Stack-Based Algorithms for HDR Capture and Reconstruction

Orazio Gallo[1]*, Pradeep Sen[2]

Abstract

High-dynamic-range (HDR) images can be created with *standard* camera hardware by capturing and combining multiple pictures, each sampling a different segment of the irradiance distribution of a scene. This seemingly straightforward process involves several important steps, which will be the focus of this chapter. We start by examining the problem of selecting the set of exposures that properly measures the full dynamic range of a particular scene, a process known as metering for HDR. We then describe how to perform radiometric calibration, needed to estimate the incoming irradiance from the low-dynamic-range (LDR) images. After that, we offer an overview of methods to merge multiple LDR images into a single HDR image. Finally, we discuss methods to compensate for camera and scene motion, which would otherwise cause artifacts in the final HDR image.

Keywords
Stack-based HDR, metering, radiometric calibration, deghosting, non-rigid registration

[1] *NVIDIA Research*
[2] *University of California, Santa Barbara*
*__Corresponding author__: ogallo@nvidia.com

Contents

1. Introduction

In this chapter, we examine approaches to capture high-dynamic-range (HDR) images and video using *conventional* digital cameras. This is in contrast with cameras that are specifically designed to capture a larger dynamic range in a single exposure (see Ch. 2). Since standard digital sensors can only capture a small fraction of the incident irradiance (see Ch. 1), approaches for capturing HDR images with a standard sensor must take a stack of N sequential images $Z_1, \ldots, Z_N$ with different exposure settings, and combine their information together as a post-process to reconstruct a high-dynamic-range irradiance image, E. We refer to these as *stack-based* algorithms to HDR capture and reconstruction.

While the focus of this book is HDR video, a thorough discussion on HDR capture for still images is very important. First and foremost, a large portion of the methods proposed for video HDR use a similar strategy, in that they acquire a stream of differently exposed frames. Additionally, many of the topics we cover in this chapter are central also in the case of video HDR, even when the latter is captured with specialized sensors.

From a historical perspective, and although HDR imaging has only recently become widespread, the analog counterpart of today's stack-based approaches was introduced as early as the mid-1800s by French photographer Gustave Le Gray. To expand the limited dynamic range he could capture on film, he literally cut and pasted together multiple films, each measuring a different portion of the dynamic range. The resulting landscapes are simply breathtaking (see Fig. 1). The idea of taking multiple shots to extend the dynamic range a camera can capture re-appeared in the context of digital photography over a hundred years later: two decades ago, Mann (1993)[1] and

[1] The algorithmic details of the work by Mann were published in a later

Figure 1. *The Great Wave, Sète*, Gustave Le Gray, 1857. In one of the earliest examples of stack-based HDR imaging, French photographer Gustave Le Gray extended the dynamic range he could capture by taking two images with different exposure times and combining the two negatives into one. From: www.metmuseum.org

Madden (1993) proposed to combine multiple low-dynamic-range (LDR) pictures into a single HDR image. Since then, stack-based HDR imaging has attracted growing interest by the research community. Today, most consumer cameras, and even some high-end DSLR cameras, offer HDR shooting modes generally based on this strategy.

The layout of this chapter roughly follows the steps involved in stack-based HDR imaging generation. First, we must determine how many pictures to take and what their exposure times should be to adequately capture the dynamic range of a given scene (Sec. 2). Once the images are captured, we must then merge their information together to reconstruct the HDR result (Sec. 3). These approaches work well for static scenes captured with tripod-mounted cameras.

However, if the scene is dynamic or the camera is handheld, the slight differences between exposures in the stack will produce unacceptable ghost-like artifacts in the final result. Since this scenario is very common, a large body of research on stack-based HDR image reconstruction focuses on handling motion (Sec. 4), and two major kinds of methods have been developed: (1) methods that remove ghosting artifacts by rejecting information from images that contain motion (Sec. 4.2), and (2) methods that perform some kind of non-rigid registration to align the input images (Sec. 4.3).

Throughout the chapter, we will use the notation shown in Table 1, often rewriting equations from the different papers to match this notation for consistency and clarity.

paper (Mann and Picard, 1995).

2. Metering for HDR Imaging

When a photographer presses the shutter button to take a picture, digital cameras analyze the scene content to determine the optimal capture parameters. Collectively, the algorithms designed to select these parameters are referred to as the three A's: auto focus (AF), auto white balance (AWB), and auto exposure (AE). The first two, AF and AWB, will not be discussed in this chapter as they do not require modification for the process of HDR capture. Auto exposure, also called metering, is the process of selecting the combination of exposure time, ISO setting, and aperture that optimally capture a specific scene based on some criterion. The heuristics involved in metering algorithms range from considerations about motion blur to signal quality in terms of both signal-to-noise ratio and quantization. Additionally, they may include optimizations better suited to work with the algorithms used by the image signal processor (ISP) in later stages.

In the absence of automatic methods, metering is the photographer's responsibility, requiring both technical and artistic skills. This process is particularly involved in the context of analog photography, because of the non-linearity of the film's response. An example of a beautifully developed theory for metering is the Zone System by Ansel Adams and Fred Archer (Adams, 1948). In a nutshell, Adams and Archer suggest to divide the range of gray levels that can be captured by the camera in eleven segments, also called "zones." Metering then becomes the process of selecting an exposure time that assigns zones to the correct range of irradiance in the scene. For instance, the exposure time should be selected so that the fifth zone captures the values in the middle of the scene's irradiance distribution.

However, when capturing the scene with a single exposure and a digital camera, an optimal AE algorithm is expose-to-the-right (ETTR)[2]; in essence, ETTR selects the longest possible exposure time that does not induce saturation (or blur for hand-held cameras or dynamic scenes). The resulting image minimizes the impact of photon shot noise (PSN). Because of the discrete nature of light, the actual number of photons hitting a pixel in a given time can be modeled by a random Poisson process. Noting that for large numbers a Poisson process can be approximated by a Gaussian process, the number of incoming photons is $n_p \sim \mathcal{N}(\mu, \sigma^2 = \mu)$, where μ is the average number of collected photons. Therefore the signal-to-noise ratio (SNR) increases with a longer exposure time, as the latter increases the average number of collected photons: $\text{SNR} = \mu/\sigma = \sqrt{\mu}$. A shorter exposure time also causes a larger quantization error: the analog signal from the sensor is linearly quantized by the analog-to-digital converter (ADC), which induces a mean square error of $\Delta^2/12$, where Δ, the size of the quantization bins, decreases linearly with the exposure time, see Fig. 2(a). By encouraging a long exposure time, ETTR minimizes the quantization mean square error

[2] The name stems from the fact that longer exposure times push the center of mass of the brightness histogram towards the right, see also Fig. 2(c).

N	number or exposures in the source image stack
$\{Z_i\}_{i=1:N}$	stack of N input LDR source images
$Z_i(p)$	value of pixel p in the i^{th} exposure
$\{t_i\}_{i=1:N}$	exposure times for each of the N source images on the stack
$R_{i,j}$	exposure ratio between exposures j and i (if the exposure time is the only parameter changing, then $R_{i,j} = t_j/t_i$)
E	HDR irradiance image of the scene (W/m^2), which the algorithms in this chapter attempt to reconstruct from an input stack
$\widetilde{E}$	estimated scene irradiance
$\{X_i\}_{i=1:N}$	stack of N exposure images (J/m^2, computed as $X_i = E \cdot t_i$)
$f(\cdot)$	camera response function (CRF), which converts the pixel exposure X to the pixel value Z, i.e., $Z_i(p) = f(X_i(p))$
$g(\cdot)$	inverse camera response function (ICRF), which converts pixel values to pixel exposures, i.e., $X_i(p) = g(Z_i(p))$ (note that $g(\cdot)$ is not an exact inverse of $f(\cdot)$)
$w_i(p)$	weight matrix indicating how well exposed each LDR pixel is, e.g., for merging LDR images to form a final HDR result
Z_{ref}	reference input LDR source, for algorithms that need a reference image from the stack to be specified

Table 1. Notation used in this chapter.

and increases the average number of collected photons, thus increasing the SNR.

Stack-based HDR imaging also requires three A's. Focus and white balance, as mentioned before, need no adaptation. On the contrary, the metering strategy needs to account for the fact that different segments of the scene's irradiance must be sampled by different pictures in the stack. Note that the aperture setting affects the depth-of-field of the image, and thus should typically not change across the stack, leaving only exposure time and ISO sensitivity as the main parameters that can be adjusted[3].

Metering for HDR imaging is more involved than single-picture metering for several reasons. First, a metering algorithm needs to select the actual number of pictures required to completely sample the scene's irradiance, given the sensor's dynamic range. Note that this may be complicated by practical constraints: on the one hand a large number of captures may be impractical for memory and computational requirements. A larger stack also requires longer time to capture, making it more likely for the scene content to change or move. It may also degrade the user experience by forcing the photographer to wait while a long stream of pictures are captured. On the other hand, a sparser sampling of the range, i.e., taking fewer pictures separated by a larger number difference in exposure time, may cause registration and merging issues. Second, and perhaps more obvious, it needs to select multiple exposure times. Several strategies have been proposed to perform metering for HDR imaging. We can roughly classify these methods in three main categories.

Range-agnostic methods use a standard auto-exposure algorithm together with a simple progression of exposures, predefined or user-supplied (such as $0, +1, -1$ EV[4]). This method is the most common strategy found in commercial products since it is extremely efficient from a computational standpoint—no computation is really needed. However, these methods do not provide any guarantee that the scene's irradiance will be fully captured: over- or under-exposed regions may still appear in the final result.

Range-aware methods select the exposure time by looking at some top and bottom percentile of pixel values in the image. By constraining the number of the pixels at both ends of the range, or their maximum and minimum brightness, this class of methods guarantees that the darkest and brightest regions of the scene be covered. The method proposed by Bilcu et al. (2008) is an example of this strategy: after metering for a single image, they select the exposure time of two more images to capture the highlights and the dark areas of the scene. To find the actual extent of the range, they iteratively change the exposure time while streaming the viewfinder frames. Because they always capture three images, the resulting stacks may be larger than necessary and suboptimal in terms of the noise characteristics. Gelfand et al. (2010) use a similar strategy, but allow the number of exposures to vary based on the actual range of a specific scene, although they limit the maximum number of stops between images.

Noise-aware methods model the noise characteristics of the camera system, sometimes even accounting for the scene's radiance distribution. For instance, the noise model proposed by Hasinoff et al. (2010) shows that using a higher ISO setting is beneficial to SNR for a given time budget: the gain boosts the signal before quantization, thus reducing the effect of ADC noise. Based on this observation they propose an optimal, though scene-agnostic, selection of the exposure times for stack-based HDR. A closely related solution is the "HDR+" mode available on the Google NEXUS devices (Levoy, 2014). Rather than selecting different exposures, they take a burst of pictures with the same exposure time, selected to be as short as needed to avoid saturation. Because of the stochastic properties of photon shot noise and ISO noise, merging the different pictures yields a higher SNR in the dark regions of the scene. This strategy, however, does not seem to address the problem of quantization noise, which, as mentioned above,

[3] However, HDR reconstruction methods based on patch-based synthesis (see Sec. 4.3.2) can handle changes in aperture as well, as first shown by Sen et al. (2012).

[4] In general EV, or exposure value, indicates a specific exposure level, corresponding to a set of different combinations of exposure time and aperture setting. It is also used in a relative sense to indicate power-of-2 increments of exposure level, also called "stops." Here we use the latter definition, where 0 EV corresponds to the exposure level obtained with standard AE, and $+1$ EV indicates a picture that captures twice as many photons.

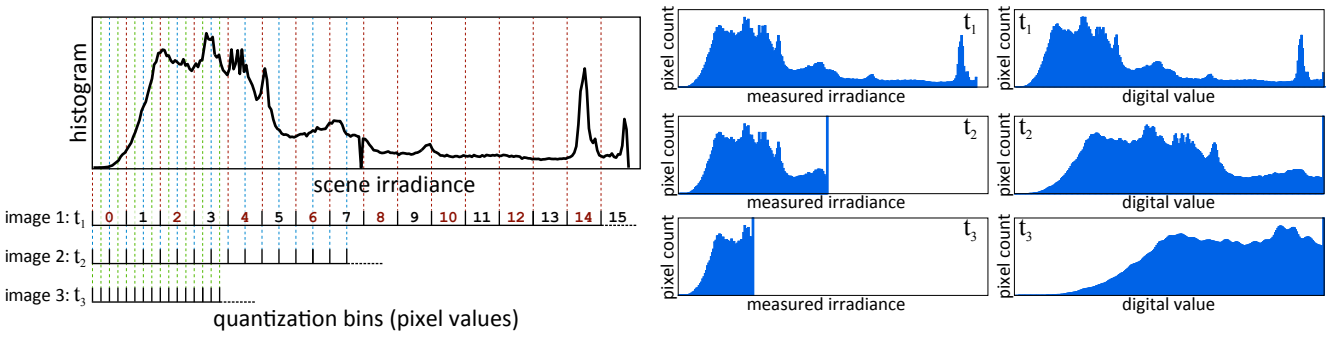

Figure 2. Changing the exposure time affects, among other things, the granularity of the quantization intervals and the shape of the histogram of the captured image. Specifically, the granularity is finer for longer exposure times: the top of the left pane (a) shows the irradiance distribution of an hypothetical scene and the quantization bins of three exposures separated by one stop ($t_3 = 2t_1 = 4t_1$) for a hypothetical 4-bit sensor which measures pixel values from 0 to 15. The last bin is indicated with a dotted line, signifying that it captures all the irradiance values from its left boundary to infinity. This causes saturation in the image, as is visible in column (b), which shows the measured irradiance distribution for these exposure times. In the digital domain, the center of mass of these histograms shifts towards the right as the exposure time is lengthened (c); this is why the process of selecting the longest exposure that does not induce saturation is called expose-to-the-right (ETTR), in this case exposure time t_1. Note that all the graphs are normalized to show details, and the saturated pixel bins are clamped to fit in the graph.

is a particularly pressing issue in the case of short exposures, where the size of the quantization bins is large.

Other noise-aware methods define the optimal sequence of exposures based on the actual distribution of irradiance from the scene. Leveraging on this knowledge, these algorithms can produce a smaller stack with a higher signal-to-noise ratio. Granados et al. (2010) perform an accurate analysis of the different sources of noise in the image formation process of linear cameras and greedily determine the optimal stack—in terms of exposure times and actual number of exposures—given a target SNR. To predict the SNR for a specific scene, they assume an *a priori* knowledge of the histogram of the scene irradiance. This is also similar to the work by Chen and El Gamal (2002). The method by Gallo et al. (2012) extends these methods in two ways. First, it proposes a strategy to compute the actual HDR histogram of the irradiance of a specific scene. Second, it finds the *globally* optimal stack for a generic camera response function, making it possible to use merging strategies designed for non-linear images. Fig. 3 shows a result from the method by Gallo et al. (2012).

When the metering process is completed, the selected images can be sequentially captured. In the following sections we will discuss the processing involved in the combination of the resulting LDR images.

Figure 3. A naïve metering strategy, even one that prevents over- and under-saturation in the final image, may lead to suboptimal results. In this example, Gallo et al. (2012) compare their method with a uniform sampling of the exposure domain that uses enough images to capture the whole dynamic range (i.e., each pixel is correctly exposed at least once in the stack). For the scene shown in the tonemapped image on the left, their method with one that captures the whole dynamic range uniformly (in this case with 5 pictures, each 2 stops apart). Nevertheless, as shown in the insets in the bottom row, their method outperforms the uniform sampling method (insets in the top row) in terms of noise. Image courtesy of Gallo et al. (2012).

3. From low dynamic range to high dynamic range

Once the necessary LDR images are selected and captured, they need to be combined into a single irradiance map. In this entire section, we assume that the camera is steady, and that the captured scene is static during the acquisition of the stack; in other words, a given pixel p in the sensor measures the same irradiance E across the whole stack (these assumptions will be relaxed in Sec. 4). We will further assume that the only parameter that changes across the stack is the exposure time t_{exp} and, possibly, the ISO gain g. Without loss of generality, we subsume both with the variable $t = t_{exp} \cdot g$.

The measured energy density (Joules/m^2), often referred to as *exposure*, can then be modeled as:

$$X_i(p) = E(p) \cdot t_i, \tag{1}$$

where i is the index of the specific LDR image in the stack. Eq. 1 is called *reciprocity assumption* because it states that the exposure $X_i(p)$ can be kept constant when the irradiance changes by a factor k, provided that the exposure time t is also changed by a factor $1/k$. This effect was first reported by Bunsen and Roscoe (1862).

There are two main approaches to the problem of combining information from the LDR images. The first works directly in the pixel's digital value, and never estimates the underlying

irradiance map (these methods are often referred to as "exposure fusion" methods, see Sec. 3.2.3). The second approach works in the irradiance domain and computes an actual HDR map. The latter requires radiometric calibration, the process of determining the mapping between the digital value of a pixel and the corresponding irradiance (up to a scale factor), which we will discuss first. Later, we will describe different strategies to merge the LDR images into the final HDR irradiance map, and conclude this section by discussing exposure fusion techniques. Note that tone-mapping, the process of compressing the dynamic range so that the image can be shown on a regular low-dynamic-range display, will not be covered in this chapter, but is discussed in the second part of this book.

3.1 Radiometric calibration

Eq. 1 describes the relationship between the irradiance $E(p)$ $\left(W/m^2\right)$ at pixel p and the corresponding energy density $X_i(p)$ $\left(J/m^2\right)$. However, we cannot always access X directly. In analog cameras, the film's opacity relates to exposure via a highly non-linear curve called the characteristic (or Hurter–Driffield) curve. In CCD and CMOS cameras, we can often access the RAW values, which are linearly related to the expoure X. However, manufacturers apply carefully-designed transfer functions that both compress the data and enhance the quality of the final image, see Fig. 4. These curves, combined with any other linear and non-linear process applied by the rest of the image processing pipeline (e.g., white balance), can be combined in a single function f, called the camera response function (CRF):

$$Z_i(p) = f(X_i(p)) = f(E(p) \cdot t_i), \tag{2}$$

where $Z_i(p)$ is the digital value associated with pixel p in the i^{th} exposure. If we know the inverse of the CRF, we can estimate the irradiance at the pixel:

$$\widetilde{E}(p) = f^{-1}(Z_i(p))/t_i. \tag{3}$$

Strictly speaking, the CRF f is *not* invertible due to saturation, since all pixels whose irradiance is beyond a certain value are mapped to the highest digital value. Furthermore, the process of quantization maps a finite set of irradiance values to the same bin. Therefore, the function f is not one-to-one and cannot be inverted; after all, if it was invertible, the full irradiance could be recovered from a single image. Despite this observation, we follow conventional notation, and say that radiometric calibration is the process of estimating the inverse of the CRF, $g = f^{-1}$. By "inverse function," we simply mean a look-up table that remaps the non-linear values Z_i to values that are linearly related to the original irradiance $\widetilde{E}$, saturation and quantization aside.

Although different algorithms have been proposed for radiometric calibration, they generally assume that the CRF is fixed; it can then be sampled by taking multiple pictures of the same scene (same irradiance at each pixel) with different exposure times. The assumption that the CRF does not change across the pictures in a stack is paramount if its estimation is to be accurate. However, it is worth pointing out that camera manufacturers spend a great effort to optimize the visual quality of the final image, sometimes adapting the CRF to a specific scene to achieve this (Kim et al., 2012); this poses limits to the overall accuracy of the estimation process.

Camera manufacturers are often reluctant to share information about CRFs, which are their "secret sauce" necessary to deal with the low quality of the pictures that popular, cheap sensors produce. However, a couple of assumptions are fairly safe to make, when performing radiometric calibration. First and foremost, it is commonly assumed that f is monotonic. It is also natural to accept that f be spatially uniform.

The literature on radiometric calibration is vast. However, based on the assumption they make about the shape of the CRF, most approaches can be classified into one of two classes: parametric and non-parametric methods. We describe a few representative methods from these two categories in Sec. 3.1.1 and Sec. 3.1.2. A small number of methods explore the possibility of estimating the CRF from a single image, but because this class is orthogonal to the previous classification, we describe it separately in Sec. 3.1.3.

3.1.1 Parametric methods for radiometric calibration

While the CRF can differ from camera to camera, and even for the same camera but different scenes, it is unlikely that its form be too exotic. Based on this observation, several methods assume a specific functional form for the CRF and attempt to estimate it using different strategies.

Farid (2001) assumes the CRF to be a simple gamma function, $Z = X^\gamma$, in which case the radiometric calibration process is reduced to estimating γ. He then observes that gamma-compressing a signal introduces higher order harmonics in the spectrum of the image. With that, he estimates the gamma as:

$$\arg\min_{\gamma} \sum_{\omega_1,\omega_2 \in [0,2\pi)} |B(\omega_1,\omega_2)|, \tag{4}$$

where B is the bicoherence of the Fourier transform of Z, a measure of the correlation of harmonically related frequencies (Farid, 2001).

In their work on HDR, Mann and Picard (1995) assume the CRF to be of a slightly more general form: $Z = \alpha + \beta X^\gamma$. To estimate α, essentially the black level of the camera, they use a picture captured with the lens cap on, usually referred to as dark frame. Then, assuming the images to be registered, they compute the cross-histogram of the intensity values of a pair of images. For 8-bit images, for example, this is a 256×256 two-dimensional histogram where bin (r,c) contains the number of pixels such that $Z_i(p) = c$ and $Z_j(p) = r$. The parameters (β,γ) can then be found by regression. Mann (2000) later extends this work by considering a number of different analytical forms for the CRF.

Mitsunaga and Nayar (1999) assume a polynomial form of

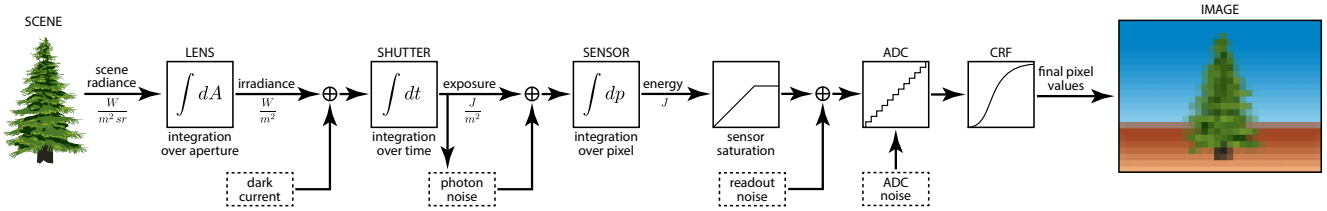

Figure 4. Imaging pipeline of a typical digital camera showing the different sources of noise. The radiant power of the scene is captured by the camera and integrated over the area of the lens aperture, over the time the shutter is open, and over the area of the pixel's footprint to be converted into energy. This signal could be cut off by the saturation of the sensor, which limits the dynamic range of the camera. The result is then quantized by an analog to digital converter (ADC), and the camera response function (CRF) is applied to get the final non-linear digital pixel values. Diagram inspired by Figs. 1 of Debevec and Malik (1997) and of Hasinoff et al. (2010). Tree model courtesy of vecteezy.com.

the inverse of the CRF. Specifically,

$$X = f^{-1}(Z) = \sum_{k=0}^{K} c_k Z^k, \tag{5}$$

where K is the order of the polynomial. Given a rough estimate of the exposure time ratio $R_{i,i+1}$, the exact ratio and the inverse of the CRF can be found as

$$\underset{\{c_n\},R_{i,i+1}}{\arg\min} \sum_i \sum_p \left(\sum_k c_k (Z_i(p))^k - R_{i,i+1} \sum_k c_k (Z_{i+1}(p))^k \right)^2. \tag{6}$$

This is a straightforward LS optimization that can be solved iteratively for $R_{i,i+1}$ and the polynomial coefficients $\{c_n\}$, until convergence.

Grossberg and Nayar (2003a) relax the assumption on the CRF having a specific analytical form. They start by observing that, under the assumption that CRFs are monotonic, the space of the CRFs is a convex space, therefore a linear combination of CRFs is still a CRF. After collecting a large number of real CRFs, $\{f_j\}_{j=1:J}$, they compute the first M eigenvectors of the covariance matrix whose elements are defined as:

$$C_{r,c} = \sum_j (f_j(X_r) - \overline{f}(X_r))(f_j(X_c) - \overline{f}(X_c)), \tag{7}$$

where $\overline{f} = 1/J \sum_{j=1}^{J} f_j$, X are different exposure values, and (r,c) index the bin in the covariance matrix. They show that as few as $M = 3$ eigenvectors can capture 99.5% of the energy, while using $M = 9$ eigenvectors produces curves that are visually indistinguishable from the ground truth. This approach is accurate and extremely efficient, which is the reason why it is used by several methods, as we will see in later sections.

3.1.2 Non-parametric methods for radiometric calibration

Making explicit assumptions on the analytical form of the CRF is not necessary. After all, radiometric calibration can be reduced to computing the look-up table that maps digital values to irradiance (or exposure) estimates, while respecting some properties.

In one of the seminal papers that perhaps most popularized modern HDR stack-based imaging, Debevec and Malik (1997) use a least square formulation to recover the inverse of the camera response function as well as the irradiance values, while imposing smoothness of the recovered response:

$$\underset{E,g}{\arg\min} \sum_{i,p} (E(p) - g(Z_i(p))/t_i) + \lambda \sum_{z=Z_{min}}^{Z_{max}} g''(z), \tag{8}$$

where $g = f^{-1}$. Essentially, the first term imposes that Eq. 3 be satisfied, while the second encourages smoothness of the recovered CRF. Changing λ in Eq. 8 has a strong impact on the overall shape of the estimated CRF; for this reason their method can be seen as a means to convert the images in the stack to the same domain, rather than to perform accurate calibration.

Lee et al. (2013) make the acute observation that the estimates for the exposures X_i from the different images in the stack should be linearly dependent. More formally, the matrix formed with the N exposure images represented as column vectors, $[X_1, X_2, \ldots, X_N]$, should have rank 1. With the matrix $O = [Z_1, Z_2, \ldots, Z_N]$, where the columns are the observed images, the inverse of the CRF g can then be then found as

$$g = \underset{g}{\arg\min}\ \text{rank}(g \otimes O), \tag{9}$$

where the operator $\otimes$ represents an element-wise application of a function. For numerical considerations, rather than minimizing the rank of the matrix, the authors suggest to minimize the ratio of the first two singular values of $g \otimes O$ (this is also called condition number). The authors propose to solve

$$g = \underset{g}{\arg\min}\ \phi(g \otimes O) + \lambda \sum_{Z_i} H\left(-\left.\frac{\partial g}{\partial Z}\right|_{Z_i} \right), \tag{10}$$

where $\phi(\cdot)$ is the first condition number of a matrix, and $H(\cdot)$ is the Heaviside step function, which is 1 if its argument is non-negative, and 0 otherwise. The second term encourages monotonicity: it adds a penalty that is proportional to the number of points where g is decreasing.

Another interesting approach is the work of Kim et al. (2012). Based on their analysis of a large database of JPEG+RAW[5]

[5] RAW images are in first approximation linear with the exposure X, and can therefore be used as ground truth to estimate the irradiance impinging the sensor.

images taken with different cameras and settings, they propose that a single CRF, as traditionally defined and estimated, is not sufficient to explain all of the in-camera processing steps. They observe that cameras perform a gamut mapping that is a function of the scene, or, more specifically, of the "picture style." Therefore, they propose the following image formation model:

$$Z = f(h(T_s T_w X)), \tag{11}$$

where $h(\cdot)$ is the gamut mapping, T_s is the conversion matrix to sRGB, and T_w is the white balance matrix. The key observation is that the gamut mapping, carefully adapted to the scene's type to improve the visual quality of the JPEG image, can be detrimental to the estimation of the inverse of the CRF; however, they hypothesize that the pixels that are affected the most by gamut mapping are the ones that highly saturated (i.e., in the HSV colorspace they have a high S value), and show that removing them from the computation of the CRF allows for a more accurate result.

Virtually all the methods described in this section require that the input stack be perfectly registered; in other words, they assume that the underlying irradiance at pixel p is the same across the stack. Grossberg and Nayar (2003b) propose to overcome this constraint by estimating intensity mapping functions (IMFs). The idea is similar to that of comparagrams (Mann and Picard, 1995): IMFs capture how the brightness values change between two images in the stack. However, rather than building the cross-histogram of two images, which implicitly assumes registration, they look at the cumulative histogram of brightness: $\mathrm{C}(\widetilde{Z}) = \sum_{Z=0}^{\widetilde{Z}} H(Z)$, where H is the histogram of the image. The advantage of the cumulative histogram of an image is that it is robust to small motions in the scene. Given the cumulative histogram of two images C_1 and C_2, the IMF $\tau_{1,2}$ is straightforward to compute:

$$\tau_{1,2}(Z_1) = C_2^{-1}(C_1(Z_1)). \tag{12}$$

More recently, Badki et al. (2015) proposed an algorithm specifically designed to tackle the problem of radiometric calibration for scenes with significant motion. Their approach builds upon both the work by Grossberg and Nayar (2003b) and the method for radiometric calibration by rank-minimization of Lee et al. (2013). First, inspired by the method of Hu et al. (2012), they extend the method of Grossberg and Nayar to large motions by proposing a new RANSAC-based method for computing the IMFs that is robust to such motions. Second, they replace the least-squares optimization for solving for the CRF in Grossberg and Nayar with the rank-minimization scheme of Lee et al. However, the original method by Lee et al. uses artifact-prone, pixel-wise correspondences in their optimization, so Badki et al. reformulate the optimization to replace these correspondences with IMFs. The result is an algorithm that can solve for the CRF even in cases of significant camera and scene motion.

3.1.3 Single-image methods for radiometric calibration

The methods we described thus far assume that multiple images Z_i of the same scene are available, essentially allowing to measure different segments of the CRF. However, radiometric calibration can also be performed on a single image, though with a potentially lower accuracy. Single-image methods can be beneficial in the context of stack-based HDR imaging when the assumption that the CRF is the same across the stack is not valid.

Matsushita and Lin (2007) leverage the fact that the different sources of noise in a camera system can be modeled with symmetric distributions: the cumulative noise distribution should then be symmetric as well, and any deviation from the overall symmetry results only from the non-linearity of the CRF. Therefore, using existing methods to estimate the noise distribution from a single image, they frame the problem of radiometric calibration as:

$$g = \arg\min_{g} \xi_\eta, \tag{13}$$

where ξ_η is a measure of the skewness induced by f to the noise distribution η. Eq. 13 essentially states that g should restore the symmetry of the noise distribution that is expected before the CRF is applied to the image.

Lin et al. (2004) propose another single-image method. They observe that, due to the finite size of the pixels, the irradiance of pixels at the boundary between two uniform regions is a linear combination of the values on either side of the edge. Moreover, moving along a direction orthogonal to the edge in image space should correspond to moving along a line in RGB space only if the CRF is linear. However, if a non-linear CRF is applied to the pixel values, these linear segments become curved. Therefore, they use the method by Grossberg and Nayar (2003a) to parametrize the CRF, and formulate an optimization problem where the solution for g maximizes the linearity in the RGB space of several segments that cross image edges.

Lin et al. (2004) extend this method to a single, grayscale image based on the same idea: the irradiance values of edge pixels should be a linear combination of the irradiances of the regions on either side of the edge. However, since color is not available, they look at the histograms of the intensities in patches lying across edges. These histograms should be roughly uniform, because the point spread function, together with the finite pixel size, turns a sharp edge in a smooth gradient. Once again, they formulate an optimization problem where the inverse CRF is the function that maximizes the uniformity of histograms of different "edge" patches.

The methods described in this section are not intended to serve as a complete survey the space of radiometric calibration methods; rather they are meant to offer insight on the strategies most commonly used. Many other relevant methods have been proposed, such as methods that model and estimate the noise characteristics of the sensor (Tsin et al., 2001; Granados et al., 2010), approaches based on a probabilistic

framework (Xiong et al., 2012), or algorithms working with video sequences, where the CRF may also vary from frame to frame (Grundmann et al., 2013).

3.2 Merging the LDR images into the final HDR result

The process of radiometric calibration we described in Sec. 3.1 essentially maps images captured with different exposure times, and processed with non-linear operators, to the same linear domain. In this domain, an estimate of the irradiance $\widetilde{E}(p)$ can be computed as a linear combination of the values of the corresponding pixels across the stack:

$$\widetilde{E}(p) = \frac{\sum_i w_i(\cdot) \cdot X_i(p)/t_i}{\sum_i w_i(\cdot)}, \tag{14}$$

where we do not make the dependency of the weights $w_i(\cdot)$ explicit because, for different methods, they can be a function of the pixel value $Z_i(p)$ or the exposure $X_i(p)$. Eq. 14 is at the heart of most methods that merge multiple LDR images into a single HDR image, with the difference between the various methods lying in the actual definition of the weights w. These weights can have a big impact on the quality of the final irradiance estimate because the different images in the stack will, in general, be affected by different amounts of quantization noise, photon shot noise, thermal noise, etc.

In their paper, Debevec and Malik (1997) observe that the non-linearity induced by clipping (saturation or underexposure) limits the accuracy with which the true exposure X of these pixels can be recovered. Therefore they empirically define a simple triangle function for w that attenuates the contribution of pixels whose exposure is close to either end of the range:

$$w_{\mathrm{DM}}(Z) = \min(Z - Z_{min}, Z_{max} - Z), \tag{15}$$

where $[Z_{min}, Z_{max}]$ is the range of the pixel values. Mann and Picard (1995) adopt a similar solution, but quantify more accurately the quality of the irradiance estimate offered by each image in the stack. Specifically, they propose to consider the granularity of the quantization induced by the CRF. Where the CRF is steeper, the mapping from X to digital value Z produces a lower quantization error; conversely, where the CRF is more flat, larger ranges of the exposure axis are mapped to the same digital value. Therefore, they define the weights as

$$w_{\mathrm{MP}}(X) = f'(X). \tag{16}$$

Note that Eq. 16 only accounts for quantization noise and ignores the other sources of noise. Mitsunaga and Nayar (1999) extend the work by Mann and Picard by explicitly considering the SNR in the weight computation:

$$w_{\mathrm{MN}}(X) = \mathrm{SNR}_X \cdot w_{\mathrm{MP}}(X) = \frac{X}{\sigma_X} \cdot f'(X) = \frac{g(Z)}{\sigma_X} \cdot \frac{1}{g'(Z)} \approx \frac{g(Z)}{g'(Z)} \tag{17}$$

where, again, $g = f^{-1}$ and, in the last step of the equation, the noise σ_X is assumed to be independent of the level itself, and is therefore dropped. As pointed out by Granados et al. (2010), for linear cameras we can write $w_{\mathrm{MN}}(X) = t$, since the rest of the terms are the same in every LDR.

Robertson et al. (2003) use a weighted least square approach, where the contribution of each pixel to the error is weighted with Mann and Picard's weight, w_{MP}, from Eq. 16:

$$\mathrm{Err} = \sum_{i,p} w_{\mathrm{MP}}(X_i(p)) \left(X_i(p) - t_i \widetilde{E}(p) \right)^2, \tag{18}$$

which can be minimized leading to the weights:

$$w_{\mathrm{R}} = w_{\mathrm{MP}} \cdot t^2. \tag{19}$$

3.2.1 Maximum likelihood estimation

A more theoretically-founded approach is to compute the maximum likelihood (ML) estimate of the irradiance $\widetilde{E}(p)$ (Tsin et al., 2001; Granados et al., 2010). Given two irradiance estimates $E_i(p) = X_i(p)/t_i$ from two different images in the stack, we seek to compute:

$$\widetilde{E} = \arg\max_E \; p(\widetilde{E} \,|\, E_1, E_2), \tag{20}$$

where we omitted the dependence on the pixel p for clarity. We can assume that the observations are drawn from two independent Gaussian distributions $\mathcal{N}(E_i, \sigma_i)$. We can then write:

$$\begin{aligned} p(\widetilde{E} \,|\, E_1, E_2) &= \frac{p(E_1, E_2 \,|\, \widetilde{E}) p(\widetilde{E})}{p(E_1, E_2)} \\ &\propto p(E_1 \,|\, \widetilde{E}) \cdot p(E_2 \,|\, \widetilde{E}), \end{aligned} \tag{21}$$

where we made the common assumption of a uniform prior distribution. Plugging Eq. 21 into Eq. 20 and taking the logarithm, we can write:

$$\begin{aligned} \widetilde{E} &= \arg\max_{\widetilde{E}} \; p(E_1 \,|\, \widetilde{E}) \cdot p(E_2 \,|\, \widetilde{E}) \\ &= \arg\min_{\widetilde{E}} \; \frac{(E_1 - E)^2}{\sigma_1^2} + \frac{(E_2 - E)^2}{\sigma_2^2} \end{aligned} \tag{22}$$

We can find the ML estimate for $\widetilde{E}$ by setting the derivative with respect to $\widetilde{E}$ of Eq. 22 to zero:

$$\begin{aligned} \widetilde{E} &= \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} E_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} E_2 \\ &= \frac{1/\sigma_1^2 E_1 + 1/\sigma_2^2 E_2}{1/\sigma_1^2 + 1/\sigma_2^2}, \end{aligned} \tag{23}$$

from which we can see that $w_{\mathrm{ML}} = 1/\sigma_i^2$. Several methods build on this result, by observing that weights should at least account for the uncertainty of the pixel's value. The first attempt in this direction is the work of Tsin et al. (2001). After modeling white balance as an affine transformation of the exposure X, and calibrating the sensor for photon shot noise and thermal noise, they define the weights as:

$$w_{\mathrm{T}} = \frac{1}{\sigma(Z)}, \tag{24}$$

where $\sigma(Z)$ is the standard deviation of the signal, measured from the residuals of the irradiance estimation. Kirk and Andersen (2006) use the ML weights as well:

$$w_{\mathrm{KA}} = \frac{1}{\sigma(X_i/t_i)} = \frac{t_i^2}{\sigma(X_i)} \approx \frac{t_i^2}{\sigma(Z_i)g'(Z_i)^2}. \tag{25}$$

Arguably, the most complete noise model was proposed by Granados et al. (2010). They too use the maximum likelihood weights, but improve upon previous work by considering both spatial and temporal noise, the latter being also modeled more accurately than in other approaches. Moreover, they pre-calibrate the camera noise parameters to avoid polluting the irradiance estimate with the uncertainty of the noise estimation.

3.2.2 Winner-take-all merging schemes

A few researchers have proposed a different approach to the generation of an HDR map from a stack of LDR images. Their work is based on the observation that the picture with the longest exposure in the stack is also the one with the smallest quantization noise, and the one impacted by the least photon shot noise (see also Sec. 2). Following this logic, in his work preceding the work of Mann and Picard (1995), Madden (1993) suggested to combine the different images in an HDR stack by using the longest, non-saturated exposure available for each pixel p. A similar approach was proposed by Tocci et al. (2011); however, they also suggest to blend the irradiance estimates at the very top and bottom of the useful range of each LDR image to prevent banding artifacts in the transition areas. Additionally, Tocci and colleagues work in Bayer domain to prevent artifacts due to demosaicing when a subset of the color channels saturate, and assess the reliability of a pixel's estimate also based on its neighborhood.

3.2.3 Exposure fusion methods

The approach to HDR we have described so far consists of radiometric calibration, followed by a merging process that produces the final HDR result. To be displayed on regular monitors, the HDR map needs to be tonemapped. An orthogonal approach is that of *fusing* the images directly in the non-linear brightness domain. The most popular method in this category is *exposure fusion* by Mertens et al. (2007). Their simple and effective method side-steps the estimation of the exposure values X_i, and blends the digital values Z_i directly. To reflect the quality of a pixel value, the authors define

$$w_{\mathrm{EF}} = w_s \cdot w_c \cdot w_e, \tag{26}$$

where w_s, the color saturation weight, encourages more vivid colors, w_c, the contrast weight, penalizes low contrast, and w_e, the well-exposedness weight, prefers pixels close to the middle of the range. A naïve application of the method directly to the image may cause visible seams due to abrupt changes in the values of the weights of neighboring pixels. To prevent such artifacts, Mertens et al. decompose the image into a Laplacian pyramid, and combine it with a Gaussian pyramid decomposition of the weight maps to create the final image. Once again, unlike Eq. 14, w_{EF} is used in the weighted average of the digital values Z_i. Merging images directly in the non-linear brightness domain has advantages and disadvantages. In general, it creates natural-looking results, whereas the tonemapping procedure required for the standard HDR pipeline often produces unnaturally contrasted pictures. Moreover, artifacts due to mis-registration are often attenuated by the weighting process. At the same time, it never produces an actual HDR irradiance map, which can be beneficial for computer vision tasks. Finally, when the difference in brightness between the images in the stack is too large, it can introduce artifacts caused by the Gaussian pyramid decomposition of the weights. Several methods build on this idea to increase computational efficiency (Gelfand et al., 2010), to embed deghosting (Zhang and Cham, 2010; An et al., 2011; Gallo et al., 2015), or simply to propose different weights (Shen et al., 2011).

4. Handling artifacts from motion for HDR imaging

The algorithms described in the previous section assume the scene to be static and the camera to be steady. However, when the stack of LDR images is captured in the presence of camera or scene motion, the misalignment between different exposures produces ghost-like artifacts in the final HDR result (see Fig. 5(b)). Since this is a common scenario, addressing motion artifacts is an important problem for practical HDR capture. Indeed, there is a large body of research on the subject, some of which we will survey here. These methods are often known as HDR "deghosting" algorithms, because they deghost (or remove ghosting artifacts from) the final HDR result. Readers seeking detailed explanation of the individual algorithms or thorough comparisons are referred to the original papers cited, as well as survey papers in the field (Srikantha and Sidibe, 2012; Hadziabdic et al., 2013; Tursun et al., 2015).

Before we begin, we note that stack-based approaches to HDR reconstruction cannot always recover the actual HDR image when the scene is dynamic, at least not like an actual HDR camera would. For example, consider the situation shown in Fig. 6, where the volleyball in front of the bright window occupies different positions across the two-image stack. In the long exposure, which has been selected as the reference, the window is almost entirely saturated and offers no useful detail. Ideally, we would recover this information from the short exposure, which properly captures the scene outside the window. Unfortunately, in the second frame, the ball has moved and blocks part of the view of the window, making it impossible to capture the scene behind it. Because this information is not available in any picture of the stack, we cannot reconstruct an HDR image that would exactly reproduce the structure of the scene as it was when the reference image was captured, as shown in Fig. 6(c). However, some of the deghosting algorithms we will discuss are able to reconstruct *plausible* HDR results, even in extreme cases. Furthermore, they offer the only practical way to capture HDR images of dynamic scenes using

(a) Low-dynamic-range inputs **(b)** HDR Result

Figure 5. Example of ghosting with stack-based HDR imaging for dynamic scenes. (a) stack of input images $Z_1, \ldots, Z_5$. Some of the input images were captured while people were in the scene. (b) HDR result from traditional merging (Sec. 3), with objectionable ghosting artifacts. Images courtesy of Gallo et al. (2009).

conventional digital cameras.

Previous deghosting work can be divided into two major categories: (1) *rejection-based* algorithms and (2) *alignment* algorithms. Rejection-based algorithms assume the scene to be mostly static and use a rejection technique to eliminate motion artifacts, while alignment algorithms perform some kind of non-rigid registration between the images so that they can be merged to produce the final HDR result. Each kind of algorithm has advantages and disadvantages, which we will discuss at a high level below. Before we begin discussing the two major kinds of deghosting algorithms, however, we note that either approach can first address artifacts from small camera motions through simple, rigid-alignment approaches as described in the next section.

4.1 Simple rigid-alignment methods

A simple rigid-alignment pre-process (e.g., using a rotation, translation, or homography matrix to align the images) can often eliminate many of the artifacts from small camera motions, making it easier to deghost images that contain mostly static objects. Of course, such rigid registrations do not address the problem of parallax (caused by camera translation), or artifacts caused by highly dynamic scenes. However, they usually work reasonably well when the camera motion is relatively small and the scene does not undergo significant changes.

To our knowledge, the first method which performed a simple rigid-alignment pre-process is the work of Bogoni (2000), who applied a global affine alignment prior to his optical flow alignment (we will discuss this method in more detail in Sec. 4.3.1). Another early method is the work by Ward (2003), which targets artifacts from camera translations. To compare the differently-exposed images, Ward proposes to first convert them into median threshold bitmaps (MTBs), which are binary images with 1's for pixels greater than the images' median. This strategy stems from the observation that MTBs from differently-exposed images resemble each other more closely than when other potential transformations, such as edge operators, are applied to the images.

MTBs can be used to measure the registration quality by simply XORing the pixels of the MTBs to see where they are different. The optimal translation is the one that maximizes the number of 1's in the MTB. To minimize the impact of noise induced by the pixels close to the median threshold, Ward excludes the pixels whose distance from the threshold is within the noise tolerance. This process can be accelerated with a pyramidal approach, where the translational alignment is computed on coarse versions of the images and then refined at higher resolutions. This multi-scale approach also reduces the chance of converging to a local minimum.

Subsequent work by Tomaszewska and Mantiuk (2007) proposed a different method of rigid alignment by using SIFT to extract key points in each image and finding correspondences between them. They then eliminate spurious matches using RANSAC to estimate the homography that can be used to pre-warp the images. These warped images can then be merged using any of the methods described in Sec. 3. In the end, different flavors of methods like these are common pre-processing steps for more advanced algorithms, as we will see in the next section.

4.2 Rejection algorithms for HDR deghosting

Rejection-based algorithms assume minimal scene motion and a static camera so that only few pixels actually exhibit motion. If the camera shakes slightly, a simple rigid registration process, such as those described in the previous section, can be applied as a pre-process to align the images and satisfy this assumption. Since most of the pixels will exhibit no motion under these assumptions, the majority of the final HDR image can be computed with the standard HDR merging process for static scenes described in Sec. 3. To prevent artifacts at the pixels affected by motion, only the images that are deemed to be static at those locations are combined.

The challenge for these rejection algorithms, therefore, is to detect pixels affected by motion and select from the stack the pixels that can be used in the corresponding locations. These algorithms are usually easy to implement and fairly fast, as they only have to detect motion pixels that deviate from the predicted value. Furthermore, because of their design,

(a) Long Exposure (reference) **(b)** Short Exposure **(c)** Actual HDR scene

Figure 6. The problem with stack-based HDR imaging when the scene is dynamic. The region marked in the reference image (a) is occluded in the exposure that captures the highlights (b), making it impossible to reconstruct the actual content of the scene, which is shown in (c).

they are usually successful at completely removing ghosting artifacts, but sometimes have to compromise on the extent of the dynamic range reconstructed in certain regions.

However, rejection algorithms do have serious shortcomings. Perhaps most importantly, these methods cannot handle moving HDR content since they typically discard from the stack any pixels that contain motion. For example, consider a scene with a moving object whose radiance has a dynamic range too high to be captured by a single image (e.g., a moving person who is partly in the shadows and partly under direct sunlight). Rejection-based techniques cannot reconstruct this HDR image correctly, as these methods only merge *corresponding* pixels across the stack of images, rather than compensating for motion (i.e., they do not move content around). In these cases, different portions of the HDR irradiance range may be measured by non-overlapping regions of the images in the stack, and therefore the values from a single pixel across the stack cannot be combined to get a proper HDR result. Therefore, in general, rejection algorithms have not been as effective in reconstructing HDR results from complex dynamic scenes as the registration-based algorithms we will examine later in Sec. 4.3.

Nevertheless, rejection-based techniques are useful to study because the results they produce are generally not affected by motion artifacts. We can classify rejection-based methods into two categories, which we will describe in the subsequent sections: (1) those that do not select a reference image and try to use information from all images equally (often producing an image only from the static parts of the scene), and (2) those that select an image in the stack as the reference (with the goal of producing an HDR result that resembles this image).

4.2.1 Rejection methods without a reference image

Rejection methods that do not define a single reference image are based on the observation that small moving objects tend to affect different regions of the images across the stack. Therefore, if the stack of LDR images is large enough (usually 5 or more images), a pixel p is likely to capture the irradiance from the static parts of the scene in most of the pictures. Methods from this category then propose a model for how pixel p should behave across the stack if it represented a static object, and discard the values $Z_i(p)$ across the stack that do not follow this model, as they are likely to be affected by motion. However, these methods run into the problem that neighboring pixels may come from a different subset of exposures where objects might be in different positions, which would introduce visible discontinuities. To minimize these effects, these algorithms generally identify clusters or groups of pixels that can be drawn coherently from one (or more) of the input LDR images.

One of the first methods to do this was described in Sec. 4.7 of the book by Reinhard et al. (2005). In this method, the CRF is assumed to be known so that the LDR images Z_i can be converted into their corresponding irradiance images E_i. The different images E_i should theoretically be the same, except for noise, saturation, and motion, which may alter some of the pixels' values from image to image. Therefore, the authors propose to compute the *weighted normalized variance* of the values at each pixel p to determine which pixels are affected by motion:

$$\sigma^2(p) = \frac{\sum_{i=0}^{N} w_i(p) E_i(p)^2 / \sum_{i=0}^{N} w_i(p)}{\left(\sum_{i=0}^{N} w_i(p) E_i(p)\right)^2 / \left(\sum_{i=0}^{N} w_i(p)\right)^2} - 1. \quad (27)$$

This equation, explained only verbally by Reinhard et al. (2005) and later presented mathematically by Jacobs et al. (2008), uses weights w_i to exclude over- or under-exposed pixels from the computation as their divergence from the true irradiance may bias the estimate of the variance. Note that unlike traditional variance, the variance in Eq. 27 is normalized to the actual size of the signal.

The key observation is that, when looking across the image stack, pixels that are not affected by motion should have a smaller variance than those that measure irradiance from different objects. Of course, one could set a simple threshold for this variance to distinguish between these two cases. However, this naïve approach has the problem that the image would suffer from discontinuity artifacts when neighboring pixels are selected from different images with different objects.

To avoid this problem, rejection methods that do not define a reference image must group pixels together into larger clusters, where all the pixels in a cluster are drawn coherently from the same image in the stack. In the particular case of

Reinhard et al. (2005), morphological operators like erosion and dilation are used to grow the binary image after thresholding the variance to create larger, contiguous regions that are identified to have motion. To decide which exposure to use in each region, they generate a histogram of irradiance values in each region and find the maximum value that is not in the top 2%, which they consider to be outliers. They then find the longest exposure that still includes this maximum value within its valid range, and interpolate between this exposure and the original HDR result using the per-pixel variance as a mixing coefficient. In this way, pixels with lower variance across the stack will use the original HDR result, while pixels with larger variance will use the single exposure. This algorithm is able to produce deghosted images and, at the same time, ensure that each region is drawn coherently from one exposure.

In another method, Eden et al. (2006) first use a SIFT-based feature registration technique to align the input images in the presence of varying exposure levels. Once the stack is aligned, they map the images to the irradiance domain, where they draw each pixel of the final composite from one of the input images. This is done in two steps. In the first step, they use a subset of the aligned input images to create a reference panorama that covers the full angular extent of all the inputs using graph-cuts (Boykov et al., 2001). However, because of over- or under-exposure this reference image could have areas of missing information, so they introduce detail from images that are better exposed while solving for a smooth transition between regions in a second pass. This problem is minimized via max-flow graph-cut to produce the final result, which can be smoothed out to remove any remaining seams.

The approach of Khan et al. (2006) attempts to compute a ghost-free image through several iterations of kernel density estimation that modify the blending weights w_i of Eq. 14, by assuming that background (static) pixels are the most common. Essentially, they compute the probability that a given pixel is part of the background, and use this weight when blending so pixels from dynamic objects (and not the background) get a smaller weight. To do this, they represent each pixel in the stack of images with a five-dimensional vector $\mathbf{x}_i(p)$, where i is the index of the image in the stack and p is the pixel location. This vector contains the 3 LDR color channels of the pixel value (in Lab space) as well as the coordinates of the pixel on the image.

For a given pixel p, they select all pixels $\mathbf{y}_j(q)$ in its 3×3 neighborhood over all the images in the stack, denoted by $\mathcal{N}(p)$. Note that the pixels at position p across the stack are not included in this neighborhood. They begin by assuming that all $\mathbf{y}_j(q)$ are equally likely to be part of the background. The probability that a pixel p belongs to the background B (given by $P(\mathbf{x}_i(p)\,|\,B)$) can then be calculated using a kernel density estimator:

$$P(\mathbf{x}_i(p)\,|\,B) = \frac{\sum\limits_{j,q\,\in\,\mathcal{N}(p)} w_{j,q} K_H(\mathbf{x}_i(p)-\mathbf{y}_j(q))}{\sum\limits_{j,q\,\in\,\mathcal{N}(p)} w_{j,q}}, \quad (28)$$

where the kernel K_H is a 5-D multivariate Gaussian density function, and the weight $w_{j,q}$ indicates the probability of the pixel belonging to the background. For the first iteration, these weights are initialized to a "hat" function similar in spirit to that of Debevec and Malik (1997). For subsequent interations, the value of the weights can be set to the probability that the pixel belongs to the background, as computed by Eq. 28 in the previous iterations. However, each time the newly computed weights are multiplied by the initial weights from the hat function to continually diminish the probability that pixels that are over- or under-exposed are used in the final estimates. Upon convergence, the weights are plugged into Eq. 14 to merge the LDR images into an HDR result.

Jacobs et al. (2008) extended the deghosting algorithm of Reinhard et al. (2005) in several ways. First, they pre-align the images as in the earlier work of Ward (2003), but in this case iteratively solving for the translation *and* rotation that maximizes the XOR score between the two median threshold bitmaps. In the second stage, they replace the variance metric of Eq. 27 with a local entropy measure that indicates movement in the scene. Specifically, they measure the local entropy at each pixel in the LDR image Z_i by looking at the pixel values z within a 2D window around pixel p:

$$H_i(p) = -\sum_z P(Z=z)\log\left(P(Z=z)\right), \quad (29)$$

where the probability function $P(Z=z)$ is computed from the normalized histogram of the intensity values of pixels within the window. Using these entropies, they compute an uncertainty image U, which is the local weighted entropy difference between the images:

$$U(p) = \sum_{i=1}^{N-1}\sum_{j=0}^{i-1} \frac{v_{ij}}{\sum\limits_{i=1}^{N-1}\sum\limits_{j=0}^{i-1} v_{ij}} \left|H_i(p)-H_j(p)\right|, \quad (30)$$

where $v_{ij} = \min(w_i(p), w_j(p))$, and weights $w_i(p)$ and $w_j(p)$ are computed using Debevec and Malik's triangle function in Eq. 15, with $Z_{min} = 0.05$ and $Z_{max} = 0.95$. The intuition is that static regions would have similar local entropy measures across the LDR images, even if they are near edges, which might increase the variance because of slight camera motions. This method also does not need an *a priori* knowledge of the CRF, as the entropy measurement can be done in the LDR domain. As with previous methods, this uncertainty image is thresholded and the resulting binary image is eroded and dialated to produce contiguous regions that are affected by motion. At this point, each region is filled with values from one of the irradiance images E_i that is not over- or under-exposed in that region, and blended with the original HDR value to avoid artifacts at the borders.

Sidibe et al. (2009) observe that the value of pixel p across the stack should increase with the exposure time, since the camera response curve is monotonically increasing: $Z_i(p) \leq Z_j(p)$ if $t_i < t_j$. Therefore, they propose to identify regions

where this order relation is broken at least once as ghosted regions. Of course, there might be motions that preserve this order which would not be detected. In the ghosted regions, they use the input images that they deem to have captured the background, which is assumed to appear in the majority of images. To do this, they effectively compute the histogram of irradiance values at each ghosted pixel and compute the mode of this distribution, which is the value that appears the most often. The mode is assumed to be the background and the values are merged together (ignoring pixel values close to saturation or zero) to form the final HDR image. In order to have enough samples at each pixel to compute the mode, they require at least 5 images in the stack.

In another approach, Pece and Kautz (2010) first compute median threshold bitmaps for each image in the stack as proposed by Ward (2003) and accumulate these binary maps for each pixel over all the exposures. Values that are neither 0 nor N are considered motion, and the morphological operators of dilation and erosion are applied to this result to generate the final motion map. In the paper, Pece and Kautz show results using exposure fusion (Mertens et al., 2007), where they select the best available exposure for each of the clusters in the motion map to produce their results.

Zhang and Cham (2012) present a technique similar to exposure fusion (Mertens et al., 2007) (see Sec. 3.2.3) because they fuse the images without generating an HDR image first, but use a novel consistency metric that uses the image gradient to detect movement. To begin, they compute the magnitude $M_i(p)$ and direction $\theta_i(p)$ of the gradient around every pixel of each image in the stack. Next, they observe that the magnitude of the gradient can be used to determine saturated or underexposed pixels, as these regions typically have lower gradient magnitude. Therefore, they propose a *visibility* measure that indicates how well exposed and visible a particular pixel is:

$$V_i(p) = \frac{M_i(p)}{\sum_{i=1}^{N} M_i(p) + \varepsilon}, \tag{31}$$

where the ε is a small value (e.g., 10^{-25}) to avoid division by zero. Finally, they observe that the gradient direction can serve as a consistency measure to detect motion across the exposure stack because of its invariant property over different exposures. Therefore, they compute the gradient direction difference of the i^{th} image with respect to the j^{th} image as follows:

$$d_{ij}(p) = \frac{\sum_{k \in \mathcal{N}} \left| \theta_i(p+k) - \theta_j(p+k) \right|}{M^2}, \tag{32}$$

where $\mathcal{N}(p)$ is the set of offsets of the pixels in an $M \times M$ square neighborhood around pixel p. Using this, a consistency score S_i can be computed for every image. This is done by accumulating a Gaussian weight for each pixel based on the difference of its gradient direction across the stack:

$$S_i(p) = \sum_{j=1}^{N} \exp\left(-\frac{d_{ij}(p)^2}{2 \cdot 0.2} \right). \tag{33}$$

Given these scores, a consistency score for each pixel p in the stack image i can then be calculated as:

$$C_i(p) = \frac{S_i(p) \cdot \alpha_i(p)}{\sum_{j=1}^{N} S_j(p) \cdot \alpha_j(p) + \varepsilon}, \tag{34}$$

where $\alpha_i(p)$ is simply a 1 if the pixel is well exposed and 0 if not. Here, we use the term "well exposed" to define a pixel whose value is in the middle of its range, say between 0.1 and 0.9 in a normalized pixel value range. These consistency scores can then use used to compute the final weights for the fusion process (Eq. 26):

$$w_{EF}(p) = \frac{V_i(p) \cdot C_i(p)}{\sum_{j=1}^{N} V_j(p) \cdot C_j(p) + \varepsilon}. \tag{35}$$

The final image can then be fused together without the need of tonemapping, but does not produce a true HDR result.

Granados et al. (2013) propose to use a noise-aware model to determine whether the image stack values for a particular pixel are *consistent*, which means that they measure the same static irradiance. They observe that, for a pixel in a static region, the exposure values across the stack should all be within an error margin based on the noise of the imaging system. Therefore, rather than using an arbitrary threshold to detect motion, they characterize the noise in the imaging system (both shot noise and readout noise) as a Gaussian distribution that enables them to determine the probability that the difference between two pixel values is caused by scene motion or noise. This idea can be extended to the N images in the stack to produce *consistent* subsets, which will not introduce ghosting artifacts when combined together.

Once these consistent subsets have been identified, the next challenge is to ensure that neighboring pixels draw coherently from the subsets to avoid artifacts. To do this, they pose the irradiance reconstruction problem as a labeling problem, solved by minimizing an energy function with two terms. The first, a consistency term, encourages the pixels to be selected from consistent subsets of the image to reduce ghosting. The second is a prior term that penalizes incoherency across neighboring pixels by enforcing that neighboring pixels should draw from the same consistent subset. They solve this labeling problem using the expansion-move graph-cuts algorithm and then merge the consistent sets together at each pixel to produce the final HDR result. However, despite this graph-cut optimization, their method still cannot always guarantee a semantically consistent result, and thus it requires a manual intervention to resolve remaining issues.

Finally, in recent work, Oh et al. (2015) propose a clever rank minimization strategy to solve for the final HDR image. They begin by assuming that there are two kinds of motion between the images in the stack. The first is global motion due to camera movement, which they assume can be modeled with a homography. The second is local motion, which

they want to eliminate, and is caused by the non-rigid movement of objects in the scene. Their key observation is that if global motion is accounted for, the stack of exposure images $X_1, \ldots, X_N$ should be linearly dependent. In other words, barring local motion, saturation, or noise, the globally aligned exposure images would simply be scaled versions of E, i.e., $X_i = E \cdot t_i$. Therefore, they attempt to eliminate motion artifacts by enforcing that the matrix whose columns are the input LDR images should be of rank 1 (i.e., all columns should be linearly dependent).

Oh et al. first account for the global motion by modeling the hypothetical process of capturing "globally aligned" LDR source images, as if the camera was not moving. This can be written as $\widetilde{Z}_i = f(\widetilde{X}_i + \eta_i)$, where $\widetilde{Z}_i$ are the LDR images that would have been taken with a static camera, $\widetilde{X}_i$ is the ideal exposure image that contains only static scene information, and η_i is a "noise" term representing the local motion in the scene. Since we can apply a homography operator $\circ h_i$ to perform global alignment on each of the inputs Z_i, we can write $\widetilde{Z}_i = Z_i \circ h_i$. Once the camera has been calibrated so that its response curve is linear (see Sec. 3.1), the capture process can be modeled as:

$$\widetilde{Z}_i = Z_i \circ h_i = a\widetilde{X}_i + a\eta_i. \tag{36}$$

We can then vectorize the terms in this equation and combine them into matrices using all of the N captured images: $\mathbf{Z} \circ \mathbf{h} = \mathbf{X} + \eta$. Since all the columns of $\mathbf{X}$ are simply scaled versions of the static scene irradiance E, it is a rank-1 matrix. At the same time, η is sparse if we assume that most of the scene is static and only a few areas are affected by motion. Therefore, the problem of removing motion artifacts from the HDR image is equivalent to the problem of solving for a rank-1 matrix $\mathbf{X}$ and sparse matrix η through the following optimization:

$$\begin{aligned} \mathbf{X}^*, \eta^*, \mathbf{h}^* = &\underset{\mathbf{X},\eta,\mathbf{h}}{\arg\min}\ p_2(\mathbf{X}) + \lambda \|\eta\|_1 \\ &\text{subject to } \mathbf{Z} \circ \mathbf{h} = \mathbf{X} + \eta. \end{aligned} \tag{37}$$

Here, $p_2(\mathbf{X}) = \sum_{i=2}^{N} \sigma_i(\mathbf{X})$ is the sum of the singular values from the second to the last[6] which measures the rank of the matrix. The L1 norm $\|\cdot\|_1$ is a measure of sparsity, and weighting coefficient λ balances the contribution of the two terms. This constrained optimization problem can be solved using augmented Lagrange multipliers (Peng et al., 2012), where the problem is divided into three different sub-problems for $\mathbf{X}$, η, and $\mathbf{h}$ and minimized iteratively.

As discussed earlier, rejection-based algorithms have their drawbacks, but this subset of algorithms that do not specify a reference image have other additional problems. For example, they can often produce images that contain duplicate objects or other artifacts, because the semantic meaning of objects is lost when the consistent sets computed in neighboring pixels are not coherent. These artifacts typically require a manual correction. Furthermore, since they typically strive to use only "background" pixels from each image, this type of rejection methods will suppress dynamic objects from the HDR result.

Finally, because these algorithms produce images that do not adhere to a ground truth reference (i.e., an HDR picture taken at a specific moment in time), they cannot be easily extended to the capture of HDR video. The reason for this is two-fold. First, they do not guarantee temporal continuity since each frame is individually computed, and may use a pixel cluster that is not temporally coherent with the neighboring frames. Second, even if temporal coherency could be enforced, the fact that dynamic objects are usually suppressed defeats the purpose of taking a video in the first place.

4.2.2 Reference-based rejection methods

The algorithms in this category select a single image from the stack as the *reference* and use it as the foundation of the final image. In other words, the HDR result will be geometrically consistent with this reference, at least in the parts where it is well exposed. The other images in the stack will be tested against the reference, and pixels deemed to have been affected by motion will be rejected. For regions where all the images in the stack are rejected, the HDR result would be reconstructed using only the reference.

One of the first examples of these algorithms is the work of Grosch, which takes two differently exposed images and first aligns them using a variant of the method proposed by Ward (2003), extended to consider both translation and rotation. He then computes the camera response function on the largely aligned images using the method by Grossberg and Nayar (2003b), and uses the first image (the reference) to predict the estimated values in the second:

$$\widetilde{Z}_2(p) = f\Big(\frac{t_2}{t_1}\, g(Z_1(p))\Big). \tag{38}$$

If the predicted color $\widetilde{Z}_2(p)$ is beyond a threshold from the actual color in the second image (i.e., $|\widetilde{Z}_2(p) - Z_2(p)| > \varepsilon$), the algorithm assumes that $Z_2(p)$ would introduce motion artifacts, and falls back to using only the first image at these locations. This produces an artifact-free result because it largely follows the reference, and has the advantage that it does not need *a priori* knowledge of the camera response function. However, if the scene contains large moving objects, then the radiometric calibration step could fail as well, unless a more robust calibration procedure, such as the algorithm by Badki et al. (2015), is used (see Sec. 3.1.2).

Gallo et al. (2009) propose a similar approach. They first define the reference as the image in the stack with the fewest over- and under-exposed pixels; they then compare the values of the pixels of the different images in the stack against it. They perform the comparison in the log-irradiance domain, where the following relationship holds:

$$\ln(X_{\text{ref}}) = \ln(X_i) + \ln(t_i / t_{\text{ref}}), \tag{39}$$

[6] Assuming that the number of images N is smaller than the number of pixels in each image.

where the dependence on pixel p is omitted for clarity. Pixels whose exposure $X_i(p)$ is farther than a threshold from the value predicted by Eq. 39 belong to moving objects. However, for increased robustness, rather than working directly with pixels, the authors propose to work with patches; a patch from the i^{th} image in the stack is merged with the corresponding patch in the reference if the number of its pixels obeying Eq. 39 is above threshold. The patches are defined on a regular grid; because two neighboring patches in the reference image can be merged with a different subset of patches from the stack, visible seams may exist at the patches' boundaries. To address this issue, the patches are blended with a Poisson solver (Pérez et al., 2003).

Raman and Chaudhuri (2011) extend the work by Gallo et al. (2009) by replacing squared patches with superpixels, which are inherently more edge-aware. However, rather than computing the high-dynamic-range irradiance map, they fuse the images directly using their non-linear digital pixel values. To begin, the authors compute the weighted variance proposed by Reinhard et al. (2005), see Eq. 27, to identify the pixels that may have measured irradiance from moving objects. Then, using only the pixels that are deemed to have captured static objects, they fit fourth-order polynomials to create a set of $N-1$ intensity mapping functions (IMFs) that map the pixel values of each exposure in the stack to the reference.

In the next step, all the images but the reference are segmented into super-pixels with homogeneous color and texture. The idea is to blend the super-pixels that are static with respect to the reference with the well-exposed reference information. To identify a superpixel as static, the authors use the IMF and compare its pixels with those of the superpixel in the reference; to make the process more robust to noise they also threshold the distance of each pixel from the predicted value. If 90% of the pixels are within this threshold, then the super-pixel is considered to be static with respect to the reference. These static super-pixels are then decomposed into 6×6 patches with an overlap of one pixel on each edge. The patches with more than 90% of pixels within the static super-pixel are considered static as well, and their gradients are merged using a Gaussian weighting function based on exposure. Finally, a Poisson solver is used to reconstruct the final color information from the gradients (Pérez et al., 2003).

Wu et al. (2010) propose a set of criteria for detecting moving pixels. First, they use a criterion that ensures that the pixel values are monotonically increasing as we lengthen exposure time, similar to the earlier work of Sidibe et al. (2009). Next, they use a criterion similar to Grosch that compares a pixel's value to that predicted from another exposure after compensating for the CRF and the exposure time ratio. If a pixel violates any of these criteria then it is considered to be affected by motion. The final motion map is generated by using the morphological operators such as opening and closing. Once the pixels affected by motion have been identified, the authors proceed to compute the final HDR image. Specifically, they select image k as a reference and use it to fill in the pixels affected by motion in the neighboring images $k-1$ and $k+1$ with the value predicted by the camera response curve as in Eq. 38. These new images are then used to predict the next images, and so on until the entire stack has been processed. Finally, boundary artifacts near the edges of the regions in the motion map are corrected by convolving the images with a low-pass kernel, and using the result to replace the values calculated originally in these regions. The HDR image is then computed using the standard merging equation (Eq. 14).

In the work by Heo et al. (2010), the images in the stack are first globally aligned to the reference image using a homography estimated with SIFT features using RANSAC. Next, $N-1$ joint histograms are computed between the values in the reference and those in the other images. These histograms are then converted into smooth joint probability distribution functions, pdfs, through a Parzen windowing process using a 5×5 Gaussian filter followed by a normalization to enforce that the subtended area sums to one. Pixels in the other images in the stack with a joint probability less than a fixed threshold are labeled as ghost pixels. This simple thresholding of the joint probability to determine ghost regions can be very noisy, however, so the authors further refine the ghost regions using an energy minimization that enforces smoothness between neighboring pixels, and which is solved using graph-cuts.

The refined ghost regions can be used to compute new joint histograms that are not affected by motion artifacts; therefore, the algorithm iteratively alternates between computing the joint pdfs and detecting the ghost regions. The pixels not affected by motion are then used to compute the CRF with the method of Debevec and Malik (1997). To further reduce artifacts, this CRF is used to refine the radiance values of all the pixels in the other images in order to make their values more consistent with the reference image. Finally, the different exposures are blended together to generate the final HDR result using a weighted filtering step. These weights are computed by applying a bilateral filter (Tomasi and Manduchi, 1998) to all the samples in a patch around a pixel, using a global intensity transfer function to compare the differently exposed pixel values.

Rejection-based methods that use a single reference image generally reduce or completely remove ghosting artifacts from the final HDR image. They do, however, have some of the shortcomings of all rejection-based methods we discussed earlier, such as not being able to handle dynamic HDR content. Furthermore, if the regions where the reference is over- or under-exposed are large, these algorithms could have problems recovering the full dynamic range because of their heavy reliance on the reference, see Fig. 7.

4.3 Non-rigid registration for HDR deghosting

Rather than simply rejecting content that could generate ghosting, one can compensate for motion by means of non-rigid registration. To do this, researchers have proposed two kinds of algorithms: (1) algorithms based on a flavor of optical flow

Z_1 | Z_2 (reference) | Z_2 tonemapped | Reference-based rejection method

Figure 7. Rejection-based HDR reconstruction methods cannot "move" information around the image. Here, the façade of the building is completely saturated in the reference image, as seen in the tonemapped version of Z_2. A reference-based rejection method, such as that shown in the last image, produces a gray halo in the final result because it falls back to the reference when motion is detected. Since the reference is saturated in this region, the measured irradiance is much lower than the actual irradiance measured by the low exposure, resulting in the artifact visible in the right-most image. Images courtesy of Sen et al. (2012).

to align the images, and (2) algorithms based on patch-based synthesis. Note that while non-rigid registration algorithms have the potential to preserve a larger dynamic range from the stack, they tend to introduce objectionable artifacts when the estimation of the displacement between the images fails. This is particularly true for flow-based algorithms, which we will discuss first.

4.3.1 Optical flow and correspondence registration methods

Bogoni (2000) presents perhaps the earliest known method to register a stack of images for HDR reconstruction. First, he applies an affine motion estimation to globally align the images. This process, based on earlier work on registration for image mosaics (Hansen et al., 1994), operates in a multiresolution fashion from coarse to fine, using a Laplacian pyramid scheme. At each iteration, the optical flow field is computed from one image to another using local cross-correlation analysis, and then an affine motion model is fit to the flow field using weighted least-squares regression. The affine transform is then used to warp each image to align it roughly to the reference. At this point, a second step performs unconstrained motion estimation with optical flow between each source image and a pre-defined reference. This resulting field is used to warp the individual sources to compute the final registration with the reference.

Jinno and Okuda (2008) propose to address the problem of ghosting using Markov random fields. After selecting the reference image, they estimate three arrays (the same size as the images) for each of the other images in the stack. The first is a displacement field $\mathbf{d}$, and the second is a binary occlusion field $\mathbf{o}$ that indicates the parts of the reference that are occluded in the second image. This is computed by thresholding the maximum search distance for the displacement field: if a pixel cannot be found in a neighborhood $\mathcal{N}(p)$ around a given pixel that has a luminance within a specific threshold, then pixel p is considered occluded. The third is a saturation field, a binary mask that keeps track of the regions where the second image is over- or under-exposed. Since these arrays are spatially coherent, they can be modeled as Markov random fields (MRFs) and computed using Bayes rule as an estimation problem that finds the most likely fields $\mathbf{d}$, $\mathbf{o}$, and $\mathbf{s}$ given observed images Z_{ref} and Z_i:

$$\begin{aligned}\max_{\mathbf{d},\mathbf{o},\mathbf{s}} P(\mathbf{d},\mathbf{o},\mathbf{s}\,|\,Z_{\text{ref}},Z_i) &= \max_{\mathbf{d},\mathbf{o},\mathbf{s}} \frac{P(Z_{\text{ref}}\,|\,\mathbf{d},\mathbf{o},\mathbf{s},Z_i)P(\mathbf{d},\mathbf{o},\mathbf{s}\,|\,Z_i)}{P(Z_{\text{ref}})} \\ &= \max_{\mathbf{d},\mathbf{o},\mathbf{s}} \frac{P(Z_{\text{ref}}\,|\,\mathbf{d},\mathbf{o},\mathbf{s},Z_i)P(\mathbf{d}\,|\,\mathbf{o},\mathbf{s},Z_i)P(\mathbf{o}\,|\,\mathbf{s},Z_i)P(\mathbf{s}\,|\,Z_i)}{P(Z_{\text{ref}})}.\end{aligned} \tag{40}$$

This problem is analogous to that of finding:

$$\max_{\mathbf{d},\mathbf{o},\mathbf{s}} P(Z_{\text{ref}}\,|\,\mathbf{d},\mathbf{o},\mathbf{s},Z_i)P(\mathbf{d}\,|\,\mathbf{o},\mathbf{s},Z_i)P(\mathbf{o}\,|\,\mathbf{s},Z_i)P(\mathbf{s}\,|\,Z_i), \tag{41}$$

which they approximate by first finding $\mathbf{s}$ through thresholding, and then iteratively solving for $\mathbf{d}$ and $\mathbf{o}$. Once they have these fields, they can use them during the merging stage to produce the final HDR result.

Zimmer et al. (2011) align images in the stack to a specified reference using an energy-based optical flow optimization that is more tolerant to changes in exposure. To achieve this invariance, they define an energy function that leverages the gradient constancy, similar to Brox et al. (2009) and Brox and Malik (2011). Specifically, for each image i in the stack they compute a dense displacement field $\mathbf{u}_i(p) = [u_i(p), v_i(p)]^T$ that specifies an offset at every pixel by minimizing an energy function of the form:

$$E(\mathbf{u}_i(p)) = \sum_{p\in\Omega} D(\mathbf{u}_i(p)) + \lambda S(\nabla\mathbf{u}_i(p)), \tag{42}$$

where $D(\mathbf{u}_i(p))$ is the data term that tries to align the image to the reference, and S is the smoothness term (regularizer) that encourages smooth flow in places where the reference image is unreliable (i.e., over- or under-exposed). Because

the brightness constancy across the stack is violated in this application, they propose that the data term $D(\mathbf{u}_i(p))$ should try to match the gradient of the offset region in image Z_i to that of the reference:

$$D(\mathbf{u}_i(p)) = \Psi\left(\frac{1}{n_x}\left|\frac{\partial}{\partial x}Z_i(p+\mathbf{u}_i(p)) - \frac{\partial}{\partial x}Z_{\text{ref}}(p)\right|^2 + \right. \\ \left. + \frac{1}{n_y}\left|\frac{\partial}{\partial y}Z_i(p+\mathbf{u}_i(p)) - \frac{\partial}{\partial y}Z_{\text{ref}}(p)\right|^2\right), \tag{43}$$

where Ψ is regularized L_1 norm $\Psi(s^2) = \sqrt{s^2 + 0.001^2}$ and n_x and n_y are normalization factors. For the smoothness term $S(\nabla\delta p_i(p))$, they use a regularizer based on Total Variation:

$$S(\nabla\mathbf{u}_i(p)) = \Psi(|\nabla u_i(p)|^2 + |\nabla v_i(p)|^2). \tag{44}$$

The energy equation in Eq. 42 is then optimized using a semi-implicit gradient descent scheme, and the final flows are used to warp each of the input images in the stack, which are then merged together using the method of Robertson et al. (1999).

Later, Hu et al. (2012) propose to use the patch-based, non-rigid dense correspondence (NRDC) method of HaCohen et al. (2011) to compute dense correspondences between the reference image and the other images in the stack, called the *source* images. They then use this correspondence field to warp pixels in the source images to match the appearance of the reference. However, because of occlusions and disocclusions, as well as brightness changes, the correspondences are generally incomplete; this can result in "holes," i.e., regions where the pixels' value is undefined.

To address this problem, Hu et al. first propose a robust strategy to estimate the intensity mapping functions (IMFs) (Grossberg and Nayar, 2003b) using the known pixel correspondences. For each hole in the warped source, they then attempt to paste the pixels from the original source image; however, to compensate for motion, they first apply a local projective transformation to align them to those in the reference. To ensure that the pasted pixels cause no artifacts, the authors take a bounding box larger than the hole to be transformed and pasted. If the pixels within the box, but outside the hole, do not match, the region is considered to be affected by motion. In these cases, the pixels are pasted directly from the reference after their brightness values are appropriately corrected with the estimated IMF.

Another method based on a flavor of optical flow is the approach by Gallo et al. (2015). The method is based on the observation that modern cameras offer fast bursts modes, which make arbitrarily large displacements unlikely. Therefore, instead of computing the optical flow at each pixel, they suggest to compute it only at sparse locations, and then to propagate it to the rest of the pixels. Specifically there are four stages to the algorithm.

The authors first describe a novel method to find and match corners across two images in the stack, one chosen to be the reference and the other being the source image. Their corners are based on the changes of average brightness around different pixels, which can be computed efficiently with integral images. The second stage identifies and removes matches that are either incorrect or belong to structures that move in a highly non-rigid fashion. To achieve this, the authors observe that good matches should be locally consistent with a homography, and propose a modification of the RANSAC algorithm to isolate those that are not. The set of matches that are both correct and rigid offers an estimate of the flow at sparse locations, which can be propagated to the rest of the pixels using the reference image as a guide to an edge-aware diffusion algorithm. With the dense flow, the source image can be warped to be geometrically consistent with the reference. However, to account for possible errors in the flow propagation, the authors modify the algorithm by Mertens et al. (2007), see Sec. 3.2.3, by adding a fourth weight. Specifically, they use the structural similarity index proposed by Wang et al. (2004) to account for the quality of the registration at different locations, and reduce the contribution of regions that are not correctly registered. Gallo et al. report execution times of under 150ms on a pair of 5MP images on desktop, and under 700ms on a commercial tablet. For reference, the methods described in Sec. 4.3.2 are several orders of magnitude slower. As mentioned before, this large speedup is possible thanks to the observation that arbitrarily large displacements are unlikely when the stack is captured in a fast burst.

Compared to rejection-based approaches for HDR reconstruction, these alignment methods, which rely on correspondences between the different images in the stack, have the advantage that they can move content around. This allows them to handle dynamic objects with high-dynamic-range illumination. However, finding reliable correspondences, especially in cases of complex motion and deformation, is quite difficult and can introduce new artifacts. These problems can largely be resolved using the patch-based synthesis methods discussed next.

4.3.2 Patch-based synthesis methods

The most successful kind of HDR deghosting algorithms are perhaps those that align the stack of images together by using *patch-based synthesis* to generate plausible images that are registered to the reference (Sen et al., 2012; Hu et al., 2013; Kalantari et al., 2013). Indeed, a recent state-of-the-art report on HDR deghosting techniques has shown that these algorithms produce the best results for general scenes (Tursun et al., 2015). These methods can also be considered a new kind of algorithm (different from the rejection and registration algorithms we have discussed) because they can solve for the aligned images and the HDR reconstruction simultaneously. Although patch-based synthesis had previously been shown to be very powerful for various computational imaging tasks (such as image hole filling (Wexler et al., 2007), image summarization and editing (Simakov et al., 2008; Barnes et al., 2010), morphing (Shechtman et al., 2010), and finding dense corre-

(a) Low-dynamic-range inputs (b) Aligned intermediate images (c) HDR result

Figure 8. Sample result of the HDR reconstruction algorithm of Sen et al. (2012). (a) Input LDR images (1st, 3rd, and 5th images shown from of a five-image input stack: Z_1, Z_3, Z_5). (b) Corresponding aligned images ($\widetilde{Z}_1, \widetilde{Z}_3, \widetilde{Z}_5$), computed by the algorithm. (c) Tone-mapped HDR result after the reconstruction. Images courtesy of Sen et al. (2012).

spondences between images (HaCohen et al., 2011)), these new works apply it to HDR reconstruction by posing the problem as an energy optimization.

In order to use patch-based synthesis for HDR reconstruction, the two independent methods of Sen et al. (2012) and Hu et al. (2013) make similar observations: after registration, each image Z_i from the stack should look as if it was taken at the same time as the reference Z_{ref}, but should be photometrically consistent with the original Z_i, thereby capturing all of the additional dynamic range information contained in the original image.

Sen et al. (2012) propose to do this using a new optimization equation that codifies the objective of reference-based HDR reconstruction algorithms: (1) to produce an HDR image that resembles the reference in portions where the reference is well exposed, and (2) to leverage well-exposed information from other images in the stack in places where the reference is not. This results in what they call the *HDR image synthesis equation*, which contains two terms:

$$\begin{aligned}\text{Energy}(E) = \sum_{p\in\text{pixels}} \big[\alpha_{\text{ref}}(p)\cdot\big(g(Z_{\text{ref}}(p))/t_{\text{ref}} - E(p)\big)^2 + \\ +(1-\alpha_{\text{ref}}(p))\cdot E_{\text{MBDS}}(H \mid L_1,\ldots,L_N)\big].\end{aligned} \tag{45}$$

The first term states that the ideal HDR image E should be close in an L_2 sense to the LDR reference Z_{ref} mapped to the linear irradiance domain. This should only be done for the pixels where the reference is properly exposed, as given by the $\alpha_{\text{ref}}(p)$ term, which is a trapezoid function in the pixel intensity domain that favors intensities near the middle of the pixel value range.

In the parts where the reference image Z_{ref} is poorly exposed as indicated by the $1-\alpha_{\text{ref}}(p)$ term, the algorithm draws information from the other images in the stack using a novel multi-source bidirectional equation E_{MBDS} that extends the bidirectional similarity metric of Simakov et al. (2008):

$$\text{BDS}(T \mid S) = \frac{1}{|S|}\sum_{P\in S}\min_{Q\in T} d(P,Q) + \frac{1}{|T|}\sum_{Q\in T}\min_{P\in S} d(Q,P). \tag{46}$$

Simakov et al.'s original function takes a pair of images (source S and target T) and ensures that all of the patches (small blocks of pixels) in S can be found in T (first term, called "completeness") and vice versa (second term, called "coherence"). Note that the coherence term ensures that the final target does not contain objectionable artifacts, as these artifacts are not found in the original source.

However, Eq. 46 does not work for HDR reconstruction directly; sometimes content that should be visible in the i^{th} exposure when "aligned" with the reference exposure might be occluded in Z_i and needs to be drawn from a different image. So rather than using a pairwise bidirectional similarity metric, Sen et al. introduce a multisource bidirectional similarity metric E_{MBDS} that draws information from all the images in the stack simultaneously.

To optimize Eq. 45, Sen et al. introduce auxiliary variables $\widetilde{Z}_i$ that represent the different LDR images in the stack after they have been aligned to the reference. This equation can be then solved with an iterative, two-stage algorithm that solves for the $\widetilde{Z}_1,\ldots,\widetilde{Z}_N$ and E simultaneously:

Stage 1: The algorithm first solves for the aligned LDR images $\widetilde{Z}_1,\ldots,\widetilde{Z}_N$ with a bidirectional search-and-vote process (Simakov et al., 2008) accelerated by PatchMatch (Barnes et al., 2009). This process draws information into each of the aligned LDR images from the entire stack, which has been adjusted to match the corresponding exposure level. In order to produce images aligned with the reference, the irradiance image E from the previous iteration (which has been injected with the reference in step 2) is used as the initial target for the search-and-vote process.

Stage 2: Next, the algorithm optimizes for E by merging the aligned images $\widetilde{Z}_1,\ldots,\widetilde{Z}_N$ together using a standard HDR merging process (Sec. 3) and then injects the portions of the reference image where it is well exposed into the result.

Once the new E has been computed, it is used to extract the new image targets for the next iteration and the algorithm goes back to stage 1. These two stages are performed at every iteration of the algorithm until it converges. Furthermore, as is common for patch-based methods like this (e.g., Simakov et al. (2008)), this core algorithm is performed at multiple scales, starting at the coarsest resolution and working to the finest.

Once the algorithm has converged, it returns both the desired HDR image E as well as the "aligned" images at each exposure $\widetilde{Z}_1,\ldots,\widetilde{Z}_N$. A result produced with this algorithm is shown in Fig. 8. This algorithm was later extended by Kalantari et al. (2013) to reconstruct HDR video, as described in Ch. 4.

Hu et al. (2013) propose a different patch-based synthesis algorithm, which, unlike the algorithm by Sen et al., does not require the camera calibration curve to be known *a priori*. Specifically, they calculate the aligned images $\widetilde{Z}_i$ as:

$$\widetilde{Z}_i = \arg\min_{\widetilde{Z}_i,\tau,\mathbf{u}} \Big(C_r(\widetilde{Z}_i, Z_{\text{ref}}, \tau) + C_t(\widetilde{Z}_i, Z_i, \mathbf{u}) \Big), \tag{47}$$

where $\mathbf{u}$ is the displacement field that "warps" image Z_i to match the geometric appearance of the reference, and τ is the IMF between the source image Z_i and the reference Z_{ref}. In Eq. 47, the first term, C_r for "radiance consistency," encourages the aligned image $\widetilde{Z}_i$ to be geometrically consistent with the reference, Z_{ref}:

$$\begin{aligned} C_r(\widetilde{Z}_i, Z_{\text{ref}}, \tau) = \sum_p \Big(& \|\widetilde{Z}_i(p) - \tau(Z_{\text{ref}}(p))\|^2 + \\ & + \alpha \, \|\nabla\widetilde{Z}_i(p) - \nabla\tau(Z_{\text{ref}}(p))\|^2 \Big). \end{aligned} \tag{48}$$

Note that both the images and their gradients are accounted for in Eq. 48. The second term in Eq. 47, C_t, is what Hu et al. call the "texture consistency" term:

$$\begin{aligned} C_t(\widetilde{Z}_i, Z_i, \mathbf{u}) = \frac{1}{k} \sum_p \big(& \|P_{\widetilde{Z}}(p) - P_{Z_i}(p + \mathbf{u}(p))\|^2 + \\ & + \alpha \, \|\nabla(P_{\widetilde{Z}}(p) - \nabla P_{Z_i}(p + \mathbf{u}(p))\|^2 \big), \end{aligned} \tag{49}$$

where k is a normalization factor. The texture consistency term enforces similarity between the patch around pixel p in the warped source, $P_{\widetilde{Z}}(p)$, and the corresponding patch in the source image, $P_{Z_i}(p+\mathbf{u}(p))$. This helps enforce that the synthesized content is plausible and free of artifacts.

Hu et al. tackle this optimization iteratively using a coarse-to-fine approach, which helps in two ways. First, it prevents the optimization from falling in a local minimum. Second, it allows the algorithm to deal with large over- or under-saturated regions: a patch that is entirely saturated at the finest level could include information from neighboring non-saturated pixels at the coarser levels, thus allowing information to propagate inward. To do this, they propose an iterative, three-stage algorithm:

Stage 1: First, they estimate τ using the intensity histograms of the images (Grossberg and Nayar, 2003b) at the coarsest level of the pyramid and initialize $\widetilde{Z}_i = \tau(Z_i)$ for the same level. The displacement $\mathbf{u}$, which only appears in C_t, can then be estimated with PatchMatch (Barnes et al., 2009).

Stage 2: In a second step, the authors propose to refine the estimate of $\widetilde{Z}_i$ by minimizing C_r. However, for the areas where reference image is over- or under-exposed, they average $\tau(Z_{\text{ref}}(p))$ with the corresponding location in the source image $Z_i(p+\mathbf{u}(p))$ with a weight that accounts for how likely the latter is to become over- or under-exposed in the reference image.

Stage 3: In the third and last step, with the new $\widetilde{Z}_i$, they refine the IMF, τ. Moving to the next finer level, τ is left unchanged and u is linearly interpolated. The latent image $\widetilde{Z}_i$, instead, is initialized with a weighted average of $\tau(Z_{\text{ref}})$ and $Z_i(p+u(p))$.

Results from this approach can be seen in Fig. 9.

As discussed earlier, these patch-based synthesis methods have the advantage that they work very well for scenes with complex, arbitrarily large motion where other algorithms would normally fail. However, they are expensive and require considerable time and hardware resources to evaluate: the reference implementations provided by the authors take over a minute for VGA images. Furthermore, although they can produce plausible results, they are only *hallucinating* the final result as compared to the true HDR result that would have been captured by a hypothetical HDR camera.

5. Conclusion

In this chapter, we have examined approaches to capture high-dynamic-range (HDR) images and video by taking a stack of multiple images at different exposure settings. We began by studying algorithms for metering, which set the exposure levels for the different images in the stack. Next, we studied the process of merging the LDR images into a final HDR result, which included a radiometric calibration process (to compute the irradiance images from the original pixel values) and merging schemes (which compute the weights of the different irradiance images to compute the final HDR). Finally, we examined algorithms developed to handle artifacts from motion when capturing stack-based HDR images, which included rejection algorithms and registration algorithms.

6. Acknowledgments

P. Sen was partially funded by NSF grants IIS-1342931 and IIS-1321168. The authors would also like to thank the many researchers within the computational imaging community who published the work described in this chapter. Without their innovations, none of this would be possible.

References

Ansel Adams. The negative: Exposure and development. *Morgan and Lester*, 98, 1948.

Jaehyun An, Sangheon Lee, Junggap Kuk, and Namik Cho. A multi-exposure image fusion algorithm without ghost effect. In *The IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

(a) Low-dynamic-range inputs **(b)** Aligned intermediate results **(c)** HDR Result

Figure 9. Sample result of the HDR reconstruction algorithm of Hu et al. (2012). (a) Input LDR images Z_1, Z_2, Z_3. (b) Corresponding aligned images ($\widetilde{Z}_1, \widetilde{Z}_2, \widetilde{Z}_3$), computed by the algorithm. (c) Tone-mapped HDR result after the reconstruction. Images courtesy of Hu et al. (2013).

Abhishek Badki, Nima K. Kalantari, and Pradeep Sen. Robust radiometric calibration for dynamic scenes in the wild. In *The IEEE International Conference on Computational Photography (ICCP)*, 2015.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 28(3): 24:1–24:11, July 2009.

Connelly Barnes, Eli Shechtman, Dan B. Goldman, and Adam Finkelstein. The generalized PatchMatch correspondence algorithm. In *The European Conference of Computer Vision (ECCV)*, 2010.

Radu C. Bilcu, Adrian Burian, Aleksi Knuutila, and Markku Vehviläinen. High dynamic range imaging on mobile devices. In *The IEEE International Conference of Electronics, Circuits, and Systems (ICECS)*, 2008.

Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *The IEEE International Conference of Pattern Recognition (ICPR)*, 2000.

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(11):1222–1239, November 2001.

Thomas Brox and Jitendra Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(3):500–513, March 2011.

Thomas Brox, Christoph Bregler, and Jitendra Malik. Large displacement optical flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Robert W. Bunsen and Henry E. Roscoe. Photochemical researches–Part V. On the measurement of the chemical action of direct and diffuse sunlight. *Proceedings of the Royal Society of London*, 1862.

Ting Chen and Abbas El Gamal. Optimal scheduling of capture times in a multiple-capture imaging system. In *The International Society for Optics and Electronics (SPIE)*, 2002.

Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH*, 1997.

Ashley Eden, Matthew Uyttendaele, and Richard Szeliski. Seamless image stitching of scenes with large motions and exposure differences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

Hany Farid. Blind inverse gamma correction. *IEEE Transactions of Image Processing (TIP)*, 10(10):1428–1433, October 2001.

Orazio Gallo, Natasha Gelfand, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. In *The IEEE International Conference on Computational Photography (ICCP)*, 2009.

Orazio Gallo, Marius Tico, Roberto Manduchi, Natasha Gelfand, and Kari Pulli. Metering for exposure stacks. In *Eurographics*, volume 31, pages 479–488, 2012.

Orazio Gallo, Alejandro Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. Locally non-rigid registration for mobile HDR photography. *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.

Natasha Gelfand, Andrew Adams, Sung Hee Park, and Kari Pulli. Multi-exposure imaging on mobile devices. In *The ACM Conference on Multimedia (MM)*, 2010.

Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik P. A. Lensch. Optimal HDR reconstruction with linear digital cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. Automatic noise modeling for ghost-free

HDR reconstruction. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 32(6):201:1–201:10, November 2013.

Thorsten Grosch. Fast and robust high dynamic range image generation with camera and object movement. In *The International Symposium on Vision, Modeling and Visualization*.

Michael D. Grossberg and Sree K. Nayar. What is the space of camera response functions? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003a.

Michael D. Grossberg and Sree K. Nayar. Determining the camera response from images: What is knowable? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(11):1455–1467, Nov 2003b.

Matthias Grundmann, Chris McClanahan, Sing Bing Kang, and Irfan Essa. Post-processing approach for radiometric self-calibration of video. In *The IEEE International Conference on Computational Photography (ICCP)*, 2013.

Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 30(4):70:1–70:10, July 2011.

Kanita K. Hadziabdic, Jasminka H. Telalovic, and Rafal Mantiuk. Comparison of deghosting algorithms for multi-exposure high dynamic range imaging. In *The ACM Spring Conference on Computer Graphics*, 2013.

Michael Hansen, P. Anandan, Kristin Dana, G. Van der Wal, and Peter Burt. Real-time scene stabilization and mosaic construction. In *The IEEE Workshop on Applications of Computer Vision*, 1994.

Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Yong. Heo, Kyoung M. Lee, Sang U. Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In *The Asian Conference of Computer Vision (ACCV)*, 2010.

Jun Hu, Orazio Gallo, and Kari Pulli. Exposure stacks of live scenes with hand-held cameras. In *The European Conference of Computer Vision (ECCV)*, 2012.

Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic-range image generation for dynamic scenes. *IEEE Computer Graphics and Applications (CG&A)*, 28(2): 84–93, Apr 2008.

Takao Jinno and Masahiro Okuda. Motion blur free HDR image acquisition using multiple exposures. In *The IEEE International Conference on Image Processing (ICIP)*, 2008.

Nima K. Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B. Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 32(6):202:1–202:8, November 2013.

Erum A. Khan, Ahmet O. Akyüz, and Erik Reinhard. Ghost removal in high-dynamic-range images. In *The IEEE International Conference on Image Processing (ICIP)*, 2006.

Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S. Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(12):2289–2302, December 2012.

Kristian Kirk and Hans J. Andersen. Noise characterization of weighting schemes for combination of multiple exposures. In *The British Machine Vision Conference (BMVC)*, 2006.

Joon-Young Lee, Yasuyuki Matsushita, Boxin Shi, In So Kweon, and Katsushi Ikeuchi. Radiometric calibration by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(1):144–156, January 2013.

M. Levoy. HDR+: Low light and high dynamic range photography in the Google Camera App. `http://googleresearch.blogspot.com/2014/10/hdr-low-light-and-high-dynamic-range.html` (Accessed on Jan. 1, 2015), 2014.

Stephen Lin, Jinwei Gu, Shuntaro Yamazaki, and Heung-Yeung Shum. Radiometric calibration from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

Brian C. Madden. Extended intensity range imaging. Technical report, University of Pennsylvania, 1993.

Steve Mann. Compositing multiple pictures of the same scene: Generalized large-displacement 8-parameter motion. In *The Society for Imaging Science and Technology (IS&T)*, 1993.

Steve Mann. Comparametric equations with practical applications in quantigraphic image processing. *The IEEE Transactions on Image Processing (TIP)*, 9(8):1389–1406, August 2000.

Steve Mann and Rosalind Picard. Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *The Society for Imaging Science and Technology (IS&T)*, 1995.

Yasuyuki Matsushita and Stephen Lin. Radiometric calibration from noise distributions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *The Pacific Conference on Computer Graphics and Applications*, 2007.

Tomoo Mitsunaga and Shree K. Nayar. Radiometric self calibration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999.

Tae-Hyun Oh, Joon-Young Lee, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(6):1219–1232, June 2015.

Fabrizio Pece and Jan Kautz. Bitmap movement detection: HDR for dynamic scenes. In *The Conference on Visual Media Production (CVMP)*, 2010.

Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2233–2246, 2012.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 22(3):313–318, July 2003.

Shanmuganathan Raman and Subhasis Chaudhuri. Reconstruction of high contrast images for dynamic scenes. *The Visual Computer*, 27(12):1099–1114, 2011.

Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. 2005.

Mark A. Robertson, Sean Borman, and Robert L. Stevenson. Dynamic range improvement through multiple exposures. In *The IEEE International Conference on Image Processing (ICIP)*, 1999.

Mark A. Robertson, Sean Borman, and Robert L. Stevenson. Estimation-theoretic approach to dynamic range enhancement using multiple exposures. *Journal of Electronic Imaging*, 12(2):219–228, 2003.

Pradeep Sen, Nima K. Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 31 (6):203:1–203:11, November 2012.

Eli Shechtman, Alex Rav-Acha, Michal Irani, and Steve Seitz. Regenerative morphing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Rui Shen, Irene Cheng, Jianbo Shi, and Anup Basu. Generalized random walks for fusion of multi-exposure images. *IEEE Transactions of Image Processing (TIP)*, 20(12):3634–3646, December 2011.

Désiré Sidibe, William Puech, and Olivier Strauss. Ghost detection and removal in high dynamic range images. In *The European Signal Processing Conference (EUSIPCO)*, 2009.

Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Abhilash Srikantha and Désiré Sidibe. Ghost detection and removal for high dynamic range images: Recent advances. *Signal Processing: Image Communication*, 27(6):650–662, 2012.

Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile HDR video production system. *ACM Trans. Graph.*, 30(4):41:1–41:10, July 2011.

Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *The IEEE International Conference on Computer Vision (ICCV)*, 1998.

Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. In *The International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2007.

Yanghai Tsin, Visvanathan Ramesh, and Takeo Kanade. Statistical calibration of CCD imaging process. In *The IEEE International Conference on Computer Vision (ICCV)*, 2001.

Okan T. Tursun, Ahmet O. Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in HDR deghosting: A survey and evaluation. In *Eurographics STAR Reports*, 2015.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004.

Greg Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of Graphics Tools*, 8(2):17–30, 2003.

Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time completion of video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(3):463–476, March 2007.

Shiqian Wu, Shoulie Xie, Susanto Rahardja, and Zhengguo Li. A robust and fast anti-ghosting algorithm for high dynamic range imaging. In *The IEEE International Conference on Image Processing (ICIP)*, 2010.

Ying Xiong, Kate Saenko, Trevor Darrell, and Todd Zickler. From pixels to physics: Probabilistic color de-rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Wei Zhang and Wai-Kuen Cham. Gradient-directed composition of multi-exposure images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Wei Zhang and Wai-Kuen Cham. Reference-guided exposure fusion in dynamic scenes. *Journal of Visual Communication and Image Representation*, 23(3):467–475, April 2012.

Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *Eurographics*, 2011.