Automatic Noise Modeling for Ghost-free HDR Reconstruction

Miguel Granados Kwang In Kim James Tompkin Christian Theobalt*
MPI für Informatik, Saarbrücken, Germany

Figure 1: *Our method receives a set of images taken with different exposure times (smaller images) and reconstructs a ghost-free high dynamic range image (larger images; tone mapped). The* acrobat *sequence on the left was captured hand-held with in-camera exposure bracketing. To our knowledge, our method is the first in the literature to reconstruct plausible HDR images of both highly dynamic scenes (*left*) and highly cluttered scenes (*right*) with both small and large displacements with little or no manual intervention.*

Abstract

High dynamic range reconstruction of dynamic scenes requires careful handling of dynamic objects to prevent ghosting. However, in a recent review, Srikantha et al. [2012] conclude that "there is no single best method and the selection of an approach depends on the user's goal". We attempt to solve this problem with a novel approach that models the noise distribution of color values. We estimate the likelihood that a pair of colors in different images are observations of the same irradiance, and we use a Markov random field prior to reconstruct irradiance from pixels that are likely to correspond to the same static scene object. Dynamic content is handled by selecting a single low dynamic range source image and hand-held capture is supported through homography-based image alignment. Our noise-based reconstruction method achieves better ghost detection and removal than state-of-the-art methods for cluttered scenes with large object displacements. As such, our method is broadly applicable and helps move the field towards a single method for dynamic scene HDR reconstruction.

CR Categories: I.4.8 [Image Processing And Computer Vision]: Scene Analysis—Photometry, Time-varying imagery;

Keywords: HDR deghosting, camera noise, motion detection

Links: DL PDF

*e-mail:{granados,kkim,jtompkin,theobalt}@mpii.de

1 Introduction

It is difficult to acquire high dynamic range images (HDR) of dynamic scenes without introducing *ghosting*. Even when using modern cameras with automatic exposure bracketing, the *inter-frame* capture time between input images can be long enough to cause significant object displacement between images of dynamic scenes (Fig. 1). Early HDR research implicitly assumed that both the camera pose and the scene remained static during the acquisition of a set of low dynamic range (LDR) images [Burt and Kolczynski 1993; Mann and Picard 1995]. When these techniques average images of dynamic scenes, they introduce ghosting artifacts (Fig. 8, right). Specialized HDR cameras have also been built, but these are expensive and are not widely available [Tocci et al. 2011].

Deghosting has been addressed in the literature through three different strategies: 1) aligning the scene before color averaging, 2) performing joint alignment and reconstruction using one reference image from the LDR set, and 3) detecting regions with moving objects and excluding their images from the average. All of these strategies fail under challenging real-life conditions. After performing an experimental validation of state-of-the-art deghosting methods, Srikantha et al. [2012] conclude that "there is no single best method and the selection of an approach depends on the user's goal".

1) Scene alignment Bogoni [2000], Kang et al. [2003], and Zimmer et al. [2011] perform a dense alignment of the images using optical flow prior to color averaging. Although optical flow methods can correct short displacements caused by camera shake and moving objects, they typically fail to estimate large displacements, and have difficulties with disocclusions occurring in highly cluttered and highly dynamic scenes. Flow estimation is an active area of research and has many limitations, and the success of these deghosting methods depends on the availability of accurate flow fields.

2) Joint alignment and reconstruction Sen et al. [2012] perform simultaneous alignment and HDR reconstruction. Their method defines a reference image to which all other images are patch-wise aligned. Ill-exposed regions in the reference are filled using an adaptation of the bi-directional similarity func-

tion [Simakov et al. 2008] between the remaining input images and the HDR result. Similarly, Hu et al. [2012] find dense and patch-wise correspondences between a reference image and the remaining images, and blend their aligned gradients using Poisson reconstruction for the final result. These methods can enhance the dynamic range of moving objects in cases where the object deformation is sufficiently small that reliable correspondences can be established, and this is an advantage over methods based on motion detection (including ours). However, correspondences might be difficult to establish due to the differences in the noise distribution between images (see Fig. 10). In such cases, the dynamic range of reference image objects cannot be completed. Further, a single reference might not correspond to the desired output, and a better result could be composited using parts from different images.

3) Motion detection Most HDR deghosting methods work by detecting and excluding image regions that could produce ghosting artifacts. In general, these methods assume that the images are already aligned, and rely on an ability to test if the colors observed for the same pixel in different images are *consistent*. Consistency is tested with criteria such as pair-wise irradiance difference [Grosch 2006; Silk and Lang 2012], irradiance difference to a background model [Granados et al. 2008], distance to the intensity mapping function [Gallo et al. 2009; Raman and Chaudhuri 2010], variance of the irradiance estimates [Reinhard et al. 2005; Jacobs et al. 2008], average ratio between images [Tomaszewska and Markowski 2010], probability of the distance to a background model [Khan et al. 2006; Pedone and Heikkilä 2008], correlation with a reference image [Menzel and Guthe 2007], difference of the entropy on local image patches [Jacobs et al. 2008], and difference between gradient orientations [Zhang and Cham 2012]. However, each of these consistency tests requires setting fixed thresholds that are unlikely to generalize well to the noise properties of different cameras and exposure settings.

Color quantization and bin matching techniques [Min et al. 2009; Pece and Kautz 2010], and techniques that test whether intensity increases monotonically with exposure [Sidibé et al. 2009], can be seen as strategies for dealing with noise differences within the input sequence (higher noise in shorter exposures, lower noise in the longer ones). These invariants have high *specificity* but lower *sensitivity* than other methods (Sec. 3).

In Sec. 3, we experimentally show that our method has higher accuracy than the state-of-the-art methods based on motion detection.

Our approach We claim that HDR deghosting can be significantly improved by modeling the noise distribution of the color values measured by the camera. This has been largely neglected in previous work, but provides a simple and principled approach to solving the problem.

Colors are observed at the same pixel location across different exposures in an LDR set. To test whether two colors correspond to the same irradiance (and so correspond to the same object), we must consider their noise distributions. Noise distributions depend on the camera and exposure settings, and can be modeled using Gaussian distributions. Distribution variance is proportional to the light intensity and is inversely proportional to the squared exposure time, and depends on camera parameters such as the gain factor and the readout noise parameters (Sec. 2.1).

Given that the noise depends on the scene irradiance and the camera parameters, no fixed threshold can be set reliably to detect image differences across camera models and scenes. Following this observation, we estimate the camera gain factor to predict the noise distribution of the input images and use this to normalize the color consistency tests (Sec. 2.2). This novel noise modeling approach improves the discriminative power of ghosting detection.

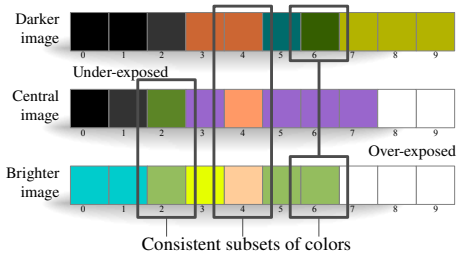


Figure 2: *1-D illustration of HDR reconstruction. An HDR image can be reconstructed by averaging the irradiance estimates derived from the color of corresponding pixel locations in the input images. Ghosting artifacts appear whenever sets of inconsistent colors are included in the average. The problem of HDR deghosting can be defined as selecting consistent subsets of colors for every pixel.*

In general, there can be multiple ghost-free HDR images that are consistent with a set of input images. Among them, we choose the final HDR image such that each pixel color 1) is reconstructed from a consistent set of input images (a single one for dynamic objects), 2) has high signal-to-noise (SNR) ratio, and 3) is spatially compatible with its neighbors in other source images (Sec. 2.3).

In summary, to our knowledge our algorithm is the first HDR reconstruction method to handle scenes with strong clutter and dynamics without introducing ghosting artifacts. This is demonstrated on very challenging scenes including crowded places with small and large object displacements and low-light shots. All these scenes are computed with fixed parameters. Furthermore, our algorithm performs on par with state-of-the-art methods for image sets with only small object displacements. As such, our method is broadly applicable and helps move the field towards a single method for dynamic scene HDR reconstruction. The contributions of our paper are:

1. A novel and simple method for estimating the camera gain factor from arbitrary images. This enables the automatic prediction of the image noise range.

2. To our knowledge, the first HDR imaging method to fully automatically take advantage of a camera noise model for performing reliable ghost-free reconstruction across different cameras and scenes.

2 HDR deghosting method

Our algorithm input is a set of images taken with a static or hand-held camera at different exposure times, where pixel values in the images are the *raw output* of the camera, i.e., before any of the camera's internal processing. If captured hand-held, we robustly register the images using a global homography computed with RANSAC [Fischler and Bolles 1981] from sparse SURF keypoint matches [Bay et al. 2008]. With an aligned image set, our method estimates an irradiance image where each pixel is constructed as a weighted average of colors of the corresponding pixels across the input images. Ghosting artifacts would be generated by averaging a set of pixels which includes an *inconsistent* subset. Our algorithm identifies a *consistent* subset of images per pixel location and reconstructs the final irradiance value as an average of consistent pixel colors (Fig. 2). This avoids having to select a reference image [Sen et al. 2012], or having to build a background model [Khan et al. 2006], which requires that the background be more likely to be observed at every image location — this is not necessarily true for cluttered scenes. To begin, we discuss our noise model and our automatic camera calibration procedure.

2.1 Image noise estimation

Even when assuming a static scene and constant camera parameters, image noise varies by exposure time. The two main temporal noise sources are known as *shot noise* and *readout noise*. Shot noise is introduced by the process of light emission, which follows a Poisson distribution where the variance is equal to the mean. Readout noise comprises several other signal-independent sources affecting the acquisition process of digital cameras (including quantization noise), and it is modeled well by a Gaussian distribution with zero mean.

In CCD/CMOS sensors, the number of photon-electrons collected by the camera at every pixel is linearly proportional to the incident irradiance. This derives from the properties of the photo-electric effect on silicon-based sensors for visible wavelengths [Janesick 2001]. The raw camera output is also linearly proportional to the number of collected photon-electrons. This relation is known as the *camera response* function f. The slope of this function corresponds to the camera's *gain* factor g. This factor is proportional to the ISO setting (e.g., the gain at ISO400 is four times the gain at ISO100).

Since the response function f is linear for raw output, it is possible to recover the number of photon-electrons collected by the camera to approximate the probability distribution of each pixel measurement [Granados et al. 2010]. For a non-saturated raw camera output $v_i(p)$ on image i and pixel p, the inverse of the response function, i.e., the amount of collected photon-electrons, is estimated by

$$\tilde{f}^{-1}(v_i(p)) = \frac{v_i(p) - b_i(p)}{g} = t_i x(p), \qquad (1)$$

where the *dark frame* b_i is an image acquired with same exposure time as v_i but without incoming light (e.g., with the lens cap on). The product $t_i x(p)$ between the image's exposure time t_i and the incident irradiance $x(p)$ is known as the *exposure*, which is proportional to the number of photon-electrons collected by the camera.

Dark frames measure the *dark current*, i.e., the pixel intensities induced by thermal energy and not by light). We assume that the dark current is negligible or, equivalently, that dark frame subtraction is performed in-camera. Thus, in Eq. (1), we replace the dark frame $b_i(p)$ with the black level L_0 of the camera, and omit the contribution of dark current to shot noise in Eq. (2) below.

The exposure $t_i x(p)$ follows a Poisson distribution, and the uncertainty in its measurement corresponds to the shot noise. We approximate this distribution using a Gaussian [Hubbard 1970] to model the variance of the irradiance estimate $x(p)$. From Eq. (1), the variance of $x(p)$ in image i can be derived as

$$\sigma^2_{x_i(p)} = \frac{g^2 t_i x(p) + \sigma_R^2}{g^2 t_i^2}, \qquad (2)$$

where σ_R^2 is the variance of the readout noise, which is also modeled using a Gaussian. To evaluate Eq. (2), we need to estimate the parameters g, L_0, σ_R^2, and t_i. The exposure time t_i can be obtained directly from the digital image file; next, we explain the estimation of the remaining parameters.

Readout noise The black level L_0, and the readout variance σ_R^2 are calibrated using the method described in [Janesick 2001; Granados et al. 2010]. This method estimates L_0 and σ_R^2 as the mean and variance, respectively, of the pixel values of a *black frame*, i.e., an image taken with no incident light and no integration time (practically, a very short exposure time). In principle, this data could be provided for every camera model by the manufacturer.

Camera gain If not provided by the manufacturer, the camera gain g can be calibrated. Janesick [2001] and Granados et al. [2010] suggested to calibrate it using *flat fields*, i.e., images exposed with a constant illumination at every pixel, such that every pixel color can be assumed to be a sample of the same random variable. Under this assumption, the mean and variance of the observed color can be approximated using the spatial mean and variance of a flat field. Using this approximation, the gain can be derived by exploiting the equivalence between the expected value and the variance of the exposure. This *flat-field calibration* is the best method available, and it can be applied to any digital camera. However, in practice,

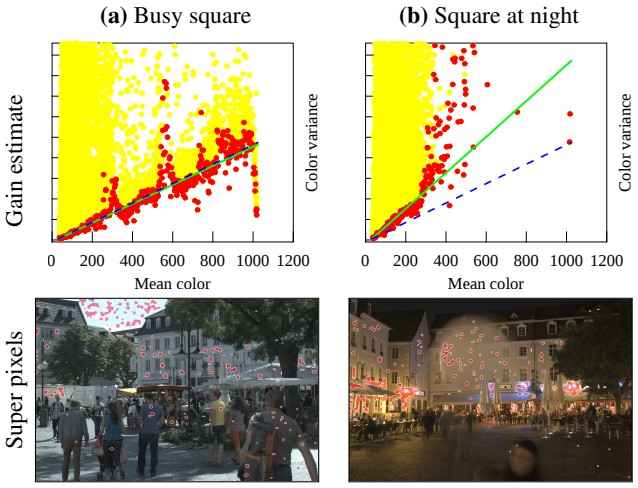


Figure 3: *Image-based gain calibration. Red dots (top and bottom) correspond to low-variance super pixels used for calibration. Yellow dots represent the remaining super pixels. Green lines show the predicted noise by image-based calibration, blue dashed lines show the prediction by flat-field calibration. Our deghosting method is robust to calibration errors, so even in cases where the gain is over-estimated (b), the final images are still free of ghosting artifacts (see Fig. 10). See supplementary material for additional results.*

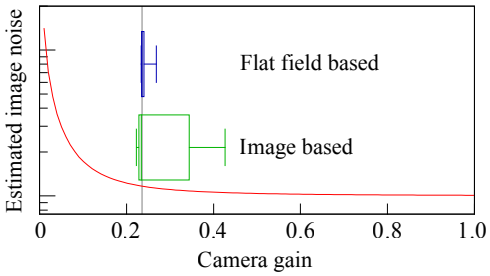


Figure 4: *Confidence of camera gain estimation. The box plots show the 1st, 25th, 50th, 75th and 99th percentiles of the distribution of gain factors obtained from flat-field calibration (a 36 samples of flat field images), and the distribution of factors obtained from image-based calibration (a sample of seven images, each from a different scene; two shown in Fig. 3). The gray line denotes the true gain of the camera. The expected gain for both methods is very close, but the variance of image-based calibration is higher. Despite this, our gain estimate can still be used to reconstruct ghost-free HDR images (see Fig. 5). The red curve illustrates the dependency between the gain factor and the image variance prediction. In general, when the camera gain is over-estimated, the predicted noise for the input images is under-estimated. This makes ghost detection stricter, thus reducing the SNR of the final HDR image because smaller consistent subsets will be found. As such, no ghosting artifacts are introduced by this error (see Fig. 5).*

(a) $g - 6\sigma_g$, SNR=28.5dB **(b)** $g - 3\sigma_g$, SNR=27.2dB

(c) $g + 0\sigma_g$, SNR=26.5dB **(d)** $g + 24\sigma_g$, SNR=21.1dB

Figure 5: *Sensitivity of our deghosting method to gain calibration accuracy. Here, g, σ_g denote the mean and standard deviation of the flat-field gain calibration. Our method is robust to slight under-estimation (b) and large over-estimation (d) of the camera gain: When it is under-estimated (which occurs seldom, see Fig. 4), ghosting artifacts can appear (a, magenta arrow). Conversely, when the gain is over-estimated, it leads to low SNR (d), but it does not introduce ghosting artifacts. See supplementary material for additional tests at intermediate error levels.*

this requires additional flat field images, which may be cumbersome for inexperienced users to acquire.

Therefore, we propose an alternative *image-based calibration* that does not require flat fields at all and works directly from the input image set of the scene. The idea is to use regions of constant illumination in the input images as proxies for the flat fields. We divide an input image (e.g., the central exposure) into *super pixels* [Veksler et al. 2010]; which have a predefined patch size and follow image edges. From the mean-variance scatter plot of the super pixel colors (Fig. 3–top), we select the minimum variance for each digital value, and use RANSAC [Fischler and Bolles 1981] to fit a line that passes through (L_0, σ_R^2), i.e., through the expected variance at the black level. The idea of using super pixels to estimate the lower bound of image variance was first proposed in [Liu et al. 2008] for image denoising. Our method uses a simpler noise model tailored to raw camera output, and a simpler inference method (i.e., RANSAC instead of Bayesian inference) that is very straightforward to implement. Figure 3 illustrates this process: The top row shows the mean and variance color value of each super pixel (yellow and red dots). Among them, we select the super pixels with minimum variance as proxies for flat fields (shown in red). This selection is justified as only shot noise and readout noise contribute to the variance of image regions with constant illumination and, therefore, these noise sources determine the lower bound of the color variance.

Figure 4 compares the performance of each gain calibration method: Our image-based calibration is sufficiently accurate and is comparable with flat-field calibration in terms of predicted image noise. Importantly, since a wide range of scenes contain locally flat regions, this calibration approach allows our deghosting algorithm to be directly applied without requiring users to capture flat field images. However, its accuracy is content dependent; Fig. 3b shows an example image from which the gain could not be estimated precisely: Since flat regions in the image cover a limited color band, the slope estimation is misled (Fig. 3b–top). That said, ghosting artifacts typically only appear when the variance within super pixels (and thus the gain) is underestimated (e.g., $6\sigma_g$ below the true gain, see Fig. 5), which is a highly unlikely scenario in practice.

2.2 Consistency test

Next, we introduce consistency measures for pairs of pixels and a group of pixels, respectively: two pixels at corresponding locations in different images are consistent if the corresponding color difference follows the predicted color difference distribution, and a group of pixels is self-consistent if all the pixels are pair-wise consistent.

Consistency test for pairs of images Let us assume we are given two irradiance observations $x_i^k(p)$, $x_j^k(p)$ at pixel p and color channel k, which are derived from the pixel colors $v_i^k(p)$, $v_j^k(p)$ on images i, j, respectively, using the inverse of the camera response function (Eq. (1)). Detecting ghosting artifacts requires testing whether these irradiance observations are *consistent*, i.e., if they correspond to measurements of the same incident light. Existing algorithms solve this problem by relying on pre-determined thresholds, which are difficult to set. This requirement can be avoided by exploiting the noise model discussed in Sec. 2.1.

Our approach is to estimate the probability distribution of a difference function $d_{ij}^k(p) = x_i^k(p) - x_j^k(p)$; since $x_i^k(p)$ and $x_j^k(p)$ follow Gaussian distributions, $d_{ij}^k(p)$ has the same distribution type which, for consistent pairs, has zero mean and has variance:

$$\mathbb{V}\mathrm{ar}\, d_{ij}^k(p) = \mathbb{V}\mathrm{ar}\, x_i^k(p) + \mathbb{V}\mathrm{ar}\, x_j^k(p), \tag{3}$$

where $\mathbb{V}\mathrm{ar}\, x_i^k(p)$ and $\mathbb{V}\mathrm{ar}\, x_i^k(p)$ are obtained from Eq. (2). Given $\mathbb{V}\mathrm{ar}\, d_{ij}^k(p)$, we can estimate the probability that observations at pixel p on images i, j are consistent by comparing the corresponding irradiance *differences* with the expected noise distribution on every color channel:

$$\Pr\left(p\,|\{v_i, v_j\}\right) = \min_{k\in\mathbf{C}} \Pr\left(-\frac{|d_{ij}^k(p)|}{\mathbb{S}\mathrm{td}\, d_{ij}^k(p)} \le \mathcal{N} \le \frac{|d_{ij}^k(p)|}{\mathbb{S}\mathrm{td}\, d_{ij}^k(p)}\right), \tag{4}$$

where $\mathbf{C} = \{R, G, B\}$, $\mathcal{N}$ is the standard Gaussian random variable with mean zero and variance one. In practice, the estimate $\Pr\left(p\,|\{v_i, v_j\}\right)$ can be noisy (e.g., when the image is taken under low-light or when the camera has a high readout noise). For this reason, prior to estimating the probabilities, we smooth the difference image $d_{ij}^k(p)$ using bilateral filtering [Tomasi and Manduchi 1998]. We refer to this step as noise-adaptive difference filtering (DF). We use a distance kernel of large bandwidth, and a range kernel with variable bandwidth $\sigma_r = 2\,\mathbb{S}\mathrm{td}\, d_{ij}^k(p)$ that is proportional to the predicted image noise. This filtering introduces dependencies between the distributions of neighboring pixels. However, this dependency occurs mostly between pixels that have already similar distributions. Given this similarity, the net effect of the filtering is a attenuation of the tails of the difference distribution. This allows us to obtain a higher detection sensitivity for the same specificity level (see Sec. 3 for experimental validation).

Since the noise variance $\mathbb{V}\mathrm{ar}\, x_i(p)$ is different at every pixel and image in the sequence, the variance of the difference function $\mathbb{V}\mathrm{ar}\, d_{ij}^k(p)$ also varies for every pixel and image pair. This observation is integral to our technique: As other reconstruction and deghosting methods do not automatically model noise, they are not likely to generalize well to the noise properties of different cameras and exposure settings.

Consistency test for sets of images Let $\mathbf{V} = \{v_i\}_{i\in T}$ be the set of images in the exposure sequence. Based on the pair-wise consistency measure (Eq. 4), we define the probability that the images

in a given subset $S \in 2^{\mathbf{V}}$ are consistent at a pixel p as the minimum of the pair-wise consistency:

$$\Pr(p\,|S) = \min_{\{v_i,v_j\}\in S\times S} \Pr(p\,|\{v_i, v_j\})\,. \tag{5}$$

For the case of a singleton S (i.e., $|S| = 1$) the corresponding consistency probability is given as the probability that the corresponding observation is well-exposed:

$$\Pr(p|\,\{v_i\}) = 1 - \max\left\{\min_k \Pr_{\text{ue}}(v_i^k(p)), \max_k \Pr_{\text{oe}}(v_i^k(p))\right\}, \tag{6}$$

where $\Pr_{\text{ue}}$ and $\Pr_{\text{oe}}$ correspond to the under- and over-exposure probability, respectively, of an observation according to the distribution of the (Gaussian) readout noise, when centered at the black level and saturation level, respectively. In this definition, the probability that an observation $v_i(p)$ is inconsistent is high in two cases: When there is a high probability that *all* color channels are under-exposed, or when there is a high probability that *any* color channel is over-exposed.

2.3 Compositing of consistent sets

Since more than one subset of images can be consistent for a given pixel location, the choice of a particular subset to be averaged is under-constrained. We discuss regularizing this choice by requiring that the selected subsets be also spatially color-consistent. Together, the pixel-wise consistency test and the spatial consistency test cast the HDR deghosting problem as a Markov random field (MRF)-type global energy minimization. Consequently, to obtain a ghost-free HDR image, we minimize an energy function that promotes two criteria: Each pixel should be reconstructed from a consistent subset (encoded in a *consistency potential*, see Eq. (7) below), and given a pair of adjacent pixels, the image subsets used to reconstruct each pixel should be mutually consistent (i.e., the union of the subsets should be also consistent; encoded in a *prior potential*). Additionally, to prevent noisy reconstructions, we promote the selection of low-noise subsets whenever possible; this is encoded in a *noise potential*. Each possible HDR image is represented by a labeling $F(p) : \Omega \to 2^{\mathbf{V}}$ that assigns to each pixel p in the image domain Ω a subset $F_p := F(p)$ of the input images. We obtain a suitable labeling F by minimizing the energy functional:

$$\begin{aligned}\mathcal{E}(F) = \quad & \sum_{p\in\Omega} \Big(\underbrace{\mathbb{1}_{\{\Pr(p\,|F_p)<\alpha\}}}_{\text{consistency potential}} + \underbrace{\gamma V(F_p)}_{\text{noise potential}} \Big) + \\ & \beta \sum_{(p,q)\in\mathcal{N}} \underbrace{\mathbb{1}_{\{\Pr(p\,|F_{pq})<\alpha \vee \Pr(q\,|F_{pq})<\alpha\}}}_{\text{prior potential}},\end{aligned} \tag{7}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function, $F_{pq} \in 2^{\mathbf{V}}$ denotes the subset $F_p \cup F_q$, and $\mathcal{N}$ corresponds to the 4-neighborhood system in Ω. The scalar α denotes the confidence value (see below), and the scalars β and γ are weighting hyper-parameters.

In the consistency and prior potentials, instead of penalizing the consistency probability directly, we set a confidence value α to determine whether a set of images F_p is consistent or not. This encodes an important design choice: We want to select *any* consistent subset, not the *most* consistent one. This design gives more freedom to the optimization algorithm to construct the final composite.

The noise potential prevents the generation of trivial solutions. In Sec. 2.2, well-exposed observations from a single image are defined as consistent. Under this definition, selecting a single well-exposed image for reconstructing the whole image would create a labeling with minimum energy. This selection is undesired since the information contained in other consistent images is left out of the average, thus degrading the SNR of the resulting irradiance estimates (see Fig. 6, top row). Instead, whenever two distinct image subsets are consistent, we prefer the set that produces lower-noise estimates regardless of the set size. The noise potential $V(S)$ encodes this preference by assigning higher costs to sets that provide noisier estimates. The relative noise of each estimate is:

$$V(S) = \frac{\sigma_S}{\sum_{S'\in 2^{\mathbf{V}}} \sigma_{S'}}, \tag{8}$$

where the variance of each image set is approximated as $\sigma_S^2 = (\sum_{i\in S} 1/t_i^2)^{-1}$.

Parameter selection There are three hyper-parameters to be tuned in Eq. (7): The weight γ for the noise potential, the confidence value α of the consistency tests, and the weight β of the prior potential. We set the parameter γ to 0.1 to ensure that the noise potential in Eq. (7) produces order-of-magnitude lower costs than the consistency potential. This design instructs the algorithm to prefer consistent subsets, but when presented with several consistent options, it will prefer the one with the least noise. The other two parameters were determined based on a performance evaluation using the challenging *busy square* sequence (Fig. 8). The confidence value α was set to 0.98, which provides a good trade-off between sensitivity and specificity of ghost detection when compared to a manual annotation of the scene (see Sec. 3 for details). In our preliminary experiments, variations of α did not affect the results significantly. We set parameter β to 20, which is the lowest value that did not introduce visual discontinuities on the test sequence (see Fig. 6). Once determined, the parameters α, β, γ were fixed for all experiments presented in this paper.

Figure 6 shows the effects of varying parameters β and γ. When noisy subsets are not penalized ($\gamma = 0$; top row), the algorithm mostly selects a single image as source except for ill-exposed regions (white arrows), as only such regions are considered inconsistent. This behavior holds regardless of the weight β given to the prior potential. If noisy subsets are penalized mildly, i.e., less than inconsistent subsets ($\gamma = 0.1$; middle row), the remaining subsets of larger SNR (shaded in blue and green colors) are preferred providing they are consistent, resulting in labelings that adapt more to the scene. In this configuration, as β of the prior potential increases, visual discontinuities (marked by yellow arrows) are eliminated from the deghosted image (e.g., in $\beta = 10, 20$). When noisy subsets are penalized as much as inconsistent ones ($\gamma \geq 1$; bottom row), it becomes affordable to include objects that are partially ill-exposed (pointed by purple arrows) if they appear on the longest (less noisy) image. These results support our choice of γ.

Optimization and final reconstruction To obtain a minimum cost labeling F^*, we apply the expansion-move algorithm [Boykov et al. 2001; Boykov and Kolmogorov 2004]. With the resulting labeling, the final irradiance map is estimated as a weighted average:

$$\hat{\mu}^k_{x(p)} = \frac{\sum_{i\in F^*(p)} \Pr(p|\,\{v_i\}) W_i(p) x_i^k(p)}{\sum_{i\in F^*(p)} \Pr(p|\,\{v_i\}) W_i(p)}, \tag{9}$$

where $\Pr(p|\,\{v_i\})$ is the probability that $v_i(p)$ is well exposed (see Eq. (6)). The weighting function $W_i = (\sum_{k\in\mathbf{C}} \mathbb{V}\text{ar}\, x_i^k(p))^{-1}$ leads to a result close to the maximum likelihood solution [Robertson et al. 2003], and it is constraint to apply identical weights to every color channel in a given pixel.

Summary of pipeline The proposed pipeline for HDR deghosting is summarized as follows: (a) Take an input set of images and align them if taken hand-held; (b) if not provided by the manufacturer, estimate the readout noise using an additional black frame,

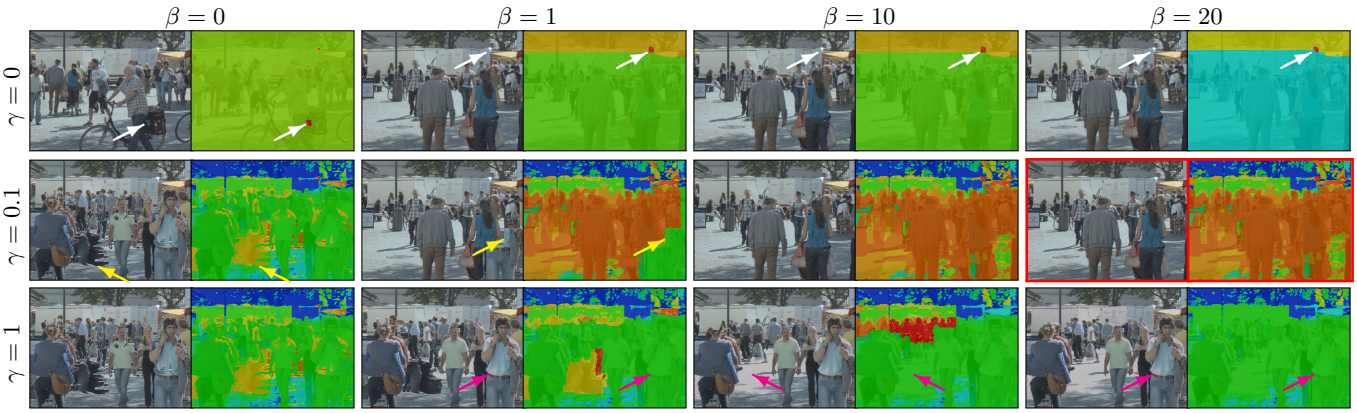


Figure 6: *Effect of varying the parameters β and γ in Eq. (7). The right-hand side colors correspond the estimated labeling, which is proportional to the noise of the selected subset (blue: higher SNR, red: higher SNR). We chose $\beta = 20$, $\gamma = 0.1$ (outlined in red) since these produce a good trade-off between low noise and spatial consistency. We kept these parameters fixed in all our experiments.*

Figure 7: *Small displacements in the* acrobat *and* street traffic *scenes. The left of each pair is the mean registered input image, thus ghosting shows small displacements. Our method on the right of each pair handles shifts of just a few pixels between exposures.*

Sequence	HH	SC	SD	LD	LL	Camera	Est. gain factor
Acrobat (Fig. 1)	×		×	×		Canon 550D	0.6597
Street traffic (Fig. 8)	×		×			Canon 550D	0.3753
Flower shop (Fig. 1)		×		×		Canon S5	0.2390
Busy square (Fig. 8)		×	×	×		Canon S5	0.2417
Café terrace (Fig. 9)			×			Canon S5	0.2250
Square at night (Fig. 10)		×	×	×	×	Canon S5	0.4125

Table 1: *Summary of test sequences. HH: Hand-held, SC: scene clutter, SD: small object displacements, LD: large object displacement, LL: low light. Gain factor for ISO100 setting.*

and the camera gain using one input image; (c) select a consistent subset of images for every pixel, and (d) reconstruct the irradiance of each pixel from the consistent sets.

3 Experimental validation

We acquired several sequences (see Table 1) using a Canon Powershot S5IS (10bit ADC) and a Canon EOS 550D (14bit ADC). Following the method in [Granados et al. 2010], the camera's black level ($L_0 = 32$ and $L_0 = 2048$, respectively) and readout variance ($\sigma_R^2 = 2.655$ and $\sigma_R^2 = 61.01$, respectively) were estimated from a black frame. The gain factor (Table 1) was estimated independently for every sequence using image-based calibration (Sec. 2.1). Although the gain needs to be estimated only once per camera model, we calibrate it per sequence to validate the robustness of our method. For reference, the gain factors obtained from flat-field calibration were $g = 0.2394$ and $g = 0.4795$, respectively.

Per scene, we captured three or five images in RAW mode at steps of one or two stops, respectively. A color image is constructed from the RGB measurements found on each 2×2 pixel block of the undemosaiced raw image (one of the measurements is not used). If captured hand-held, we robustly register the images using a global homography computed with RANSAC from sparse SURF keypoint matches. After HDR reconstruction, the images were white balanced and tone mapped using Drago et al. [2003] (*square at night* sequences) and Fattal et al. [2002] (all the remaining sequences).

The *acrobat* (Fig. 1) and *street traffic* (Fig. 8) scenes show hand-held capture with both small displacements (trees, people shifting their weight) and large displacements with fast motion (acrobat, cars). We focus on our small displacement quality in Fig. 7, showing that our method produces convincing results. The *flower shop* (Fig. 1) and *busy square* (Fig. 8) sequences show how strong scene clutter can cause severe ghosting artifacts in an HDR reconstruction which includes every image into the irradiance average. In addition, the *square at night* (Fig. 10) sequence shows that our algorithm is robust to high image noise. The *café terrace* sequence (Fig. 9) and the additional *Christmas market* sequence (supplementary material) contain relatively small object displacements for which previous reference-image-based methods are designed [Sen et al. 2012]. Even under small displacements, which are well-handled by reference-image-based methods, our method produces results with less washed out regions and lower noise.

Comparison with reference-based methods We compare our approach to the state-of-the-art methods of Sen et al. [2012], and Zimmer et al. [2011] on the *busy square* sequence using their own implementations. The method of Sen et al. finds patch-wise correspondences between the reference and the remaining input images. As the reference image is of low dynamic range, regions that are ill-exposed or contain high noise might not be matched correctly to other exposures. This is demonstrated in Fig. 9, where the dynamic range of over-exposed regions could not be enhanced (indicated by arrows). Additionally, Fig. 10 shows that strong noise in the reference may restrict correspondence finding in other images for range enhancement, leading to a noisy HDR image. In contrast, our method is designed to select sets of images that are both consistent and have low noise, resulting in HDR images with comparatively

Figure 8: Left: *Hand-held capture via in-camera bracketing. The dynamic car motions are reconstructed ghost free.* Right: *Cluttered* busy square *sequence, where naive averaging produces severe artifacts (left-hand side) and our result is ghost free (right-hand side).*

less noise. In general, our method could also generate noisy image regions (see Fig. 8, right) if this guarantees consistency, as this is weighted more than achieving low noise (see Eq. (7)).

Zimmer et al. establish correspondences using optical flow, which will fail on objects that undergo large displacements or disocclusions. This failure case is shown on the person in Fig. 11, where ghosting artifacts are introduced after two instances of a person undergoing local motion cannot be properly aligned. In contrast, our method selects a single self-consistent image, thus preventing the introduction of ghosting artifacts.

Comparison with detect-and-exclude methods We compare our method against the top four performing methods reported by Sidibé et al. [2012], according to their sensitivity score: Grosch [2006], Sidibé et al. [2009], Heo et al. [2010], and Pece and Kautz [2010]. We used our own implementation of these methods using the exact parameters specified by the respective authors; since Grosch does not provide a difference threshold, we set it robustly to the median difference plus three median absolute deviations. All detect-and-exclude methods, including ours, work in two stages: Detect inconsistent regions, and reconstruct the HDR image using consistent parts only. Since the inconsistency detection is often noisy, they apply different regularization techniques before the reconstruction stage (e.g., Gaussian smoothing, morphological operations, or MRF priors; our method applies the latter). Therefore, to exclude the effect of different regularization strategies (i.e., of different image priors), only the detection stage of every method is compared (see Fig. 12). For the comparison, we used the first two input images of the *busy square* sequence. As ground truth, we constructed a manual segmentation of their differences (Fig. 12a). Table 2 summarizes the sensitivity and specificity achieved by each method in classifying pixels as consistent or inconsistent wrt. the ground truth. For a fair comparison, we present results with and without applying the difference filtering (DF) step of our method.

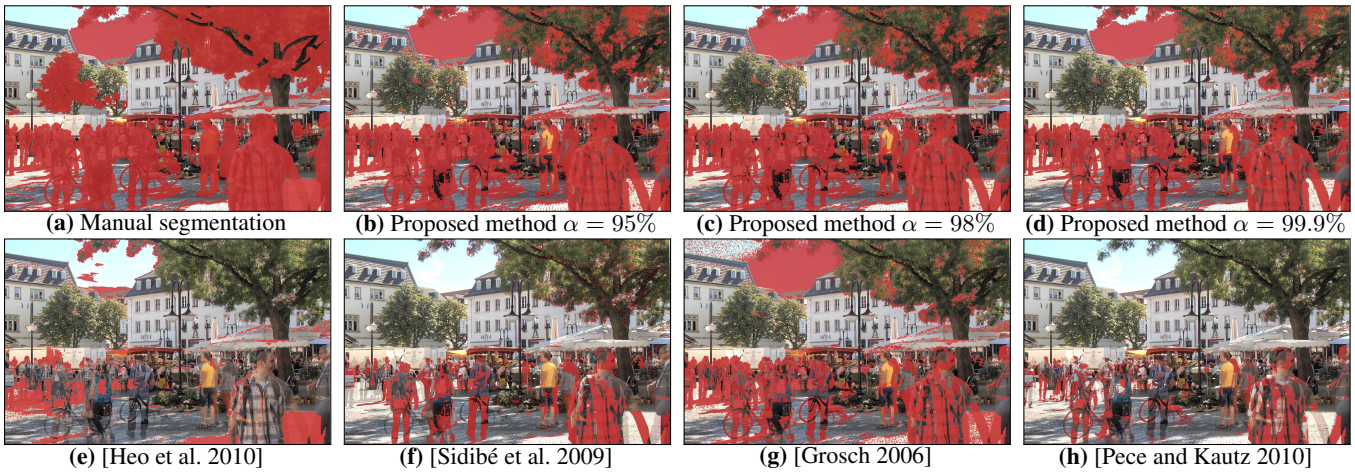
Among previous methods, Grosch's approach achieved the best sensitivity (43.5%) by thresholding the absolute irradiance difference between the images (Fig. 12g). The methods of Sidibé et al. (Fig. 12f) and Pece and Kautz (Fig. 12h) achieve the highest specificity (99.4% and 99.9%) but the lowest sensitivity (24.6% and 15.8%). This occurs as both methods are based on invariants

Figure 10: *Comparison with the method of Sen et al. on the* Square at night *sequence (top). The second exposure was selected as reference for Sen et al.'s method. Due to noise, their method finds few similar patches in other exposures. This implies that the dynamic range cannot be effectively extended using other input images (middle). Our method selects consistent sources with as low variance as possible, preventing the appearance of noise in the result (bottom).*

that are satisfied whenever two pixels correspond to the same light intensity, but this is not always violated by moving objects.

We tested our method with confidence values $\alpha = \{0.95, 0.98,$

Figure 9: *Comparison to Sen et al. on the* Café terrace *sequence (top). The third image was selected as reference for the method of Sen et al. Here, their method encounters difficulties extending the dynamic range of ill-exposed regions, which results in a washed-out appearance (indicated by arrows). In contrast, our method automatically selects well-exposed sources for every region.*



(a) Manual segmentation **(b)** Proposed method $\alpha = 95\%$ **(c)** Proposed method $\alpha = 98\%$ **(d)** Proposed method $\alpha = 99.9\%$

(e) [Heo et al. 2010] **(f)** [Sidibé et al. 2009] **(g)** [Grosch 2006] **(h)** [Pece and Kautz 2010]

Figure 12: *Comparison of our consistency detector with other state-of-the-art ghosting-detection methods. Here, the differences between a pair of images of the* busy square *(Fig. 8, right) are shown in red on top of their average color.*

(a) **(b)** **(c)** **(d)**

Figure 11: *Comparison with the method of Zimmer et al. on the* busy square *sequence:* **(a)** *Reference image,* **(b)** *optical-flow alignment of an additional input image to the reference,* **(c)** *result after HDR reconstruction using (a) and (b), and* **(d)** *our result.*

$0.999\}$, and with and without applying noise-adaptive difference filtering (DF) (see Sec. 2.2). In all cases, our was higher than previous methods (46.7–58.3% vs. 43.5% for Grosch). With our adaptive DF, the specificity was comparable to that of other methods, including those methods based on invariants. The best trade-off was obtained at $\alpha = 0.98$ with sensitivity and specificity of 51% and 95%, respectively (Fig. 12c). Our method achieves the best sensitivity, which is crucial for removing ghosts, without compromising the specificity, which is crucial for producing low-noise HDR images.

4 Discussion

Handling of challenging scenes and motion blur Our method produces plausible HDR images of scenes with small and large object displacements and clutter (Figs. 1, Fig. 8, 9 and 11), scenes taken hand-held (Fig. 1, left, and Fig. 8, left), and scenes taken during the night (Fig. 10). To the best of our knowledge, this is the first method which demonstrates ghost-free results in all of these scenarios. Furthermore, the parameters used for all results were identical. However, our method does not detect motion blur, and so blurred objects in long exposures could be selected by our algorithm. In the future, blur-detection methods can be used to exclude such objects.

Handling of HDR moving objects Our method cannot recover the dynamic range of moving HDR objects, i.e., objects that cannot be properly captured in a single exposure, as it only performs

Detection strategy	Sensitivity	Specificity	Avg. diff.	SNR (dB)
Proposed method (-DF), $\alpha = 95.0\%$	0.583	0.750	1.0x	28.30
Proposed method (-DF), $\alpha = 98.0\%$	0.542	0.881	1.2x	28.39
Proposed method (-DF), $\alpha = 99.9\%$	0.480	0.979	1.7x	28.46
Proposed method (+DF), $\alpha = 95.0\%$	0.536	0.899	1.2x	28.41
Proposed method (+DF), $\alpha = 98.0\%$	**0.513**	**0.947**	**1.4x**	**28.44**
Proposed method (+DF), $\alpha = 99.9\%$	0.467	0.987	1.8x	28.47
Absolute difference [Grosch 2006]	0.435	0.928	2.3x	28.38
IMF probability [Heo et al. 2010]	0.254	0.949	5.5x	28.38
Monotonic ordering [Sidibé et al. 2009]	0.246	0.994	9.8x	28.47
Median threshold [Pece and Kautz 2010]	0.158	0.999	9.7x	28.47

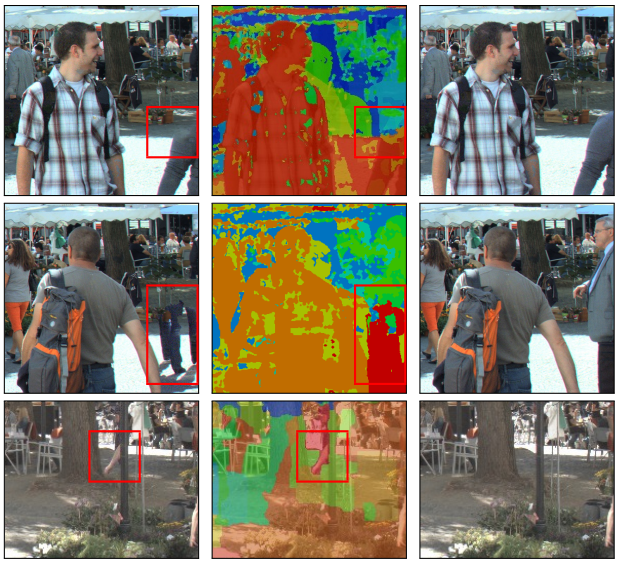
Table 2: *Comparison with existing ghost-detection methods. Unlike existing techniques, our method automatically estimates an appropriate range kernel bandwidth to bilaterally filter image difference images. This improves sensitivity for a given specificity. Other methods would require a user in the loop to estimate the filtering bandwidth and accomplish the same improvement. For comparison, we show results with and without applying difference filtering (denoted by +DF and -DF). The avg. diff. column shows the average color difference between true-negative detections (i.e., dynamic objects detected as static) as a factor of the best method's detection, where smaller factors imply a more accurate detection. The last column illustrates the decrease in SNR caused by false-positives (i.e., static objects detected as dynamic).*

a global image alignment and not a local alignment between moving objects in different exposures. As a result, moving objects are likely to be reconstructed from a single image. This could be alleviated using a correspondence-based method [Sen et al. 2012] that accounts for noise. However, in dynamic scenes with deforming objects and occlusions, there is never a guarantee that the same object surface will be observed in different exposures, and without this guarantee, correspondence-based reconstruction is sometimes impossible.

Handling of hand-held capture Our method successfully handles hand-held capture (Fig. 1, left, and Fig. 8, left) whenever the camera motion can be approximated using a homography. Other objects moving independently are not registered but are implicitly handled through a optimization procedure which selects one of the instances available in the input (usually their best exposure with respect to noise).

Time complexity The C++ implementation of our algorithm takes between one and five minutes to deghost sets of three to five LDR images at 1648×1236 resolution on an Intel Core i5 3GHz CPU. Larger image stacks will lead to higher run times as our method considers every possible combination of input images. In practice, stacks of three to five images are sufficient to reconstruct the dynamic range of most scenes if their exposure times are sufficiently separated. In addition, exposure selection methods that work at acquisition time [Gallo et al. 2012] could be used to select the best five-image-or-less subset.

Interaction for handling semantic inconsistencies In some cases, our method may produce semantic inconsistencies, such as half-included objects, or twice the same object in the final image. This may occur in three cases. In the first case, objects at the same location in different images that have consistent colors could become merged in the final HDR image. This is because observations can only be compared up to the noise level of the signal. This case is illustrated in Fig. 13–top, where the color of the shirt of the person indicated is consistent with the background color. This results into a partial inclusion of the person, as the algorithm prefers the lower-variance background image. The second case arises when all objects at a given location on different images are ill-exposed. In this case, no object can be fully included without averaging ill-exposed



(a) Deghosting **(b)** Labeling **(c)** After editing

Figure 13: *Semantic inconsistencies and interactive correction: Our algorithm may produce semantic inconsistencies (a). These can appear when the color difference falls below the noise level (top), when all objects in a given image region are partially ill-exposed (middle), or when objects are partially occluded (bottom). These inconsistencies can be corrected interactively by editing the labels (b). The results after editing are shown in (c).*

pixels, which leads to visual discontinuities. Resolving this situation requires deciding between using ill-exposed pixels or splitting objects in half. This is illustrated in Fig. 13–middle, where we provoke this case by performing the deghosting excluding the shortest and longest exposure of the *busy square* sequence. In the deghosted image, only the legs of the persons at the right are included (enclosed in red). In the last case, our algorithm may produce semantic incongruencies, either by including multiple instances of the same object, or by including only some parts of visually disconnected but conceptually whole objects. This is visible in Fig. 8–right, where the person holding a suitcase appears twice in the final HDR image, and in Fig. 13–bottom, where only the part of a person occluded by a lamp post is included. In general, these three cases can be corrected with user interaction by editing the automatic labeling (see Fig. 13). Except for the *flower shop* sequence (Fig. 1, right), all the results presented in this paper were computed fully automatically.

5 Conclusions

We have presented a robust method to model image noise and produce ghost-free HDR reconstructions. Our algorithm uses a new consistency measure that exploits the estimated noise distribution in images. This avoids the need for any reference image or a background model. The resulting consistency measure is combined with a spatial coherence prior and constitutes an MRF-type energy minimization framework. Experiments demonstrated that our algorithm can be applied to challenging dynamic and cluttered scenes which cannot be handled with existing algorithms, and also performs on par with state-of-the-art techniques for less challenging scenes. As such, our algorithm moves towards a widely-applicable algorithm for ghost-free dynamic scene HDR reconstruction.

Acknowledgments. We thank P. Sen and O. Veskler for making publicly available an implementation of their HDR deghosting and super pixel segmentation methods, respectively. We thank H. Zimmer for kindly providing us with results of their method.

References

BAY, H., ESS, A., TUYTELAARS, T., AND GOOL, L. J. V. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding 110*, 3, 346–359.

BOGONI, L. 2000. Extending dynamic range of monochrome and color images through fusion. In *Proc. ICPR*, 3007–3016.

BOYKOV, Y., AND KOLMOGOROV, V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI 26*, 9, 1124–1137.

BOYKOV, Y., VEKSLER, O., AND ZABIH, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE TPAMI 23*, 11.

BURT, P. J., AND KOLCZYNSKI, R. J. 1993. Enhanced image capture through fusion. In *Proc. ICCV*, 173–182.

DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND CHIBA, N. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *CGF 22*, 3, 419–426.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. *ACM TOG 21*, 3.

FISCHLER, M. A., AND BOLLES, R. C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM 24*, 6, 381–395.

GALLO, O., GELFAND, N., CHEN, W.-C., TICO, M., AND PULLI, K. 2009. Artifact-free high dynamic range imaging. In *Proc. ICCP*, 1 –7.

GALLO, O., TICO, M., MANDUCHI, R., GELFAND, N., AND PULLI, K. 2012. Metering for exposure stacks. *CGF 31*, 2.

GRANADOS, M., SEIDEL, H.-P., AND LENSCH, H. P. A. 2008. Background estimation from non-time sequence images. In *Proc. GI*, 33–40.

GRANADOS, M., AJDIN, B., WAND, M., THEOBALT, C., SEIDEL, H.-P., AND LENSCH, H. P. A. 2010. Optimal HDR reconstruction with linear digital cameras. In *Proc. CVPR*, 215–222.

GROSCH, T. 2006. Fast and robust high dynamic range image generation with camera and object movement. In *Proc. VMV*.

HEO, Y. S., LEE, K. M., LEE, S. U., MOON, Y., AND CHA, J. 2010. Ghost-free high dynamic range imaging. In *Proc. ACCV*, vol. 4, 486–500.

HU, J., GALLO, O., AND PULLI, K. 2012. Exposure stacks of live scenes with hand-held cameras. In *Proc. ECCV*, 499–512.

HUBBARD, W. M. 1970. The approximation of a Poisson distribution by a Gaussian distribution. *Proc. IEEE 58*, 9, 1374–1375.

JACOBS, K., LOSCOS, C., AND WARD, G. 2008. Automatic high-dynamic range image generation for dynamic scenes. *IEEE CGA 28*, 2, 84–93.

JANESICK, J. 2001. *Scientific charge-coupled devices*. SPIE Press.

KANG, S. B., UYTTENDAELE, M., WINDER, S. A. J., AND SZELISKI, R. 2003. High dynamic range video. *ACM TOG 22*, 3, 319–325.

KHAN, E. A., AKYÜZ, A. O., AND REINHARD, E. 2006. Ghost removal in high dynamic range images. In *Proc. ICIP*.

LIU, C., SZELISKI, R., KANG, S. B., ZITNICK, C. L., AND FREEMAN, W. T. 2008. Automatic estimation and removal of noise from a single image. *IEEE TPAMI 30*, 2, 299–314.

MANN, S., AND PICARD, R. 1995. Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proc. IS&T*, 422–428.

MENZEL, N., AND GUTHE, M. 2007. Freehand HDR photography with motion compensation. In *Proc. VMV*, 127–134.

MIN, T.-H., PARK, R.-H., AND CHANG, S. 2009. Histogram based ghost removal in high dynamic range images. In *Proc. ICME*, 530–533.

PECE, F., AND KAUTZ, J. 2010. Bitmap movement detection: HDR for dynamic scenes. In *Proc. CVMP*, 1–8.

PEDONE, M., AND HEIKKILÄ, J. 2008. Constrain propagation for ghost removal in high dynamic range images. In *Proc. VISAPP*.

RAMAN, S., AND CHAUDHURI, S. 2010. Bottom-up segmentation for ghost-free reconstruction of a dynamic scene from multi-exposure images. In *Proc. ICVGIP*, 56–63.

REINHARD, E., WARD, G., PATTANAIK, S., AND DEBEVEC, P. 2005. *High dynamic range imaging: Acquisition, display and image-based lighting*. Morgan Kaufmann publishers.

ROBERTSON, M., BORMAN, S., AND STEVENSON, R. 2003. Estimation-theoretic approach to dynamic range improvement using multiple exposures. *J. Elec. Imag. 12*, 2, 219–228.

SEN, P., KALANTARI, N. K., YAESOUBI, M., DARABI, S., GOLDMAN, D., AND SHECHTMAN, E. 2012. Robust patch-based hdr reconstruction of dynamic scenes. *ACM TOG 31*, 6.

SIDIBÉ, D., PUECH, W., AND STRAUSS, O. 2009. Ghost detection and removal in high dynamic range images. In *Proc. EUSIPCO*.

SILK, S., AND LANG, J. 2012. Fast high dynamic range image deghosting for arbitrary scene motion. In *Proc. GI*, 85–92.

SIMAKOV, D., CASPI, Y., SHECHTMAN, E., AND IRANI, M. 2008. Summarizing visual data using bidirectional similarity. In *Proc. CVPR*.

SRIKANTHA, A., AND SIDIBÉ, D. 2012. Ghost detection and removal for high dynamic range images: Recent advances. *Sig. Proc.: Image Comm. 27*, 6, 650–662.

TOCCI, M. D., KISER, C., TOCCI, N., AND SEN, P. 2011. A versatile hdr video production system. *ACM TOG 30*, 4, 41.

TOMASI, C., AND MANDUCHI, R. 1998. Bilateral filtering for gray and color images. In *Proc. ICCV*, 839–846.

TOMASZEWSKA, A. M., AND MARKOWSKI, M. 2010. Dynamic scenes HDRI acquisition. In *Proc. ICIAR*, vol. 2, 345–354.

VEKSLER, O., BOYKOV, Y., AND MEHRANI, P. 2010. Superpixels and supervoxels in an energy optimization framework. In *Proc. ECCV*, vol. 6315, 211–224.

ZHANG, W., AND CHAM, W.-K. 2012. Gradient-directed multi-exposure composition. *IEEE TIP 21*, 4, 2318–2323.

ZIMMER, H., BRUHN, A., AND WEICKERT, J. 2011. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *CGF 30*, 2, 405–414.