occur in radiometric sensors: random noise and pattern noise. Random noise is temporally random, can be described by statistical distributions, and can be dealt with by averaging successive frames. Pattern noise does not change frame by frame and is classified into fixed pattern noise and photoresponse nonuniformity (PRNU) noise. FPN noise is pattern noise that does not depend on illumination and is mostly present in CMOS sensors. It is typically caused by contamination during sensor fabrication or by the detector dimensions. PRNU noise is pattern noise that depends on illumination and can depend on the wavelength of illumination among other reasons. Such noise levels can affect the signal quality at low luminance levels for instance. In practice, a low shutter exposure time or a low gain value helps deal with blooming/saturation effects and provides a higher contrast image under high luminance. Similarly, under low luminance, we can obtain a higher contrast image by increasing the shutter exposure time. Small exposure times tend to lead to a decrease in the image's signal-to-noise ratio, while a high gain value magnifies some types of sensor noise.

Assume $E(x, y, s)$ denotes the accumulated charge for pixel (x, y) and over a time period s, where the charge is proportional to the number of photons falling on the response region of pixel (x, y) over this time period. Notice that for constant sources of illumination, it is reasonable to assume that $E(x, y, s)$ is a linear function of s. Given a gain setting $g > 0$ and a shutter exposure time s, an amplifier transforms the charge into a voltage $V(x, y, g, s)$, which is modeled as [15]

$$V(x, y; g, s) = (K_1(x, y)E(x, y, s) + K_2(x, y) + N_{DC}(s) + N_S(x, y, s) + N_R(x, y, s))A(x, y, g). \quad (2)$$

$K_1(x, y)$ and $K_2(x, y)$ are used to denote the sensor's fixed pattern nonuniformities and take values close to 1 and 0, respectively, for a good sensor. $N_{DC}(s)$ is the mean dark current induced noise, which is reducible by cooling the sensor, and is typically modeled as being proportional to s.

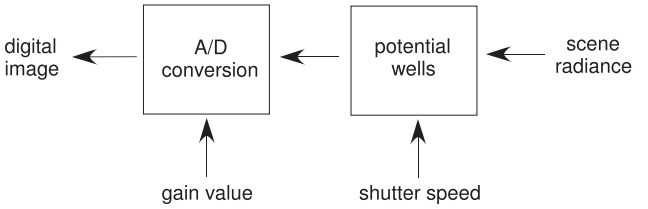

Fig. 1. The image formation process of a digital image. The potential wells accumulate electric charge that is proportional to the scene radiance over a period of time (shutter exposure time) that is inversely proportional to the shutter speed. Afterward, an analog-to-digital converter discretizes the measured photocurrent density. The gain factor affects the linear transformation of the measured voltage.

The dark current increases the average number of electrons sensed by the camera. $N_S(x, y, s)$ is called shot noise and characterizes the uncertainty in the number of electrons as a result of the quantum nature of light. It is modeled as zero mean Poisson shot noise with a variance equal to $K_1(x, y)E(x, y, s) + N_{DC}(s)$. $N_R(x, y, s)$ is called the read noise and is the result of using amplifiers to convert the charge into voltage. The read noise dominates shot noise for low illumination levels. $A(x, y, g)$ denotes the sensor gain value and is modeled by $A(x, y, g) = g + \epsilon(x, y, g)$ [12], where $\epsilon(x, y, g)$ is used to denote the fixed pattern noise associated with the sensor (see Fig. 1). Last, a quantization phase outputs the channel's integer valued image

$$I(x, y; g, s) = V(x, y; g, s) + N_Q(x, y), \quad (3)$$

where $N_Q(x, y)$ is the quantization noise. For most cameras with b-bit channels $I(x, y, g, s) \in \{0, 1, \ldots, 2^b - 1\}$. Typically, if the voltage is too high, $I(x, y, g, s) \triangleq 2^b - 1$.

From $E(x, y, s)$, we see that for constant illumination, the average accumulated charge is proportional to the shutter exposure time. Since the shot noise's variance is equal to $K_1(x, y)E(x, y, s) + N_{DC}(s)$, its variance is also proportional to the shutter exposure time. This implies that when all other noise sources disappear, each pixel's signal-to-noise ratio (defined as the mean signal divided by the noise's standard deviation) is proportional to $\sqrt{s}$. Of course, this does not take into account the saturation point of the potential wells. If too much charge is collected in a given sensor position/potential well, the charge tends to spill over into adjacent potential wells (mostly with CCDs), causing the blooming effect and saturating parts of the image, at which point the quality at that part of the image is catastrophically degraded. As we see from $V(x, y, g, s)$, the gain $A(x, y, g)$ amplifies the output voltage due to the signal $K_1(x, y)E(x, y, s) + K_2(x, y)$ and the noise sources $N_{DC}(s)$, $N_S(x, y, s)$, and $N_R(x, y, s)$. In the absence of any quantization related errors, $A(x, y, g)$ does not affect the signal-to-noise ratio of the image. In practice, a high value for $A(x, y, g)$ can cause image saturation problems, while a low value can lead to a poorly lit image. Notice that the variance of $V(x, y, g, s)$ is proportional to $A^2(x, y, g)$, demonstrating that for noisy sensors and sensitive vision algorithms, high gain values should be avoided. Often, a gain A_{cur} is expressed in decibels (dB), using the conversion $20 \log_{10}(\frac{A_{cur}}{A_{min}})$ dB, where A_{min} is the minimum discernible gain value and A_{cur} is the current gain value. Henceforth, the *shutter value* refers to the shutter's exposure time s,

which is inversely proportional to the shutter speed, and the *gain value* refers to its decibel scale value.

Color images are formed through the use of color filters with differing spectral responses. Multiple distinct color filters can be used to form color images, but such solutions are typically expensive. The most popular and cost effective method currently employed involves the use of a *color filter array*, where a *single* sensor is used, so that the sensor's surface is covered by a mosaic of colored filters [33]. Two popular color filter arrays use color sensitive filters $\{red, green, blue\}$ or $\{cyan, magenta, green, yellow\}$ as their color range. Obtaining the high-resolution color image (an RGB image, for example) is a matter of upsampling the outputs of the color filter array using various interpolation techniques such that a red, green, and blue value is extracted for each pixel position on the sensor. Naturally, such methods can blur edges or exacerbate aliasing effects. The chemical dyes making up such color filters are usually sensitive to environmental stresses such as humidity. The human visual system is known to disregard infrared wavelengths and, thus, most cameras are also built to reject the information at the infrared wavelengths using so-called "hot mirrors." However, these can affect the camera temperature, which in turn affects the sensor and dye responses, creating another source of sensor noise. The effect of different color filters' spectral responsivities and the improvements achieved by adjusting the shutter/gain values under differing scene luminance are exemplified in Fig. 2.

3 Interest Point Detectors and Saliency

We overview seven algorithms for interest-point detection and saliency computation, which are known to be fairly robust with respect to illumination changes. All the algorithms described have implementations that are publicly available online.[1,2,3,4,5] In the next sections and in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, we provide an exhaustive list of the results for all possible combinations of data sets, illumination, and interest-point/saliency algorithms used, documenting the effect of shutter/gain changes on the algorithms' robustness, under simultaneous changes in illumination. We also demonstrate the inability of the sensors' autogain/autoexposure mechanisms to consistently determine good shutter/gain values for generating the scene image.

We begin by overviewing the Harris-Affine and the Hessian-Affine region detectors [34]. Given two images of a scene viewed from different viewpoints, the Harris-Affine and Hessian-Affine detectors can detect in both images, regions of interest that are related by an affine transformation, making both methods relatively robust under viewpoint changes. This makes both methods suitable for a number of computer vision applications, such as wide-baseline stereo matching, for example. Assume $L(\mathbf{x}, \Sigma) = G(\Sigma) \otimes I(\mathbf{x})$, where $G(\Sigma)$ denotes a 2D Gaussian kernel with covariance Σ, $I(\mathbf{x})$ denotes an image, and $\otimes$ is the convolution operator. Let $\mu(\mathbf{x}, \Sigma_I, \Sigma_D) = det(\Sigma_D)G(\Sigma_I) \otimes (\nabla L(\mathbf{x}, \Sigma_D)\nabla L(\mathbf{x}, \Sigma_D)^T)$, where Σ_I denotes the region of integration of the Gaussian kernel and Σ_D is the region over which the derivative filter is defined. The Harris-Affine detector iteratively optimizes a number of metrics in order to find interest-point positions $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and covariance matrices $\{\Sigma_I^1, \ldots, \Sigma_I^m\}$, $\{\Sigma_D^1, \ldots, \Sigma_D^m\}$ such that matrices/tensors $\mu(\mathbf{x}_1, \Sigma_I^1, \Sigma_D^1), \ldots, \mu(\mathbf{x}_m, \Sigma_I^m, \Sigma_D^m)$ characterize the shape and local orientation of image regions centered at $\mathbf{x}_1, \ldots, \mathbf{x}_m$, respectively. The eigenvalues and eigenvectors of each matrix/tensor $\mu(\mathbf{x}_i, \Sigma_I^i, \Sigma_D^i)$ define an ellipse centered at point $\mathbf{x}_i$ which specifies the extent of the interest region. As is common in the literature and as we describe in more detail later in the paper, the measure of repeatability that we use to decide if two extracted interest points/regions match is based on the area of intersection of the two ellipses. This is referred to as the Harris-Affine interest-point detector since $\nabla L(\mathbf{x}, \Sigma_D)\nabla L(\mathbf{x}, \Sigma_D)^T$ is used in the popular Harris corner detector. The Hessian-Affine interest-point detector refers to a similar algorithm as the one overviewed above, with the main exception being that rather than using the Harris detector during interest-point detection, the Hessian corner detector is used instead.

Another popular interest-point detector is presented by Kadir and Brady [35]. The idea is to find scales and circular regions where the product of an entropy-based metric and the rate of change of intensity distribution with respect to changing scale are relatively high. In more detail, assume a circular region (x, R, s) is defined, where x denotes the region's center, R denotes all the pixels lying in this region, and s denotes the image scale. The Kadir-Brady interest-point detector finds triples (x, R, s) in the image that optimize the saliency metric

$$Y(x, R, s) = H(x, R, s)W(x, R, s),$$

where $W(x, R, s) = s \cdot \sum_{i \in V} |\frac{\partial P(i,x,R,s)}{\partial s}|$ and $H(x, R, s) = -\sum_{i \in V} P(i, x, R, s) \log (P(i, x, R, s))$. V is the set of values an image could take (e.g., $V = \{0, \ldots, 255\}$ in the case of an 8-bit grayscale image) and $P(i, x, R, s)$ is the probability that a pixel in region R under scale s takes a value of i. Usually, R is defined as a function of s. A clustering algorithm is applied to select representative triples (x, R, s) which do not change easily under variable imaging conditions (e.g., noise and small motions).

The fourth interest-point algorithm evaluated is the Maximally Stable Extremal Regions (MSER) algorithm by Matas et al. [36]. The algorithm can extract good correspondences from images that were acquired under different viewpoints, and it has also been applied to object recognition problems. The algorithm relies on the image's intensity level, and searches for local regions in the image such that for each region, all the pixel intensity values inside that region are larger (or smaller if it is a minimum intensity region) than the values at the boundary of the region. The algorithm attempts to make each one of these regions as large as possible by making sure that each region is next to a pixel which should not belong to this region. These regions can be clustered into sets, where each set contains regions $\mathcal{Q}_1, \ldots, \mathcal{Q}_n$ of pixels, such that $\mathcal{Q}_{i-1} \subset \mathcal{Q}_i \forall i$. For each such

1. http://www.robots.ox.ac.uk/~vgg/research/affine/index.html.
2. http://www.robots.ox.ac.uk/~timork/salscale.html.
3. http://www.vision.ee.ethz.ch/~surf/.
4. http://www.saliencytoolbox.net/.
5. http://www-sop.inria.fr/members/Neil.Bruce/.

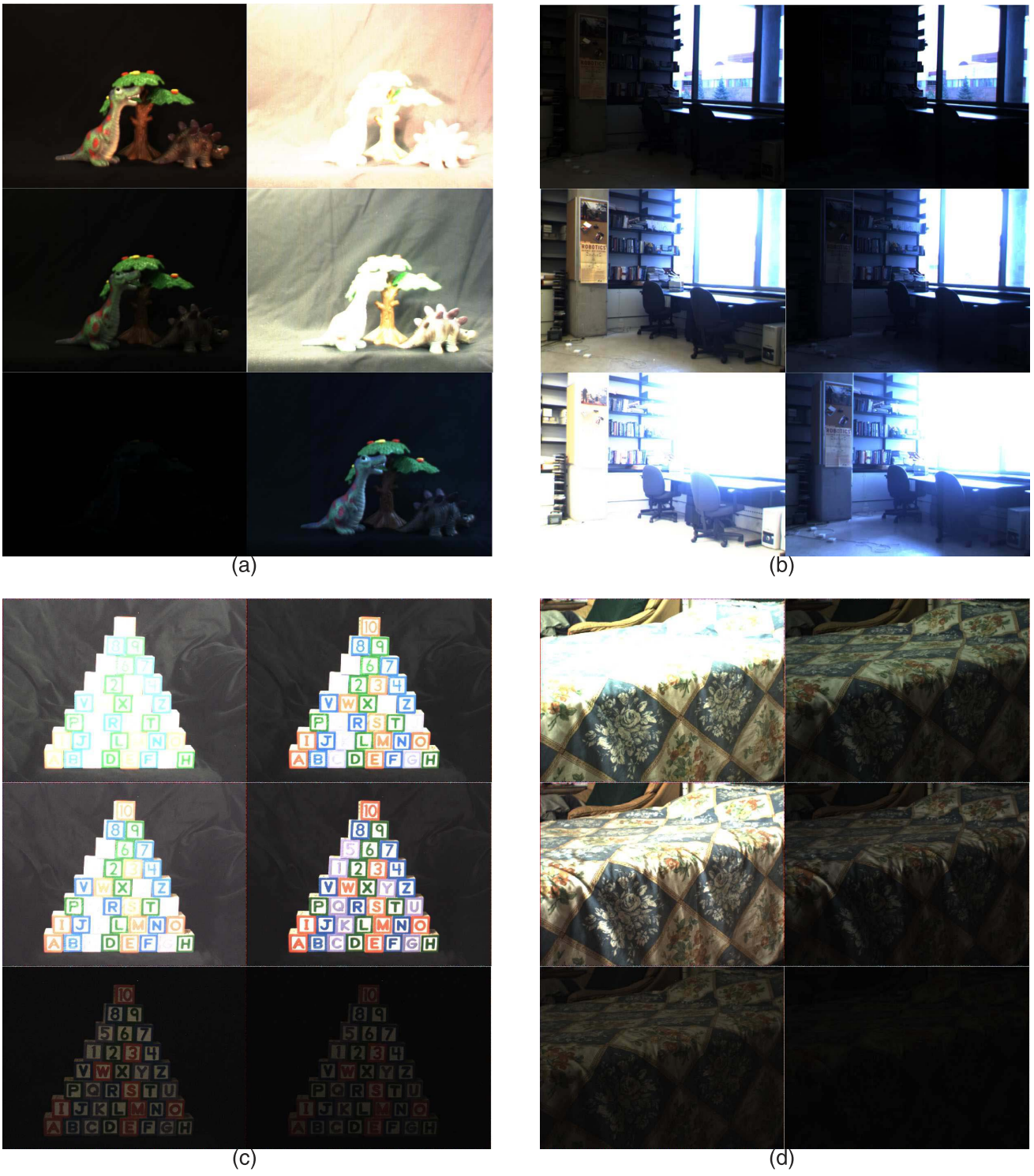

Fig. 2. Samples from all four data sets. See the supplementary documentation for a mosaic of all images acquired, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. (a) First data set—Toy dinosaur images. Each row denotes identical illumination conditions (top-bottom: illuminated, normal, and dark) and each column denotes identical shutter/gain values. (b) Second data set—Indoor versus outdoor blindness. Each column denotes identical illumination conditions (left:normal, right:dark) and each row denotes identical shutter/gain values. We cannot simultaneously discern outdoor and indoor features (limited dynamic range). (c) Third data set—Multicolored letters and numbers. Each row denotes identical illumination conditions (top-bottom: illuminated, normal, and dark) and each column denotes identical shutter/gain values. Notice the different color filters' sensitivities and the "disappearance" of many letters. (d) Fourth data set—Two perpendicular surfaces. Each row denotes identical illumination conditions (top-bottom: illuminated, normal, and dark) and each column denotes identical shutter/gain values. It is difficult to simultaneously discern both surfaces.

set, the algorithm labels a region $\mathcal{Q}_{i^*}$ as maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \backslash \mathcal{Q}_{i-\Delta}|/|\mathcal{Q}_i|$ has a local minimum at i^*, where Δ is an algorithm parameter.

The fifth algorithm is the Speeded-Up Robust Features extraction (SURF) algorithm developed by Bay et al. [37]. The algorithm combines an interest-point detector with a

descriptor for each of these interest points. Briefly, the interest-point algorithm approximates the Hessian matrix using box filters. The use of box filters makes it possible to efficiently calculate the Hessian of each image position by using the integral image and simply performing three subtractions on this image. Furthermore, the use of integral images makes it possible to calculate this Hessian matrix over multiple scales without having to calculate the scaled version of the input image. This makes the algorithm quite efficient. A nonmaximum suppression on the determinant of the Hessian matrix, across $3 \times 3 \times 3$ space and scale neighborhoods, helps reduce the number of interest-point candidates. The interest points are described in terms of their image position and scale (i.e., circular image regions). The distribution of intensity values within the interest-point's neighborhood is used as a descriptor. Similarly to the gradient information used by SIFT features, Haar wavelet responses in the x and y directions are calculated, and these responses are transformed into feature vectors which serve as neighborhood descriptors of the filter response and of the polarity of intensity changes.

The sixth method that we evaluate is the topographical saliency map created by the Itti-Koch-Niebur attention model [5]. We evaluate the effect that shutter/gain control and illumination has on this saliency map. The interesting feature of this model is that it ranks each image region based on a number of distinct image features such as color data, local orientation, and image intensity, providing a test case of how shutter/gain control affects a more complex vision system. Briefly, the model creates multiscale maps of color intensity and local orientation. At the next level, center-surround operators are applied in conjunction with a normalization process to create conspicuity maps, which in turn are fused to form a so-called saliency map. The higher the value in a particular position of this saliency map, the more salient/important this image position is according to the attention model.

The final method we evaluate is the *AIM* saliency algorithm described by Bruce and Tsotsos in [6]. Briefly, a set of learned filters, reminiscent of the ones found in area V1 of the human visual system, are used to extract a cascade of firing rates from an input image. In the publicly available implementation of AIM that we used, these filters were learned from natural images in the Corel Stock Photo Database, using ICA and PCA as a preprocessing step. In practice, these learned filters are similar to oriented Gabor filters and color opponent cells. A set of basis coefficients is extracted from each local image neighborhood on which these filters are applied. The coefficients around all local image neighborhoods lead to the creation of distributions of coefficients. For each image pixel, the corresponding local coefficient values that are extracted from the image are assumed independent (due to the use of ICA) and thus their probabilities are multiplied with each other, leading to a tractable joint probability value. The product of each local image region's probabilities is translated into an information theoretic measure using Shannon's measure of self-information. This results in an "information map" which corresponds to the calculated measure of saliency.

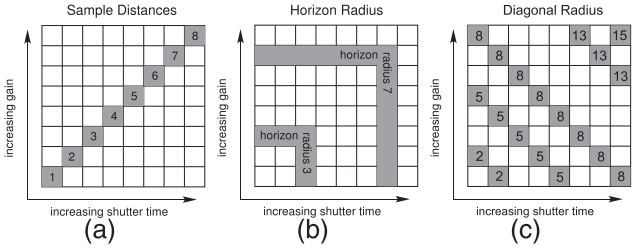

Fig. 3. The three strategies used to sample the space of the 8×8 shutter/gain value pairs out of which the results in Figs. 5, 6, 7, 8, 9, 10, 11, and the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, are constructed. (a) Each *sample-distance* value $1 \leq r \leq 8$, maps to the cell with shutter/gain values corresponding to bin (r, r). (b) A *horizon-radius* value r maps to the set of shutter/gain bins (s, g) such that $s \leq r$, $g \leq r$, and either $s = r$ or $g = r$. (c) A *diagonal-radius* value $1 \leq r \leq 15$ maps to the set of all shutter/gain bins (s, g) for which there exists an $i \geq 0$ such that $s = r - i$ and $g = i + 1$.

4 Experimental Setup

4.1 Test Data

We test the effect of shutter speed and gain control on four data sets acquired from four different scenes. We use one CCD sensor (a Bumblebee stereo camera by PointGrey Research) and one CMOS sensor (a FireflyMV camera by PointGrey Research) to obtain the data sets of images. The permissible shutter exposure time and gain ranges for the Bumblebee camera are 2-128 ms and 0.49-26.17 dB, respectively. The corresponding ranges for the FireflyMV camera are 0.06-33.31 ms and 0-12.04 dB, respectively. These permissible ranges are uniformly sampled using eight samples in each dimension. The ith sample from a range $[a, b]$ is set as $a + \frac{b-a}{8}(i-1)$, where $i \in \{1, \ldots, 8\}$, leading to 8×8 candidate settings for the shutter/gain parameters under which the corresponding images are acquired. We use these cameras to acquire four data sets, each data set sampling a different scene across all the 8×8 pairs of shutter/gain parameters and across different scene illumination conditions. In Fig. 3, we provide a diagram demonstrating the sampling strategies used to sample the 8×8 shutter/gain values out of which Figs. 4, 5, 6, 7, 8, 9, 10, and 11 are constructed. For each of the four data sets, a sample of images acquired under different shutter, gain, and illumination conditions is shown in Figs. 2a, 2b, 2c, and 2d, respectively, and the full data set of images is displayed in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. For each data set, we present mosaics of all the images acquired, under the different illumination conditions and the 8×8 shutter/gain settings. Figs. 2a and 2b show samples from the two data sets (data sets 1 and 2) captured using the CCD sensor and Figs. 2c and 2d show samples from the two data sets (data sets 3 and 4) captured using the CMOS sensor. For data sets 1, 3, and 4, we acquire samples under "dark," "normal," and "illuminated" scene conditions, which denote, as their names suggest, progressively better illuminated scenes. For data set 2, we acquire images under "dark" and "normal" illumination conditions. Each image in a data set is identified by an ordered pair (n, l) where $1 \leq n \leq 64$ maps to the shutter/gain value under which the

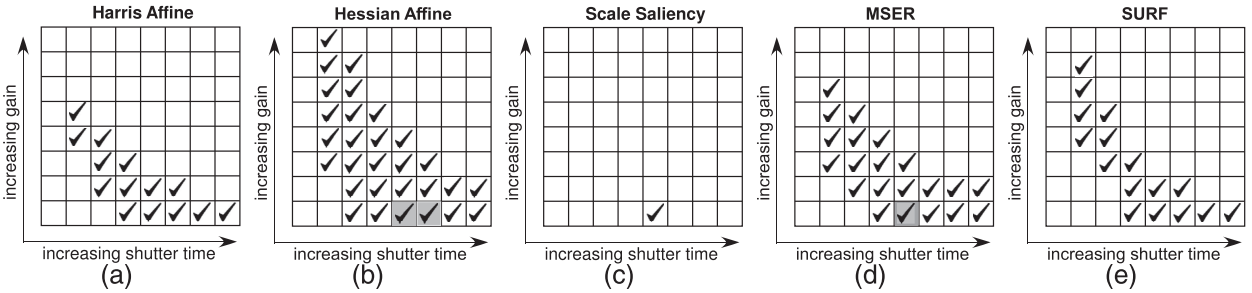

Fig. 4. Diagrams demonstrating the shutter/gain bins under which the average precision and recall rates for various interest-point algorithms (and across all five scenarios) are both above 0.5. Cells with check marks denote shutter/gain bins with average precision and recall rates above 0.5, where the precision and recall rates are calculated by limiting ourselves to the scenario images acquired under illumination conditions identical to the illumination conditions under which the scenarios' target images are acquired, thus limiting illumination variability. Gray colored cells have an average precision and recall rate above 0.5 across all the illumination conditions corresponding to the scenarios (a more challenging task). (a) The Harris-Affine algorithm results. (b) The Hessian-Affine algorithm results. (c) The Scale-Saliency algorithm results. (d) The MSER algorithm results. (e) The SURF algorithm results.

image was acquired and $l \in \{dark,\ normal, illuminated\}$ denotes the scene illumination conditions under which the image was acquired. We refer to (n, l) as an *image key*. The data set is also available online at http://www.cse.yorku.ca/LAAV/datasets.

In Fig. 2a, we see that under identical illumination conditions (dark/low illumination), certain image features are easily discernible if we appropriately adjust the shutter and the gain values, while the shutter/gain values that were quite appropriate under more illuminated conditions become utterly inappropriate in low illumination. In Fig. 2b, a qualitative observation is that we cannot discern well both exterior scene structure (outside the window) and interior scene structure, under differing illumination conditions. By adjusting the shutter/gain values, we obtain a more suitable setting for either the interior or exterior part of the scene. However, the sensor's dynamic range is too small and, thus, we cannot simultaneously discern interior and exterior space. In Fig. 2c, we notice that the different color filter sensitivities of the particular CMOS camera we are using have a significant effect on image contrast and feature detectability. A number of purple colored letters (C, K, Q, for example) end up returning saturated values, leading to a nonuniform change in image contrast—in this case, the change in contrast depends on the scene color—making the localization of all the cube letters impossible without some adjustment of the shutter/gain values. As Fig. 2c again demonstrates, by adjusting the shutter/gain values, we can

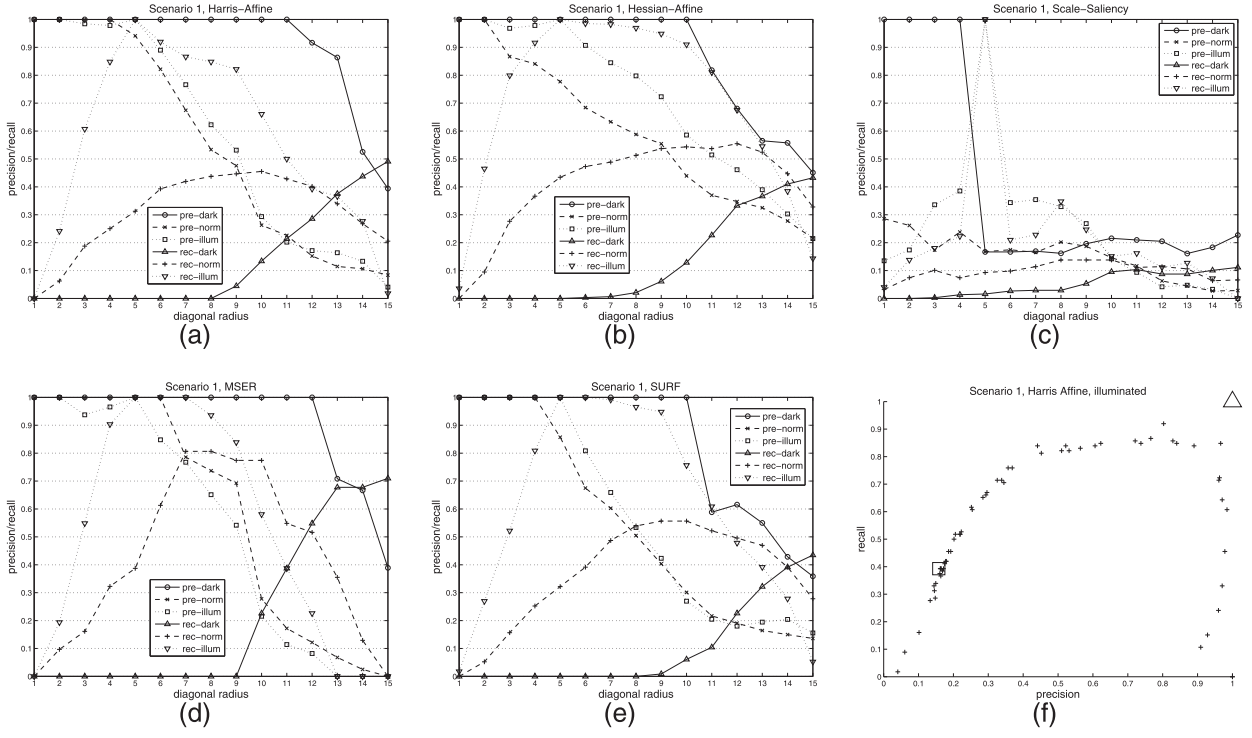

Fig. 5. (a)-(e) Scenario 1 results for the precision and recall rates of five interest-point algorithms under three different scene illumination conditions ("illuminated," "normal," and "dark"). All graphs were created using the diagonal-radius sampling strategy (see Fig. 3) and show the maximum precision/recall rates of the images that lie in each diagonal-radius value. (f) The precision-recall scatter plot of scenario 1, the Harris-Affine algorithm, and under "illuminated" conditions (the scenario's target image was also acquired under "illuminated" scene conditions).

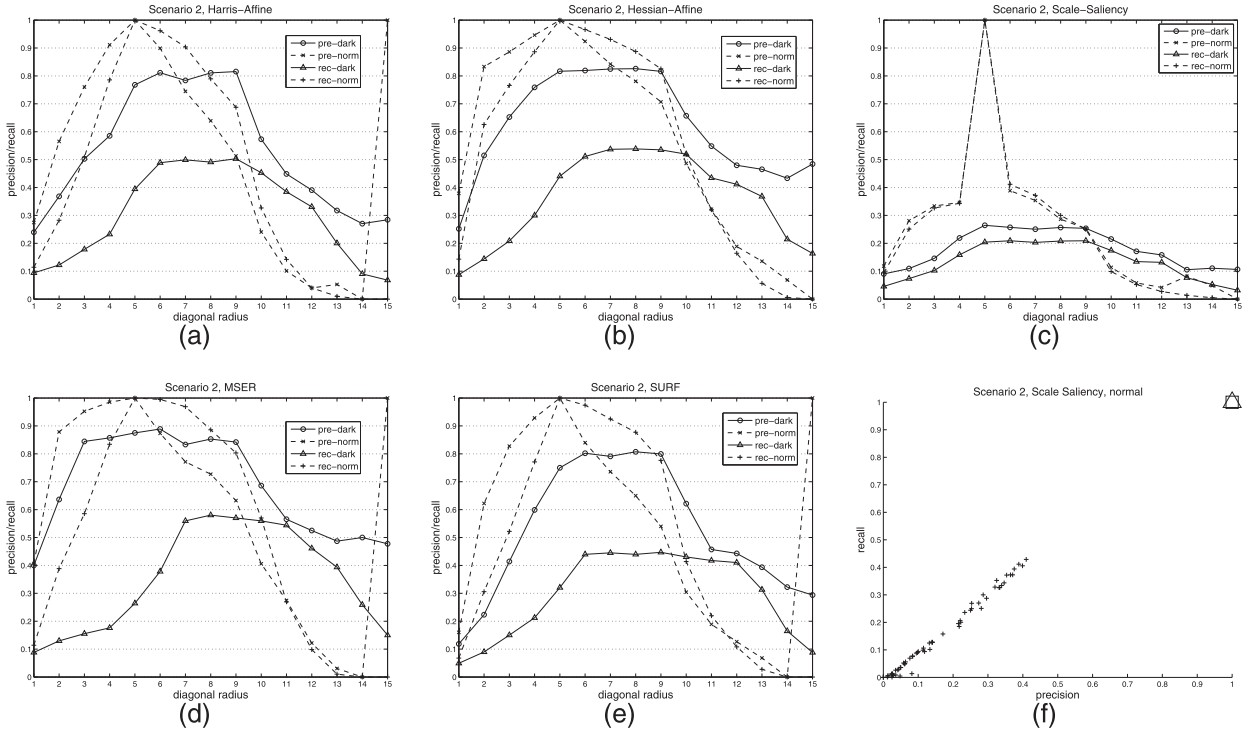

Fig. 6. (a)-(e) Scenario 2 results for the precision and recall rates of five interest-point algorithms under two different scene illumination conditions ("normal" and "dark"). All graphs were created using the diagonal-radius sampling strategy (see Fig. 3) and show the maximum precision/recall rates of the images that lie in each diagonal-radius value. (f) The precision-recall scatter plot of scenario 2, the Scale-Saliency algorithm, and under "normal" conditions (the scenario's target image was also acquired under "normal" scene conditions).

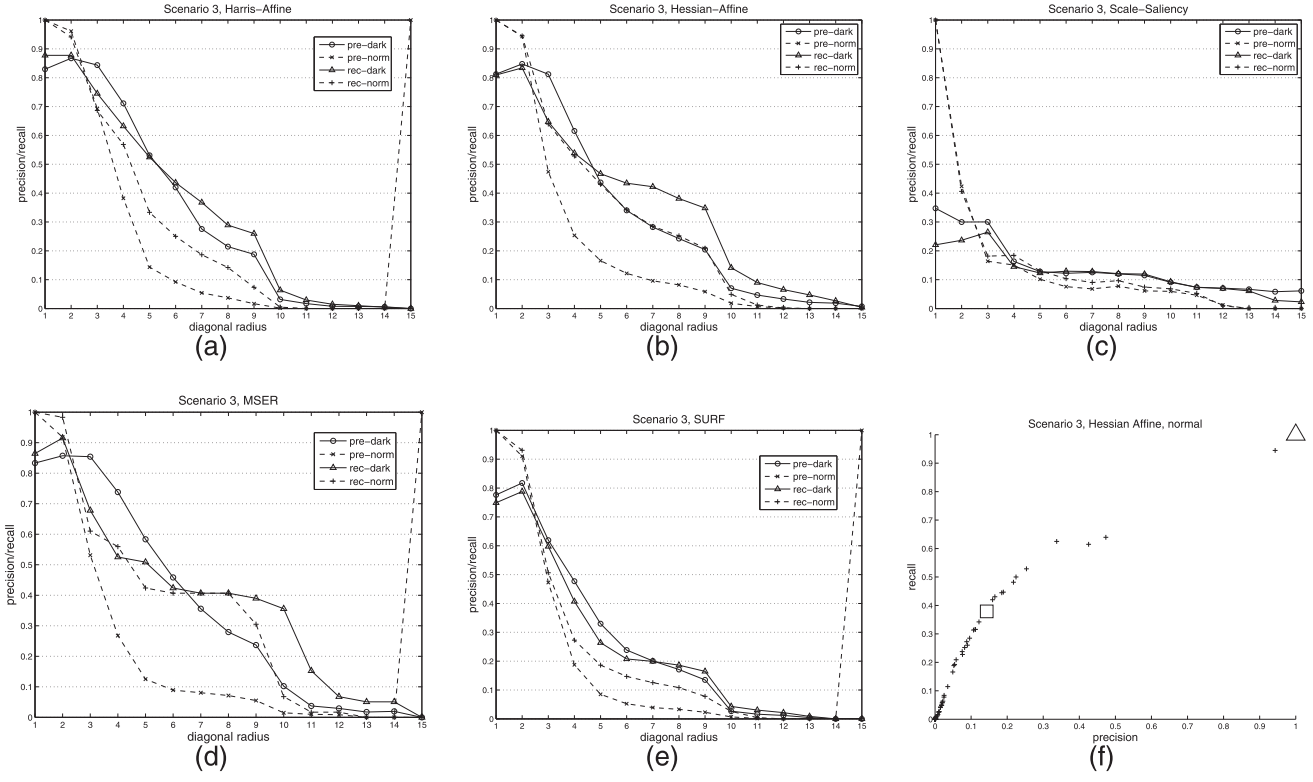

Fig. 7. (a)-(e) Scenario 3 results for the precision and recall rates of five interest-point algorithms under two different scene illumination conditions ("normal" and "dark"). All graphs were created using the diagonal-radius sampling strategy (see Fig. 3) and show the maximum precision/recall rates of the images that lie in each diagonal-radius value. (f) The precision-recall scatter plot of scenario 3, the Hessian-Affine algorithm, and under "normal" conditions (the scenario's target image was also acquired under "normal" scene conditions).

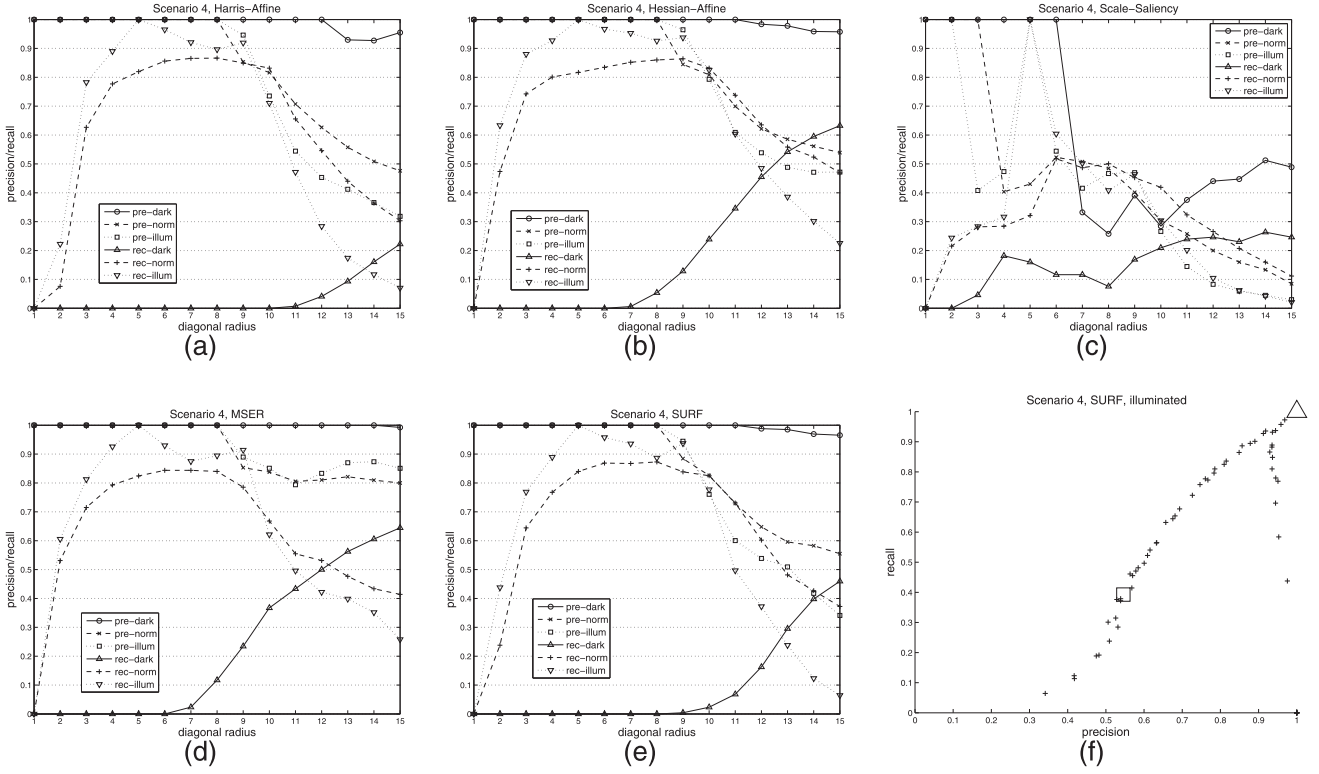

Fig. 8. (a)-(e) Scenario 4 results for the precision and recall rates of five interest-point algorithms under three different scene illumination conditions ("illuminated," "normal," and "dark"). All graphs were created using the diagonal-radius sampling strategy (see Fig. 3) and show the maximum precision/recall rates of the images that lie in each diagonal-radius value. (f) The precision-recall scatter plot of scenario 4, the SURF algorithm, and under "illuminated" conditions (the scenario's target image was also acquired under "illuminated" scene conditions).

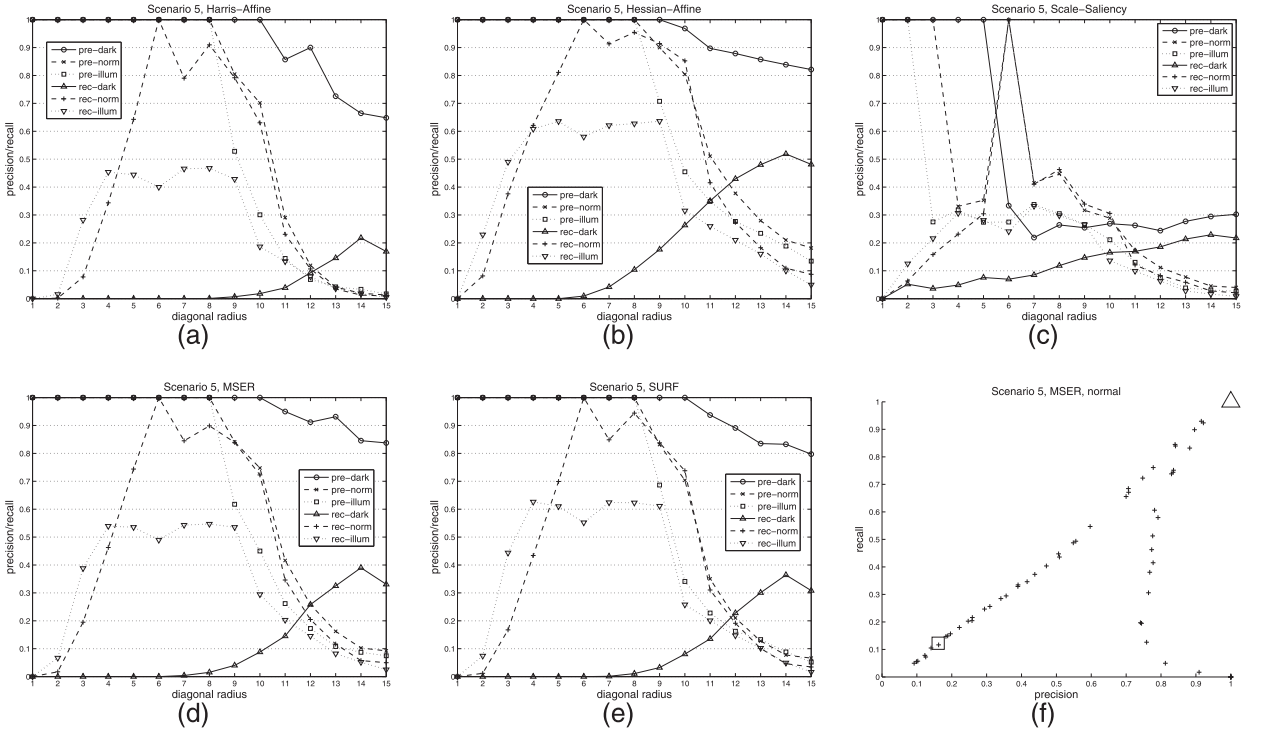

Fig. 9. (a)-(e) Scenario 5 results for the precision and recall rates of five interest-point algorithms under three different scene illumination conditions ("illuminated," "normal," and "dark"). All graphs were created using the diagonal-radius sampling strategy (see Fig. 3) and show the maximum precision/recall rates of the images that lie in each diagonal-radius value. (f) The precision-recall scatter plot of scenario 5, the MSER algorithm, and under "normal" conditions (the scenario's target image was also acquired under "normal" scene conditions).

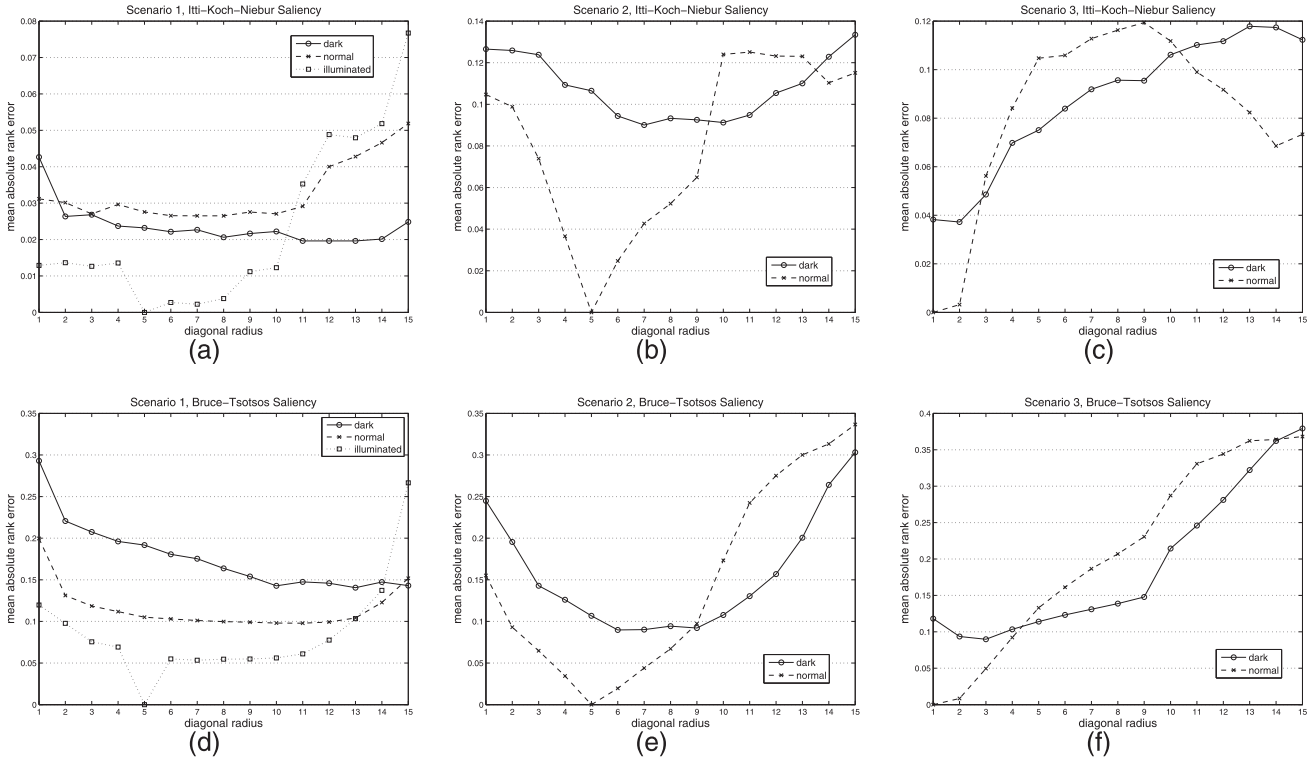

Fig. 10. Results of scenarios 1-3 on the Itti-Koch-Niebur (first row) and the Bruce-Tsotsos saliency maps (second row), under a variety of illumination conditions and under the diagonal-radius sampling strategy. For each scenario and saliency algorithm, we plot the minimum mean absolute rank error of all images in a particular diagonal radius.

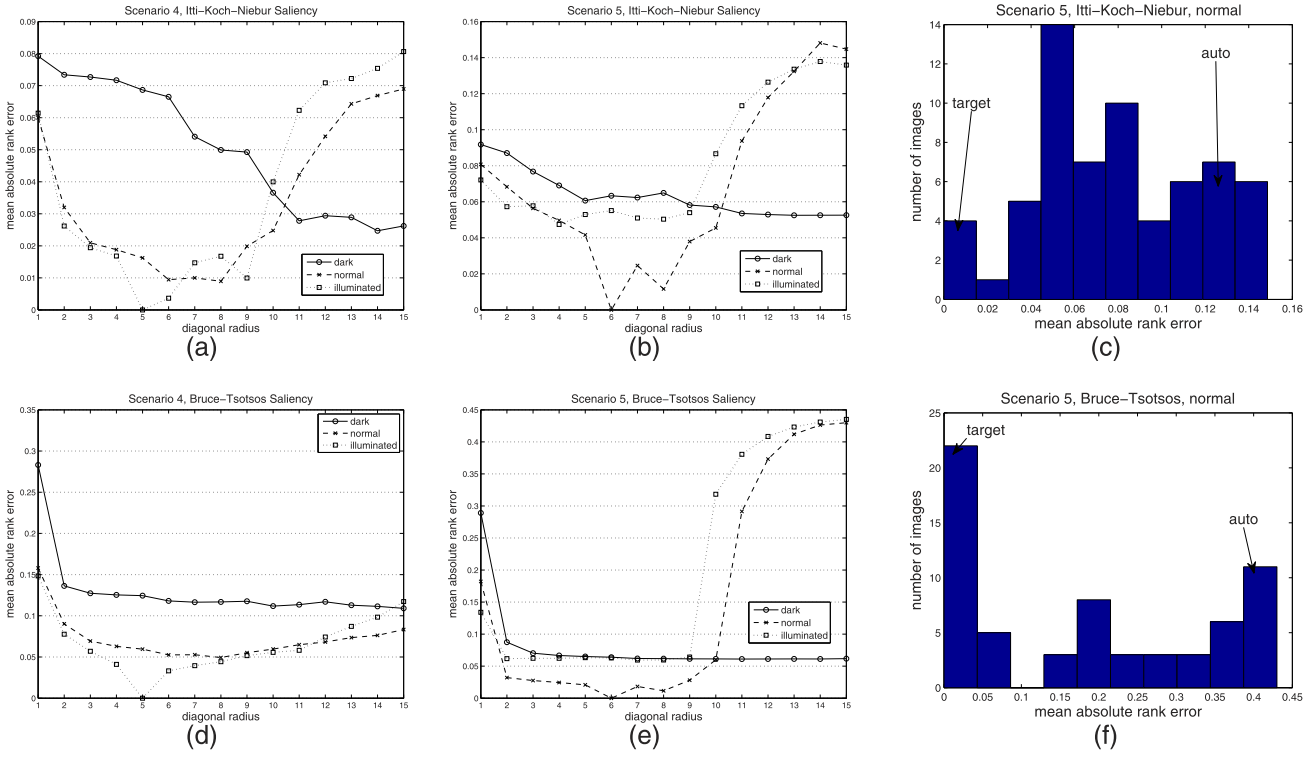

Fig. 11. (a), (b), (d), and (e): Results of scenarios 4 and 5 on the Itti-Koch-Niebur (first row) and the Bruce-Tsotsos saliency maps (second row), under a variety of illumination conditions and under the diagonal-radius sampling strategy. For each scenario and saliency algorithm, we plot the minimum mean absolute rank error of all images in a particular diagonal radius. (c) and (f): Distributions of the mean absolute rank errors for Scenario 5, and normal illumination. Notice the poor results of the autogain/autoexposure mechanism ("auto" label).

increase the image contrast in low illumination conditions, even though these shutter/gain values lead to a saturated image under high illumination. As we previously argued, a fundamental aspect of a good vision/active vision system is that it should neither be sensor specific nor too brittle under a change of sensors, and should be capable of compensating

for the sensor's inadequacies at a software or hardware level. In Fig. 2d, we show that the orientation of an object surface with respect to the light source has a decisive effect on detection of its surface patterns, demonstrating the need for adaptation techniques even in the presence of constant sources of illumination. The top surface, which is facing the light source, tends to suffer from saturation effects under illuminated conditions, while the other surface whose normal is perpendicular to the light source direction, reflects less light, resulting in lower contrast on average. A qualitative observation is that both surfaces are not optimally discernible under a single shutter/gain value.

4.2 Test Protocol

We examine the performance of seven different interest-point and saliency detectors (the Harris-Affine and Hessian-Affine detectors [34], Kadir and Brady's detector [35], the MSER detector [36], SURF's detector [37], the Itti-Koch-Niebur saliency algorithm [5], and the *AIM* saliency algorithm [6]), under variable illumination and simultaneous variations in the camera's shutter/gain values, using their publicly available implementations. Unless otherwise indicated, the default algorithm parameters are used in all the test cases. Notice that the algorithms we test use various operators, such as Difference of Gaussians, Laplacian detectors, Gabor filters, or sparse filters, that were mined from offline data sets in the case of [6], so we are also implicitly testing the repeatability of such low-level feature extraction algorithms. The four data sets (a subset of which is shown in Figs. 2a, 2b, 2c, and 2d) are used to construct five different testing scenarios. For each scenario, we select a *target image*, which is the image with respect to which we evaluate the change in the detected features under different shutter/gain values and different scene illumination. The target image is identical, regardless of the interest-point or saliency algorithm being tested, so that the results acquired with different algorithms are comparable. We thus let each scenario's target image be an image which is determined to have good contrast and few saturation or blooming effects present. A validation phase verifies that the selected target images return a good number of detected features with respect to the interest-point/saliency algorithms tested and the regions of interest, without fitting the image noise. In scenario 1, we select the corresponding target image from the first data set (see Fig. 2a) and refer to this image as target image 1. In scenarios 2 and 3, we select two images from the second data set (see Fig. 2b). We select the images which are determined to have among the greatest number of high contrast features visible inside the room and outside the room, respectively. We refer to these images as target images 2 and 3, respectively. Similarly, for scenarios 4 and 5, we select good quality target images from data sets 3 and 4, respectively (see Figs. 2c and 2d), which are referred to as target images 4 and 5, respectively. In the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, we list the exact settings under which each target image was acquired.

Notice that it is difficult to test each scene using two different sensors (which would provide a nice benchmark of CCD versus CMOS sensor performance) since the two sensors could have different fields of view, pixel sizes, focal lengths, depths of field, as well as different spectral response functions (differences in such camera properties are not inherently due to CCD and CMOS technologies). This would make the validity of any results questionable, as it is not possible to show that any potential differences in the features extracted by the two sensors are due to inherent differences in CCD and CMOS technologies and are not due to any other latent variables. However, with some caveats and through image resampling, image cropping, and very precise measurements, it might be possible for the differences in the two sensors to be mitigated somewhat.

Later in this section, we define the error metrics with respect to which the results are constructed, namely, the *precision rate*, the *recall rate*, and the *mean absolute rank error*. Briefly, for every image in every scenario, we calculate the corresponding precision rate, recall rate, and mean absolute rank error, quantifying the quality of the features extracted from that image with respect to the scenario's target image. We first define the images in each scenario and illumination condition that correspond to each *sample-distance* value, to each *horizon-radius* value, and to each *diagonal-radius* value (see Fig. 3 for a relevant diagram). Each *sample-distance* value $1 \leq r \leq 8$ maps to a single shutter/gain bin $(r, r) \in \{(1,1), (2,2), \ldots, (8,8)\}$. A *horizon-radius* value r corresponds to the set of $2r-1$ shutter/gain bins $(s, g) \in [1,8] \times [1,8]$ such that $s \leq r$, $g \leq r$, and either $s = r$ or $g = r$ (see Fig. 3b). Each *diagonal-radius* value $1 \leq r \leq 15$ corresponds to the set of all shutter/gain bins (s, g) for which there exists an $i \geq 0$ such that $s = r - i$ and $g = i + 1$ (see Fig. 3c).

Recall that an increase in s or g by 1 indicates an increase in the corresponding shutter exposure time or gain value (dB scale) by an additive constant. Notice that for each horizon-radius/diagonal-radius value, for each scenario, and each illumination condition, there corresponds a set of images that were acquired under the corresponding shutter/gain values. For each such radius value, for each scenario, each illumination condition, and each interest-point algorithm, we show in the appropriate graphs of Figs. 5, 6, 7, 8, and 9 (as well as Figs. 41-50 in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91) the maximum precision and recall rates of the images that lie in the corresponding set of images. Henceforth, we identify a figure x from the supplementary documentation using the postscript $\mathcal{S}$, as Fig. $x\mathcal{S}$, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. For each horizon-radius/diagonal-radius value, for each scenario, each illumination condition, and each saliency algorithm, we also output the minimum mean absolute rank error of the images lying in the corresponding set of images (see Figs. 10 and 11 and Figs. $42\mathcal{S}$, $44\mathcal{S}$, $46\mathcal{S}$, $48\mathcal{S}$, and $50\mathcal{S}$). In general, a high horizon/diagonal radius provides images with better contrast under low illumination and a low horizon/diagonal-radius value provides images with good contrast under high illumination. This allows us to investigate the existence of shutter/gain settings for the camera parameters that improve the feature detection under differing scene illumination. The radius approach enables us to parameterize using a single variable, the entire 8×8 space of shutter/gain values

in such a way that we could superimpose multiple precision/recall graphs in a single pair of axes so that an increase in the parameter value leads to a "brighter" image. A radius approach allows us to sample the full space of shutter/gain values and to demonstrate in a single graph, the effect of progressively changing the camera parameters under multiple distinct scene illumination conditions. In general, the images in each diagonal radius have a more uniform brightness than the images in each horizon radius (see the image mosaics in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91). The precision rate, recall rate, and mean absolute rank error of the image corresponding to each sample distance (for various scenarios, illumination conditions, and interest-point/saliency algorithms) are shown in the corresponding subgraphs of Figs. $41\mathcal{S}$-$50\mathcal{S}$. A sample-distance approach allows us to directly compare, in a single graph, precision and recall rates that were achieved for identical images. Notice that while the full spectrum of horizon/diagonal-radii samples the full 2D range of shutter/gain values and allows us to show the corresponding results in a single graph, the sample distance provides a 1D correlated set of progressively increasing exposure times and gain values.

We first describe the testing strategy for the five interest-point algorithms. As discussed in Section 3, each one of these interest-point detectors returns an ellipse describing the detected feature's position and its orientation (or support region in [35] and [37]). Assume we have extracted an interest point/ellipse from a target image t. Given that we use one of the above interest-point algorithms, let $\mathbf{X}_t^i$ denote a set containing all the image pixels lying inside the ith interest point/ellipse of target image t. Given another extracted interest point/ellipse from an image located in the same data set as target image t—if this second image has key (n,l), we use $\mathbf{X}_{n,l,t}^j$ to denote the set of points lying inside the jth ellipse extracted from this second image—we say that the two interest points/ellipses *match* if

$$\frac{|\mathbf{X}_t^i \cap \mathbf{X}_{n,l,t}^j|}{|\mathbf{X}_t^i \cup \mathbf{X}_{n,l,t}^j|} \geq \theta$$

for some threshold $0 \leq \theta \leq 1$. We use $\theta = 0.6$, which is neither too low nor too constraining, allowing us to draw certain conclusions on the tested algorithms. Let $\delta_t(\mathbf{X}_{n,l,t}^j) \in \{0,1\}$ take a value of 1 iff there exists an interest point i located in target image t so that $\mathbf{X}_{n,l,t}^j$ matches $\mathbf{X}_t^i$, where image (n,l) is located in the same data set as target image t. Similarly, let $\delta_{n,l,t}(\mathbf{X}_t^i) \in \{0,1\}$ take a value of 1 iff there exists an interest point j in an image with key (n,l) (where (n,l) is located in the same data set as target image t) so that $\mathbf{X}_t^i$ matches $\mathbf{X}_{n,l,t}^j$. For each scenario's target image and each illumination condition under which we acquired the corresponding data set, we evaluate the precision and recall rate of the target image's interest points as we vary the shutter and gain values. We test the ability of detecting the same interest points in the corresponding target image, under differing shutter exposure times, gain values, and scene illumination conditions. Given an interest-point extraction algorithm, the precision and recall rates corresponding to a given target image t and an image from the same data set as target image t and with image key (n,l) are given by

$$\text{Precision}(n,l,t) = \frac{\sum_{j=1}^{N} \delta_t(\mathbf{X}_{n,l,t}^j)}{N},$$
$$\text{Recall}(n,l,t) = \frac{\sum_{i=1}^{M} \delta_{n,l,t}(\mathbf{X}_t^i)}{M},$$

where N and M are the number of interest points in image (n,l) and target image t, respectively. If $N = 0$ (as might occur if no interest points are detected in image (n,l) due to very low luminance, for example), we define $\text{Precision}(n,l,t) \triangleq 1$. If $M = 0$ for target image t, $\text{Recall}(n,l,t) \triangleq 1$. A good interest-point detection algorithm would consistently display high precision and recall rates of over 0.5 since, otherwise, the signal potentially has poor information content. In the learning theory literature for example, an error rate of less than 0.5 (corresponding to a precision and recall of over 0.5) is usually necessary for a learning algorithm to be well behaved, justifying the use of 0.5 as a lower bound on the minimum quality constraints that good interest-point algorithms should satisfy.

We also measure the effect that shutter/gain control and changing illumination has on the order with which one would attend to locations on the saliency maps returned by the Itti-Koch-Niebur attention model [5] and the saliency model by Bruce and Tsotsos [6] (see Figs. 10 and 11). As before, to each scenario we assign the same *target image* and evaluate the minimum *mean absolute rank error* of all the images lying in each horizon-radius/diagonal-radius value and each sample-distance value: Assume we have chosen a target image t and an image key (n,l) from the data set corresponding to image t. For each pixel (i,j) in the saliency map of (n,l), we extract its rank $\mathbf{R}_{n,l,t}(i,j)$, which is defined as the number of pixels in the saliency map that have a smaller value than that of saliency map pixel (i,j). $\mathbf{T}(i,j)$ denotes the rank of pixel (i,j) in the target image's saliency map with respect to all the other saliency map values in the target map. The mean absolute rank error of image (n,l) is defined as the average value of $\frac{|\mathbf{T}(i,j)-\mathbf{R}_{n,l,t}(i,j)|}{P}$ for all P pixel indices (i,j) of the two saliency maps, where P is a normalization constant denoting the number of pixels in the saliency map. P makes the mean absolute rank value lie in range $[0,1]$ and thus makes it independent of the size of the saliency maps. In practice, with natural images, the occurrence of pixels with identical ranks is rare. An alternative approach for images where many such ties might occur is to sort all pixel values based on their saliency values (by breaking ties arbitrarily so that no two ranks are identical) and then resetting the ranks of all pixels with identical saliency values to be equal to their average rank. For a given horizon-/diagonal-radius value r which consists of a set of images acquired under different shutter/gain values with identical scene illumination conditions, we evaluate the minimum mean absolute rank error across all these images. Similarly, for each sample-distance value, we evaluate the mean absolute rank error of the unique image corresponding to the shutter/gain pair.

In Figs. 5f, 6f, 7f, 8f, and 9f and Figs. $21\mathcal{S}$-$35\mathcal{S}$, we evaluate how the precision and recall rates of the images designated

by the autogain and autoexposure mechanisms—represented by a square (□) symbol in the scatter plots—compare with all the other data set images which were acquired under the full spectrum of candidate sensor states. These other images are represented by a plus (+) sign in the scatter plot, and the image acquired under the same sensor state that the scenario's target image was acquired is represented by a triangle (△). Similarly, in Figs. 11c and 11f, we present the distribution of mean absolute rank errors for two of the worst performing scenarios (in Figs. $36\mathcal{S}$-$40\mathcal{S}$, we present an exhaustive list of all the results). In each histogram, we mark the bin of the image designated by the autogain and autoexposure mechanism (a label of "auto") and the bin of the image captured under the same shutter/gain values as the scenario's target image. This provides a demonstration of the inability of simple autogain/autoexposure mechanisms which rely on the mean image intensity to optimally choose the appropriate sensor state for sensing the features of an arbitrary object of interest. The inadequacies of such autogain/autoexposure mechanisms are further exemplified in Section 2 of the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, where we demonstrate the potentially catastrophic effects that background biases can have on the ability of a vision system to process the foreground data under variable illumination. These results support one of the paper's theses, that an active vision system that emulates the excellent abilities of the human visual system in arbitrary environments must purposively control the related sensor parameters. For brevity, we include some of the worst performers in the paper's main text, and the full set of results is available in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91.

5 Results and Discussion

Fig. 4 shows the sampled shutter/gain values for which both the average precision and the average recall rates of the interest-point algorithms are above 0.5. From Fig. 4, we notice that there is simply too much variation in the scenes (illumination-wise and structure-wise) to guarantee that a single camera setting can provide reliable results for all scene and algorithm variations. As we see from Fig. 4, there exist almost no constant camera settings (gray colored cells) that guarantee a minimum average precision and recall rate of over 0.5, under the greatest amount of scene variability (there only exist three gray cells in total in Figs. 4b, 4c, and 4d). Even if we limit ourselves to the illumination conditions under which the target images were acquired (normal and illuminated conditions), we notice very few checkmarked cells on average. This shows that reliable interest-point detection across all interest-point algorithms and all illumination conditions is difficult to achieve. In other words, while there exists a single shutter/gain bin $(5, 1)$ giving good results for all five algorithms (see Figs. 4a, 4b, 4c, 4d, and 4e), a small change in the shutter/gain bin is enough to give poor results in at least one of the interest-point algorithms, demonstrating that constant sensor parameters can provide an illusory confidence on the algorithms' performance since the best performing algorithm can vary as the sensor parameters change. Fig. 4 indicates that, at least for the popular interest-point algorithms tested, under a constant shutter/gain value it is difficult to acquire a robust vision algorithm using standard 8-bit channel cameras. Another implication is that if we have two sensors with constant but significantly different settings, the results can easily get affected by the sensor used due to the different saturation and contrast levels in the various corresponding image regions generated by the two sensors. As previously discussed, care must be taken since noise is magnified with increasing gain, while too small a shutter value also tends to make the camera's shot noise and dark current noise more prominent. On close examination of Figs. 4a, 4d, and 4e (and to a lesser extent Fig. 4b), we see that a linear/correlated pattern of check marks emerges, demonstrating that a decrease in the gain is successfully compensated by an increase in the shutter time, thus allowing us to better match the target image under variable sensor settings. As discussed in Section 1 and the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, under a sufficiently large range of illumination conditions all gain and shutter values are easily achievable when using an autogain/autoshutter algorithm which relies on the mean image intensity to adjust the sensor settings. For at least half of the sensor states in each subfigure in Fig. 4, the precision or recall rate is at most 0.5, corresponding to unreliable feature detection for the popular and illumination-robust algorithms investigated. This supports the claim that an evaluation of an algorithm's performance that does not take into consideration the effects of the camera parameters is incomplete.

In Figs. 5, 6, 7, 8, 9, 10, and 11, we present a compact subset of the most informative results that support this paper's thesis on the need for purposive control of the sensor parameters. An expanded set of results with a corresponding discussion is provided in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. Notice that in a number of graphs, we have precision rates of 1 and recall rates of 0. This can occur because of the absence of any detectable features in the set of images corresponding to the particular horizon radius and diagonal radius, due to low illumination or camera saturation for example. We now concentrate on Figs. 5, 6, 7, 8, and 9, which deal with the five interest-point detectors. For each pair of identically illuminated precision/recall graphs that are embedded in a given figure, we define their equal error position (EEP) as the diagonal radius (or the horizon-radius/sample-distance value, if it is a different sampling strategy) under which the two graphs intersect each other with the highest precision/recall rate—or at least are the closest to intersecting each other if they do not intersect. The position where the precision and recall rates are the closest to each other parallels, in certain respects, the equal error rate, which is often used in ROC-curve analysis as a metric of algorithm quality. A decreasing equal error rate in ROC-curve analysis typically corresponds to an improving algorithm, while an EEP which occurs at an increasing precision/recall value

corresponds to improving algorithm reliability. Notice that in scenario 1, the target image was acquired under "illuminated" conditions (see the supplementary documentation for a list of the target images used in the scenarios, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91). In each of subfigures Figs. 5a, 5b, 5d, and 5e, we notice that the precision and recall rates achieved at the EEP positions of "normal" and "dark" conditions are about the same. In Fig. 5a, for example, the EEP for the "normal" and "dark" illumination graphs occurs at diagonal radius 9 and 14, respectively. This does not hold for "illuminated" versus "normal" conditions however, since the target image was acquired under "illuminated" conditions, which makes it much easier to achieve high precision/recall rates under identical illumination. Thus, the precision/recall rates at the EEP of "normal" and "dark" conditions are lower than the precision/recall rates at the EEP of "illuminated" conditions. Similar results are evident by inspecting the corresponding sample-distance graphs of Scenario 1 in the supplementary documentation (Figs. $41(a)\mathcal{S}$, $41(d)\mathcal{S}$, $41(j)\mathcal{S}$, $42(a)\mathcal{S}$), which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. This implies that by adjusting the shutter/gain values, the interest-point detection can compensate for the fact that the target image was acquired under different/higher illumination conditions. We can thus achieve similar precision/recall rates at the EEP of "normal" and "dark" illumination. In other words, even though there is significantly lower scene illumination under "dark" conditions, we end up with equally reliable interest-point detection as under "normal" illumination. This points perhaps to the need to selectively choose and discard features from multiple images, acquired under variable shutter/gain values (precipitating the need for more task-directed knowledge). The relevant graphs (see Figs. 5, 6, 7, 8, 9, and the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91) show that adaptively selecting different shutter/gain values for determining if a feature is present or absent can lead to improved results, even under illumination conditions that are different from those under which the target image was acquired. Even in the cases where the precision or recall rate does not surpass 0.5, there is a clear pattern of significant improvement by adapting the sensor parameters. See, for example, how the recall rates under "dark" illumination conditions in Figs. 5a, 5b, 5d, and 5e improve as we increase the diagonal radius. Similar results hold for the corresponding sample-distance and horizon-radius graphs in Figs. $41(a)\mathcal{S}$, $41(b)\mathcal{S}$ as well as other precision/recall graphs. Different shutter/gain values lead to significantly different errors, so, depending on the problem and on whether we are interested in better precision or recall rates, we can adjust the shutter and gain values to the detriment of the other error metric.

Notice in Fig. 5f that the autogain/autoexposure mechanism result—the square (□) symbol in the scatter plot—provides some of the worst results among all other sensor states, demonstrating the need for task-directed knowledge to guide the sensor state. Similar conclusions hold by inspecting most of Figs. 6f, 7f, 8f, and 9f, and the results in Section 2 of the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. Notice in Figs. 6f and 7f, in particular, the precipitous decrease in the precision/recall as we choose sensor states different from the ones under which the target image was acquired—the triangle (△) sign in the scatter plot—demonstrating the instability of this algorithm, at least for the data sets we were using. In Fig. 6f, for example, even though the autogain/autoexposure setting lies in the same bin as the target image's setting, all other sensor states lead to vastly poorer precision/recall rates. From Fig. 5c and Fig. $41(h)\mathcal{S}$, we notice that for scenario 1, the Scale-Saliency algorithm is sensitive to appearance changes, which is reasonable since the algorithm is appearance-based and depends heavily on the histogram of image intensities. For example, in Fig. 5c, around diagonal radius 5, we notice the precipitous decrease in the precision and recall rates of the "illuminated" images, caused by a small change in the shutter/gain value—the perfect precision and recall rate at diagonal radius 5 is achieved because we are matching the target image with itself. By inspecting Figs. 6c, 7c, 8c, and 9c, similar conclusions are reached, demonstrating that for at least some algorithms, the sensor parameters can have a tremendous effect.

Scenarios 2 and 3 (see Fig. 2b and Figs. 6 and 7) investigate the ability to match a target image where all the internal scene structure is visible in the image and the external scene structure is saturated (scenario 2), and of matching a target image where the external scene structure is visible and the internal scene has too low a contrast to be visible (scenario 3). In scenario 2, we notice that the results tend to improve as we increase the diagonal radius since this provides better contrast for the internal scene structure. After a certain limit, however, the internal structure also starts becoming saturated (notice how the precision/recall rates start dropping after diagonal radius 6-9 in Figs. 6a, 6b, 6c, 6d, and 6e), thus leading to a "peak" of optimal sensor settings. Conversely, in scenario 3 (Figs. 7a, 7b, 7c, 7d, and 7e), by decreasing the diagonal radius, the interest points corresponding to the external scene structure are matched better. As can be seen by inspecting Figs. $43\mathcal{S}$-$46\mathcal{S}$ in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91, similar results hold for most interest-point algorithms sampled using the sample-distance strategy. For example, if we increase the sample distances too much in Figs. $45(a)\mathcal{S}$, $45(d)\mathcal{S}$, $45(g)\mathcal{S}$, $45(j)\mathcal{S}$, $46(a)\mathcal{S}$, and $46(d)\mathcal{S}$, we end up with a degraded image quality. Notice that even though the target image in Scenario 2 was acquired under shutter/gain bin $(5,1)$, this bin does not lie in the range $\{(1,1),(2,2),\ldots,(8,8)\}$ of sample-distance bins. As a result, the maximum precision/recall rates of the sample-distance graphs of the five interest-point algorithms in Figs. $43\mathcal{S}$-$44\mathcal{S}$ occur close to sample distance 3 (bin $(3,3)$), which is close to bin $(5,1)$. Notice the significantly different behavior of the error graphs in Figs. 6 and 7, despite the fact that scenarios 2 and 3 deal with the same data set of images (only the target image changes).

The target image of scenario 4 was acquired under "illuminated" conditions, while the scenario 5 target image was acquired under "normal" luminance conditions. We observe an improvement in the recall rates (Figs. 8 and 9) as we increase the diagonal radius in "dark" illumination conditions. Notice, however, the big decrease in many of the precision and recall rates for the diagonal-radius graphs of the "illuminated" and "normal" conditions, as the sensor parameters become increasingly different from the parameters under which the target image was acquired. Similar observations hold for the sample-distance and horizon-radius graphs of scenarios 4 and 5 in Figs. $47\mathcal{S}$-$50\mathcal{S}$. The Scale-Saliency graphs of Figs. 8 and 9 demonstrate the algorithm's sensitivity to changing appearance and some improvement—especially in the recall rates—under "dark" scene conditions.

Notice that in certain figures, there is either a sudden upward spike in the precision (Figs. 6a, 6d, 6e, 7a, 7d, and 7e) or there are large intervals of precision 1.0 (see Figs. 5, 8, and 9). This occurs because the particular diagonal radii contain either saturated or extremely low-contrast images, from which no features are extracted. As a result, the precision defaults to a value of 1.0 (see the mosaic of images in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91). Depending on the target images, their features, and their amount of noise/saturation (e.g., Fig. 6 versus Fig. 7), the precision/recall graphs can vary drastically. For example, in Fig. 7, if the target image had a high shutter/gain setting, we would have a saturated target image and would thus expect a low precision when the target image is matched to an image with good contrast. Furthermore, the recall would increase as more of the target image features become present in the generated image.

The results related to the saliency detection algorithms are shown in Figs. 10 and 11, and the corresponding supplementary documentation figures, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. A general observation is that the AIM algorithm [6] tends to detect more local features, thus indicating that under the AIM method, a greater proportion of each image contains salient features. This is reflected in the greater mean absolute rank errors returned for the AIM algorithm. However, as is seen in Figs. 10 and 11, by changing the diagonal radius, this difference in the absolute errors between the two methods is significantly decreased due to the effect that illumination and shutter/gain control has on the image contrast. Similar observations hold for the sample-distance and horizon-radius sampling strategies in the supplementary documentation, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.91. For example, let us compare Figs. 10b and 10e. We notice that for appropriate shutter/gain values, the mean absolute rank error in Fig. 10e drops from 0.25 to 0.10 for "dark" conditions, making the error comparable to that in Fig. 10b. The relative change in the errors as the illumination and shutter/gain values change provides a measure of the algorithm's sensitivity to these parameters. For illumination conditions that are the same as those corresponding to the target image, we observe a sensitivity of the saliency as we change the shutter/gain values. This becomes more pronounced in Figs. 10b, 10c, and 11b and Figs. 10e, 10f, and 11e, where we are dealing with extreme changes in illumination. Given a constant shutter/gain bin, a change in illumination leads up to an (approximately) eightfold increase in the error. Given constant illumination and for both saliency algorithms, we notice that changes in the diagonal/horizon radii and sample distances also lead to large increases in the average rank errors, demonstrating the sensitivity of both algorithms to the sensor parameters. From Fig. 10c, we observe that in the target image, the exterior scenery is clearly discernible, while the interior is too dark. As a result, as the diagonal radius decreases, the error rates for the "dark" illumination conditions improve. It is also interesting to point out how quickly the "normal"-illumination line's error increases after diagonal radius 2, indicating that the three 8-bit channels forming the RGB images lead to a very limited range of shutter/gain values for which the image regions of interest have sufficient contrast.

In Figs. 10 and 11, we notice the formation of approximately "V"-like or monotonically decreasing/increasing graphs, demonstrating a small range of optimal shutter/gain values for matching the target images, where the optimal camera settings change as the illumination conditions change, also demonstrating the existence of different optimal illumination conditions for different sensor settings. This again demonstrates the inability of a single constant shutter/gain value to optimally discern the most salient image regions under modest changes in the illumination or as the surface reflectance properties change. For scenario 5, this demonstrates the inability of an arbitrary constant shutter/gain value to optimally discern both perpendicular surfaces under arbitrary surface orientation or under modest changes in the illumination direction and intensity since a change in either can easily cause saturation or low-contrast problems in at least one of the two surfaces, precipitating the need for at least two different shutter/gain values so that our camera can optimally match both surfaces with the target image. Since AIM learns its filters during an offline training phase, it is likely that learning those filters using images acquired with the same camera and under the same settings and illumination conditions would lead to improved performance (compared to a nonprobabilistic approach such as [5]) when tested on images generated by the same camera and under similar settings. It is thus reasonable to predict that saliency algorithms which use nonlocal image statistics rather than local image statistics (e.g., [38] or AIM if the neighborhood of features was modified to be the entire image) would perform poorer under an arbitrary camera setting, due to the greater uncertainty in the resulting feature distributions. The histograms in Figs. 11c and 11f show some of the worst results achievable by the two saliency algorithms when using the autogain/autoexposure mechanism. These figures demonstrate that such sensor mechanisms, which rely on the mean image intensity, can offer extremely poor and inconsistent results in degenerate cases (resulting in errors of over 0.12 and 0.4, respectively). This further strengthens the paper's thesis on the need for

more task-directed knowledge during the image formation process. In [28], it is indicated that "all curves are nearly horizontal, showing good robustness to illumination changes." By varying the aforementioned sensor parameters, which were not tested in [28], we reached significantly different conclusions, supporting our hypothesis that the effects of the internal camera parameters are far from trivial and have been overlooked in the literature. This divergence in the results is due to the effects of gain and shutter speed on signal quantization, random noise, image saturation levels, and contrast.

6 Conclusion

An implication of the presented results is that many image data sets used to evaluate vision algorithms are inherently biased since the sensor settings used in acquiring such data sets are typically unknown. This suggests that experimental methodologies used to evaluate many vision algorithms could be significantly improved. Better data sets for evaluating vision algorithms could involve the use of images with higher dynamic ranges. Preferably, rather than representing each scene using a single image, as is usually done, a mosaic of images could be acquired by densely sampling the spectrum of intrinsic sensor parameter settings.

Solving the recognition problem necessitates the construction of non-camera-specific vision modules capable of adapting to changing scene luminance and providing excellent low-level feature detection that far exceeds the state of the art for current systems. Switching between two vision sensors and modest changes in illumination should ideally have a small effect on vision algorithms. This argument becomes stronger as the range of luminance values within which the sensor operates increases and exceeds the sensor's dynamic range. The presented results support a number of hypotheses:

1. Under varying illumination conditions and for both CCD and CMOS sensors, image contrast improves and interest-point detection becomes more reliable by intelligently adjusting the shutter/gain values.
2. For modest changes in the luminance levels of typical indoor scenes, the sensor saturation levels are easily reached with standard cameras and shutter/gain control is a simple way of compensating for this.
3. As we observed with our CMOS sensor, different color filters have different sensitivity levels, leading to the occurrence of localized saturation which is not exclusively based on the scene radiance, but on the spectral density of individual colors present in a particular subset of the scene. Multiple shutter/gain values can capture multiple images such that all features are discernible in at least one of the images.
4. Even under constant illumination, we see that changes in the camera parameters and foreshortening effects due to surface orientation can have a drastic effect on feature detection and saliency.

As indicated previously, while purely software-based approaches to brightness equalization have been proven useful in practice [29], such approaches cannot alleviate catastrophic image degradations that might arise due to pixel saturation, or due to the prominence of dark current noise effects in low-contrast image regions. Even though such degradations are not prominent in offline data sets (e.g., Labeled Faces in the Wild and Caltech-256), which are typically prescreened for good image quality, such degradations can easily arise in an active vision system since the sensor's coordinate frame is free to change in an underconstrained manner. The requirement that active vision platforms operate in such underconstrained environments can compound the problems discussed. As indicated in Section 1, there exist indications that the human visual system uses specialized processes that deal with multiple signal exposures of varying duration. In conjunction with the paper's results, this provides motivation for a form of multiscale processing in the image intensity domain (rather than just the spatial domain, as is usually done in practice) through purposive control of the signal exposure duration. Furthermore, as supported by the results, global brightness equalization techniques can get biased by the background intensity, affecting the ability of a vision algorithm to process the foreground data. Intelligent control of the sensor parameters offers a simple and inexpensive way to compensate for such biases.

As we have argued, a vision system should be insensitive to the camera used. Switching vision sensors should ideally have a small effect on the vision algorithm, as should modest changes in illumination have a limited effect on the algorithm's reliability. This leads to the problem of efficiently selecting the optimal sensor parameter values for a particular problem since simply enumerating all values is not efficient in a real-time vision system that uses standard low dynamic range sensors. Within the context of an active vision system, the choice of a set of optimal sensor parameters which are purposively manipulated subject to a cost constraint gives rise to a number of challenging complexity problems [39]. Within this context, the paper has demonstrated the benefits of a vision system which has direct access to tune the camera's intrinsic parameters. The results in this paper indicate that mosaics of visual maps which sample the permissible range of a sensor's shutter exposure times and gain values can have a nontrivial effect on higher level vision modules since the reliability of recognition algorithms is constrained by the strength of the low-level feature and interest-point detectors used. One could argue that if we use high dynamic range images, many of the problems would be solved since this would decrease the likelihood of encountering saturation/blooming effects in images. However, this would shift the problem to the complex domain of intelligent/adaptive contrast and threshold sensitivity control for the low-level filters used to extract features such as lines and edges. This is in many ways equivalent to the problem of attending to an appropriate visual map in the above described mosaic. This constitutes a strong hint on the need for attentive mechanisms in vision algorithms.

References

[1] R. Bajcsy, "Active Perception versus Passive Perception," *Proc. IEEE Workshop Computer Vision Representation and Control,* 1985.
[2] J. Aloimonos, A. Bandopadhay, and I. Weiss, "Active Vision," *Int'l J. Computer Vision,* vol. 1, pp. 333-356, 1988.

[3] D. Ballard, "Animate Vision," *Artificial Intelligence,* vol. 48, pp. 57-86, 1991.
[4] J.K. Tsotsos, "On the Relative Complexity of Active versus Passive Visual Search," *Int'l J. Computer Vision,* vol. 7, no. 2, pp. 127-141, 1992.
[5] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[6] N.D. Bruce and J.K. Tsotsos, "Saliency, Attention and Visual Search: An Information Theoretic Approach," *J. Vision,* vol. 9, no. 3, pp. 1-24, 2009.
[7] J.K. Tsotsos, *A Computational Perspective on Visual Attention.* MIT Press, 2011.
[8] J. Valeton and D. van Norren, "Light Adaptation of Primate Cones: An Analysis Based on Extracellular Data," *Vision Research,* vol. 23, no. 12, pp. 1539-1547, 1983.
[9] K. Purpura, E. Kaplan, and R. Shapley, "Background Light and the Contrast Gain of Primate P and M Retinal Ganglion Cells," *Proc. Nat'l Academy of Sciences USA,* vol. 85, no. 12, pp. 4534-4537, 1988.
[10] C. Lok, "Seeing without Seeing," *Nature,* vol. 469, pp. 284-285, 2011.
[11] W. Porod, F. Werblin, L. Chua, T. Roska, A. Rodriguez-Vazquez, B. Roska, P. Fay, G. Bernstein, Y. Huang, and A. Csurgay, "Bio-Inspired Nano-Sensor-Enhanced CNN Visual Computer," *Annals New York Academy of Sciences,* vol. 1013, pp. 92-109, 2004.
[12] S. Lim and A.E. Gamal, "Gain Fixed Pattern Noise Correction via Optical Flow," *IEEE Trans. Circuits and Systems,* vol. 51, no. 4, pp. 779-786, Apr. 2004.
[13] D. Litwiller, "CCD versus CMOS: Facts and Fiction," *Photonics Spectra,* vol. 35, no. 1, 2001.
[14] A. Theuwissen, *Solid State Imaging with Charge-Coupled Devices.* Kluwer Academic Press, 1995.
[15] G. Healey and R. Kondepudy, "Radiometric CCD Camera Calibration and Noise Estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 3, pp. 267-276, Mar. 1994.
[16] S.K. Nayar and T. Mitsunaga, "High Dynamic Range Imaging: Spatially Varying Pixel Exposures," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2000.
[17] P.J. Burt and R.J. Kolczynski, "Enhanced Image Capture through Fusion," *Proc. IEEE Int'l Conf. Computer Vision,* 1993.
[18] S. Mann and R. Picard, "On Being Undigital with Digital Cameras: Extending Dynamic Range by Combining Differently Exposed Pictures," *Proc. IS&T's 48th Ann. Conf.,* 1995.
[19] P.E. Debevec and J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," *Proc. ACM SIGGRAPH,* 1997.
[20] T. Mitsunaga and S.K. Nayar, "Radiometric Self Calibration," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1999.
[21] M.A. Robertson, S. Borman, and R.L. Stevenson, "Dynamic Range Improvement through Multiple Exposures," *Proc. IEEE Int'l Conf. Image Processing,* 1999.
[22] M. Aggarwal and N. Ahuja, "High Dynamic Range Panoramic Imaging," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.
[23] S. Nuske, J. Roberts, and G. Wyeth, "Extending the Dynamic Range of Robotic Vision," *Proc. Int'l Conf. Robotics and Automation,* 2006.
[24] S.W. Hasinoff and K.N. Kutulakos, "A Layer-Based Restoration Framework for Variable-Aperture Photography," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.
[25] B.A. Wandell, A.E. Gammal, and B. Girod, "Common Principles of Image Acquisition Systems and Biological Vision," *Proc. IEEE,* vol. 90, no. 1, pp. 5-17, Jan. 2002.
[26] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *Int'l J. Computer Vision,* vol. 37, no. 2, pp. 151-172, 2000.
[27] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 10, pp. 1615-1630, Oct. 2005.
[28] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool, "A Comparison of Affine Region Detectors," *Int'l J. Computer Vision,* vol. 65, pp. 43-72, 2005.
[29] M. Gevrecki and B. Gunturk, "Illumination Robust Interest Point Detection," *Computer Vision and Image Understanding,* vol. 113, no. 4, pp. 565-571, 2009.
[30] F. Faille, "Adapting Interest Point Detection to Illumination Conditions," *Proc. Int'l Conf. Digital Image Computing: Techniques and Applications,* 2003.
[31] R. Unnikrishnan and M. Hebert, "Extracting Scale and Illuminant Invariant Regions through Colour," *Proc. British Machine Vision Conf.,* 2006.
[32] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J.K. Tsotsos, and E. Körner, "Active 3D Object Localization Using a Humanoid Robot," *IEEE Trans. Robotics,* vol. 27, no. 1, pp. 47-64, Feb. 2011.
[33] J. Adams, K. Parulski, and K. Spaulding, "Color Processing in Digital Cameras," *IEEE Micro,* vol. 18, no. 6, pp. 20-30, Nov./Dec. 1998.
[34] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," *Int'l J. Computer Vision,* vol. 60, no. 1, pp. 63-86, 2004.
[35] T. Kadir and M. Brady, "Scale, Saliency and Image Description," *Int'l J. Computer Vision,* vol. 45, no. 2, pp. 83-105, 2001.
[36] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *Proc. British Machine Vision Conf.,* 2002.
[37] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding,* vol. 110, no. 3, pp. 346-359, 2008.
[38] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. Vision,* vol. 8, no. 7, pp. 1-20, 2008.
[39] A. Andreopoulos and J.K. Tsotsos, "A Theory of Active Object Localization," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.

Alexander Andreopoulos received the BSc Honors degree in 2003 in computer science and mathematics, with high distinction, from the University of Toronto. In 2005, he received the MSc degree and in January 2011 received the PhD degree, both in computer science, from York University, Toronto, Canada. His research interests include active vision, visually guided robotics, and medical imaging. He has received the DEC Award for the most outstanding student in computer science to graduate from the University of Toronto, a SONY science scholarship, NSERC PGS-M/PGS-D scholarships, and a Best Paper Award.

John K. Tsotsos received the PhD degree in 1980 from the University of Toronto. He was on the Faculty of Computer Science at the University of Toronto from 1980 to 1999. He then moved to York University, appointed as the director of York's Centre for Vision Research from 2000 to 2006, and is currently a distinguished research professor of vision science in the Department of Computer Science and Engineering. He is an adjunct professor in both ophthalmology and computer science at the University of Toronto. He has published many scientific papers and six conference papers receiving recognition. He currently holds the NSERC Tier I Canada Research Chair in computational vision and is a fellow of the Royal Society of Canada. He has served on the editorial boards of *Image and Vision Computing*, *Computer Vision and Image Understanding*, *Computational Intelligence and Artificial Intelligence and Medicine*, and on many conference committees. He served as the general chair for the IEEE International Conference on Computer Vision 1999. He is a senior member of the IEEE.