https://doi.org/10.2352/ISSN.2470-1173.2018.05.PMII-311

Overview of State-of-the-Art Algorithms for Stack-Based High-Dynamic Range (HDR) Imaging

Pradeep Sen

Abstract

Modern digital cameras have very limited dynamic range, which makes them unable to capture the full range of illumination in natural scenes. Since this prevents them from accurately photographing visible detail, researchers have spent the last two decades developing algorithms for high-dynamic range (HDR) imaging which can capture a wider range of illumination and therefore allow us to reconstruct richer images of natural scenes. The most practical of these methods are stack-based approaches which take a set of images at different exposure levels and then merge them together to form the final HDR result. However, these algorithms produce ghost-like artifacts when the scene has motion or the camera is not perfectly static. In this paper, we present an overview of state-of-the-art deghosting algorithms for stack-based HDR imaging and discuss some of the tradeoffs of each.

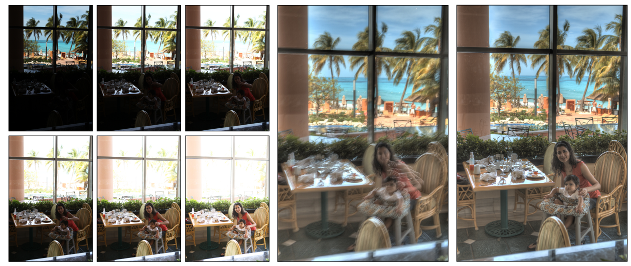
Introduction

Natural scenes have a wide range of illumination and usually contain very dark and very bright objects. Although our eyes can accurately sense this large difference in light intensity–allowing us to see dark and bright regions simultaneously–standard digital cameras suffer from a very limited dynamic range and do not have this capability. For example, a camera with an 8-bit imaging sensor (which is common, especially for mobile cameras) can only represent 255 distinct intensity levels between the darkest and brightest portions of the image, which does not give us enough bit-depth to represent detail in differently exposed regions.

Furthermore, imaging sensors are plagued by noise at low light levels (where the signal-to-noise (SNR) ratio is poor) and become saturated at high light levels (where the accumulator that counts photons simply runs out of numbers), making it difficult to recover the original scene intensity at each pixel from the final image. Theoretically speaking, if the sensor had infinite bit-depth (i.e., no quantization) and no noise, we could capture the full illumination range in a scene by simply exposing the sensor as long as possible while keeping the brightest region from becoming saturated. The result could then be simply scaled to any desired exposure level without noise or quantization artifacts.

Unfortunately, real sensors are going to suffer from both noise and coarse quantization levels for the foreseeable future, which means that they cannot capture the full dynamic range of many natural scenes. This fundamental limitation results in images that do not have the visual richness and detail of the natural world. To address this problem, several technical solutions have been explored. For example, custom camera hardware has been proposed that will capture a wider dynamic range directly. These include systems that have custom sensors (e.g., with different neutral density filters over each pixel [33]) or use beam-splitters [43] to capture differently exposed images simultaneously, which can then be used to compute the final HDR result. Although some of these systems have demonstrated impressive results, they are still not widely available and are often quite expensive.

In this paper, we will focus instead on *practical* approaches



(a) LDR input images **(b)** Ghosted HDR **(c)** Sen et al. [39]

Figure 1. **(a)** *Stack of differently exposed images of a scene. The limited dynamic range of the camera makes it unable to capture the interior and exterior simultaneously, even though it was perfectly visible to the naked eye.* **(b)** *High-dynamic range (HDR) image produced with the original merging algorithm of Debevec and Malik [6] that leverages well-exposed information from each image in the stack (see Eq. 1). Although the result contains well-exposed detail in all regions of the image, there are objectionable ghosting artifacts because of subject and camera motion while the stack was captured.* **(c)** *Result from the patch-based deghosting algorithm of Sen et al. [39]. This result captured the full range of illumination without the ghosting artifacts. All HDR results in this paper have been tonemapped for display.*

to HDR imaging with standard image sensors. This is typically done by capturing a stack of normal, low-dynamic range (LDR) images at different exposures (known as an *exposure stack*), which can then be *merged* together to produce the final result. For example, we would take long exposures to capture all the detail in the dark regions of the scene and short exposures to capture all the bright details (see Fig. 1). However, because the images are taken in a temporal sequence, there can be motion between them from either dynamic objects or a moving camera, which results in objectionable *ghosting* artifacts in the final HDR image. The bulk of this paper is dedicated for providing an overview of state-of-the-art algorithms for deghosting these HDR images. We begin, however, by providing basic background on stack-based approaches in the next section. Readers interested in more detail are referred to other, longer overview papers on the subject [11, 38, 42].

Background on stack-based HDR imaging

The idea of combining differently exposed images together to form an image with higher dynamic range is actually quite old. Early photographers such as Hippolyte Bayard and Gustave Le Gray were performing this technique in the mid 1800's, combining differently exposed negatives to produce a print with significant detail in both the bright and dark regions. In the context of digital imaging, the modern stack-based HDR approaches were first proposed by Madden [27] and later by Mann [30], but were popularized by Debevec and Malik [6] with their seminal paper.

To understand how this process works, assume we acquire a stack of N low-dynamic range images $I_1, \ldots, I_N$ with different exposure times $t_1, \ldots, t_N$. Our goal is to recover the *incident irradiance* E (the amount of light power per unit area arriving on the sensor), where the irradiance at pixel p is given by $E(p)$. This irradiance will be the same for all images assuming a static scene and camera. The camera integrates this irradiance over the different exposure times and measures *exposures* $X_1,\dots,X_N$, where each exposure is simply a linear function of incident irradiance scaled by the exposure time: $X_k(p) = E(p)\cdot t_k$.

In order to make the images look better and model the human visual response, however, most cameras do not output the X_k exposures directly[1], but rather apply a non-linear camera response function $f()$ to the exposure to get the resulting images: $I_k(p) = f(X_k(p))$. In order to perform the HDR reconstruction process, we must undo this by first converting our input images $I_1,\dots,I_N$ to linear exposures $X_1,\dots,X_N$ using the inverse of the camera response function: $X_k(p) = f^{-1}(I_k(p))$. Although this requires the camera response function to be known, many approaches have been proposed to estimate it directly from the image stack for static [6] or dynamic scenes [1, 14].

Once the exposures X_k have been recovered, we then *merge* them together and compute the estimated irradiance map $\widetilde{E}$ by computing a weighted average of the measured irradiances $E_k(p) = X_k(p)/t_k$ of each image:

$$\begin{aligned}\widetilde{E}(p) &= \frac{\sum_{k=1}^{N} w_k(p)\cdot E_k(p)}{\sum_{k=1}^{N} w_k(p)} = \frac{\sum_{k=1}^{N} w_k(p)\cdot X_k(p)/t_k}{\sum_{k=1}^{N} w_k(p)} \\ &= \frac{\sum_{k=1}^{N} w_k(p)\cdot h(I_k(p))}{\sum_{k=1}^{N} w_k(p)}, \end{aligned} \tag{1}$$

where we divide by the sum of the weights w_k to ensure they add up to one (i.e., normalized), and the last equation is written in terms of the input images using a function that maps LDR values directly to HDR irradiance values: $h(I_k(p)) = f^{-1}(I_k(p))/t_k$.

To reduce the influence of over- or under-exposed pixels, Debevec and Malik [6] proposed a simple “triangle” function for the weights: $w_k(p) = \min(I_k(p), 255 - I_k(p))$, assuming pixel values in the range from 0 to 255. The resulting irradiance map $\widetilde{E}$ can either be output as the final HDR result or *tonemapped* by applying a non-linear function for presentation on a low-dynamic range display. Note that in this discussion, we ignored color channels which are typically captured by a Bayer pattern which measures different colors at each pixel and must be first be *demosaiced* to get full RGB color. As observed by Tocci et al. [43], we must perform merging *prior* to demosaicing as certain colors might be saturated, causing discoloration artifacts if we merge afterwards.

A clever alternative to the standard HDR merge of Eq. 1 is *exposure fusion*, proposed by Mertens et al. [31]. Exposure fusion does not compute an HDR irradiance map, but rather directly *fuses* the input low-dynamic range images I_k to get another low-dynamic range image $\widetilde{I}$ with well-exposed detail everywhere:

$$\widetilde{I}(p) = \frac{\sum_{k=1}^{N} w_k(p)\cdot I_k(p)}{\sum_{k=1}^{N} w_k(p)}. \tag{2}$$

Here, weights w_k are computed by a product of contrast, saturation, and well-exposedness metrics so that pixels in the input images that do not meet these properties are weighted less or ignored altogether. However, naïve implementation of Eq. 2 does not work, as all the images are at different exposure levels and there will be visible seams where different images come together.

To address this problem, exposure fusion blends the images by first constructing a Laplacian pyramid of the input images I_k and a Gaussian pyramid of the weights w_k. At each scale, the Laplacian level of each image is multiplied by the Gaussian level of the respective weight and summed over all images, producing a Laplacian pyramid for the output which can then be reconstructed to produce the final fused result. Exposure fusion can produce nice images that do not require any tonemapping for display on a low-dynamic range media. However, they have the limitation that by definition an HDR image is never created. Therefore, it cannot be used in situations where an HDR image is needed (e.g., lighting environment maps for rendering, post-production special effects shots, for display on an HDR monitor, and so on).

Finally, so far we have proposed to change the exposure for every image in the stack to capture the full dynamic range of the scene. However, there is an alternative, stack-based approach for HDR imaging known as *burst HDR* imaging that captures a sequence of equal, short-exposure images [16, 48]. Here, the exposure is set short enough so that nothing in the image is saturated, hence the only issues to deal with are noise and quantization. These methods specifically address the noise by averaging all the images in the stack together, effectively transforming the HDR problem into a problem of denoising images.

The fundamental problem of *all* stack-based HDR methods (whether standard HDR merge, exposure fusion, or burst HDR imaging) is that they assume the scene radiance values are constant during the capture of the entire stack. Objects moving from frame from frame will cause ghost-like artifacts in the final result (see Fig. 1b). Furthermore, unlike HDR camera hardware which captures differently exposed images simultaneously, stack-based approaches cannot guarantee recovery of the true high-dynamic range information. For example, if the lady’s arm in Fig. 1 had been blocking the window in the first three LDR images where the exterior is well exposed, we would not have the necessary information to complete the window with 100% accuracy.

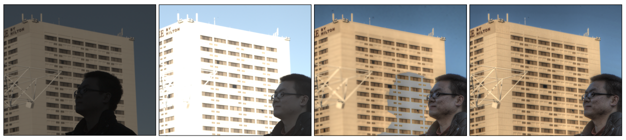
To address the ghosting problems with stack-based HDR approaches, researchers have explored various *deghosting* algorithms, which we will spend most of this paper reviewing. To discuss them, we build on the taxonomy of deghosting algorithms proposed by Sen et al. [39] and discussed elsewhere [11, 38].

Rejection methods

Some of the earliest HDR deghosting methods proposed were *rejection methods*. These methods assume that only a few pixels in the image contain moving objects, and that most of the image is of static content (or that a simple alignment process can be used to “freeze” most of the pixels). The basic idea behind these methods is to determine which pixels are affected by motion and to process them differently. Specifically, in pixels that contain static objects, the standard HDR merge given by Eq. 1 can be performed since it will not produce ghosting artifacts. This reconstructs HDR information for most of the final image. For pixels that contain moving objects, on the other hand, only a subset of the input images that are deemed to be static are merged together. In other words, information from pixels with moving objects is *rejected* from the final result at these pixel locations.

The main difference between rejection-based methods is in how they detect pixels affected by motion and how they decide what subset of images to merge together at each pixel. Some algorithms identify one of the images in the stack to be the *reference*

[1] An exception are RAW images, which are typically linear (no camera response function applied) and so are equivalent to the exposures X_k.



(a) LDR input images **(b)** Heo et al. [17] **(c)** Sen et al. [39]

Figure 2. *(a) Input images, with the brighter one chosen as the reference to match. (b) The reference-based rejection method of Heo et al. [17] detects motion in the region occluded by the head in the dark input image, so it falls back to the reference. However, the reference is saturated in this region, resulting in a gray "shadow" in the final result because the HDR irradiance values are all clamped. (c) For comparison, the patch-based synthesis method of Sen et al. [39] is able to fill in the missing information.*

image, which means the final HDR result will have that specific composition, as if an HDR camera had captured it. Depending on the algorithm, the reference is either selected by the user or computed automatically (e.g., the picture with the best overall exposure or the image in the middle of the stack).

Early reference-based, rejection algorithms used simple heuristics to detect motion and then only merged together the images that were determined to be static with respect to the reference. For example, Grosch [13] propose a two-image approach which maps pixels from the reference to the other exposure and compares them. If their difference is larger than a given threshold, only the reference is used in these regions. Gallo et al. [10] build on this idea by comparing the patches surrounding the pixels instead, and doing so in the log domain. Min et al. [32] use the input images to compute multilevel threshold maps which are compared against that of a reference to determine regions with motion. Wu et al. [47] use consistency in the radiance and color across exposure to identify static pixels. Heo et al. [17] compute joint histograms of the pixel values in the reference and the other images. These histograms are converted into a joint probability and ghosted regions are identified as those with a joint probability less than a fixed threshold. Raman and Chaudhuri [36] extend the algorithm of Gallo et al. [10] by segmenting the image into super-pixels instead of patches to avoid edge artifacts.

The second class of rejection methods do not use a reference, but must rather "stitch together" static portions of different images to produce a coherent final result. The earliest example of these *reference-free* rejection methods was the algorithm of Reinhard et al. [37], which uses the weighted normalized variance at each pixel to determine which ones are impacted by motion. Eden et al. [7] first apply graphcuts on the input images to stitch together a coherent "reference" that will not have the entire dynamic range. The HDR result is then created by adding well-exposed detail from the other exposures to regions where this reference is poorly exposed. Khan et al. [26] use several iterations of kernel density estimation to adjust the weights of the HDR merging process assuming that most pixels will be static. This reduces the contribution of pixels with dynamic content. Jacobs et al. [20] extend the approach of Reinhard et al. [37] to detect motion using an local entropy metric instead of normalized variance. Sidibe et al. [40] identify static pixels by testing if pixel values increase as exposure increases. Pece et al. [35] propose to compute the median threshold bitmap (binary images with 1's for all pixels greater than the median value, 0 elsewhere) for each image in the stack and identify motion by comparing their values. Zhang and Cham [49] use changes in the gradient between exposures to detect motion.

Figure 3. *The reference-free rejection method of Granados et al. [12] takes the stack of images on the left and produces the result shown on the right. Although most of the image is properly deghosted, reference-free approaches often suffer from semantic artifacts such as the duplicated man (red arrows) and the cut-off person (green arrow). Image courtesy of Granados et al. [12].*

One of the state-of-the-art, rejection methods is the algorithm of Granados et al. [12], which uses a model that accounts for measurement noise to identify consistent subsets of pixels in the stack that can be merged together without ghosting. Since these subsets can be different from pixel to pixel (resulting in spatial artifacts), they propose an optimization with a coherency term that attempts to select the same subsets for neighboring pixels. However, this can still result in semantic errors (e.g., a moving object appearing twice in the final image, such as shown in Fig. 3) and so manual intervention is sometimes needed.

Finally, there are clever reference-free rejection methods that use rank minimization [34]. They observe that for static scenes all exposures X_k are linearly related since $X_k = E \cdot t_k$. Therefore, after removing global motion due to camera motion using homographies (see more in the next section), they stack all the exposures X_k together to form a matrix and they compute a rank-1 matrix that represents their measured exposures as close as possible plus some sparse "noise" representing the regions affected by motion. In this way, they attempt to reject the misaligned content.

Discussion

Advantages: Rejection algorithms tend to be fairly robust. easy to implement and fairly fast. For this reason they are widely implemented in various HDR applications on mobile devices as well as in commercial HDR packages. These algorithms are best suited for largely static scenes (e.g., a set of buildings, a courtyard, etc.) with only small dynamic objects (e.g., people walking around).

Disadvantages: Rejection algorithms have many problems. First, they are unable to handle moving HDR content since they only reconstruct HDR information using the standard merge in static regions. If an object has HDR content but moves from image to image, the algorithm can only use information from one image and will lose the HDR detail. Furthermore, these algorithms assume most of the image is static, since these are the only regions where HDR information is reconstructed. For this reason, they cannot handle large moving content, such as a full-frame, moving subject. Rejection methods that use a reference have problems if motion is detected in regions where the reference is poorly exposed, since they fall back to the reference in this region (see Fig. 2). Similarly, reference-free rejection-based methods can produce images with duplicate objects or similar artifacts because they do not enforce the semantic meaning of the scene.

Alignment methods

The second kind of early deghosting algorithms were *alignment methods*. As their name implies, these methods attempt to "align" or warp the images in the stack to register them with a reference image. Once the images are aligned, standard HDR merging using Eq. 1 can be used since the scene is now essentially "static." To accomplish this, some early alignment methods used *rigid transformations* to align the images together. For example, Ward [46] observes that some image stacks could be aligned with simple translations that can be efficiently found by comparing the images' median threshold maps. Tomaszewska and Mantiuk [44] extend this idea to compute homographies by applying RANSAC on the SIFT feature matches between the images.

More sophisticated methods used more advanced, non-rigid warping techniques based on optical flow. The earliest known example of this is by Bogoni [4], which first applies an affine transform to globally align the images together. This is done by computing optical flow fields between the images in multiscale fashion from coarse to fine using a Laplacian pyramid, and then fitting affine models to these flow fields using weighted least-squares. The affine transforms are used to warp the images, and then a second optical flow computes the final deformation field to warp the individual source images to the reference.

Kang et al. [24] also propose an optical-flow-based approach for capturing HDR video from two alternating exposures (see more on this topic later in the paper). Specifically, they use gradient-based optical flow to compute the bidirectional flow from each image to its neighbors, as well as unidirectional flows from the neighboring frames to the current frame. These flows are used to compute four warped images by deforming the two neighboring frames, and the resulting images are then merged together with the reference (the current frame) using a weighting scheme that rejects pixels that are still misaligned.

Jinno and Okuda [21] use Markov Random Fields to estimate the local displacement to align the images together as well as the occlusion and saturation to reject certain regions from the merging step. Zimmer et al. [50] use an energy-based optical flow optimization that is robust to changes in exposure to align the images. The data term in their energy function tries to align the image to the reference while the regularizer ensures that the flow is smooth wherever the reference is poorly exposed. Hu et al. [18] compute dense correspondences between the images using the patch-based non-rigid dense correspondence (NRDFC) algorithm [15], and fill in holes of missing information with pasted pixels from the transformed source. Finally, Gallo et al. [9] propose an algorithm designed to do the alignment very quickly for mobile applications. They observe that for images taken in a fast burst, the motion tends to be very limited, so instead of computing optical flow at every pixel of the image, they only compute it at sparse locations and interpolate it to the rest of the pixels.

Discussion

Advantages: Unlike rejection methods that compute the HDR irradiance value at each pixel using only information from the same pixel through the stack, alignment methods can in theory handle true, dynamic HDR content because they can move information between pixels. This gives them a significant advantage in reconstructing true HDR content for dynamic scenes.

Disadvantages: However, alignment methods for HDR reconstruction suffer from several serious problems. First of all, the methods that perform simple registrations [46, 44] cannot handle the kind of general motion that happens in natural scenes. Furthermore, even algorithms based on optical flow have trouble with complex scenes, since optical flow can be fairly brittle for the cases of complex deformable motion. One example is shown in Fig. 4 where the state-of-the-art alignment algorithm of Zimmer et al. [50] produces visible artifacts compared to other methods. For this reason, alignment algorithms are often used in conjunction with a rejection method as well to handle any left over misalignments. Finally, these approaches cannot handle disoccluded content because they cannot synthesize new information. For example, in the scene in Fig. 2, no amount of warping of the dark exposure would produce the correct result because the occluded content is simply missing in that image.

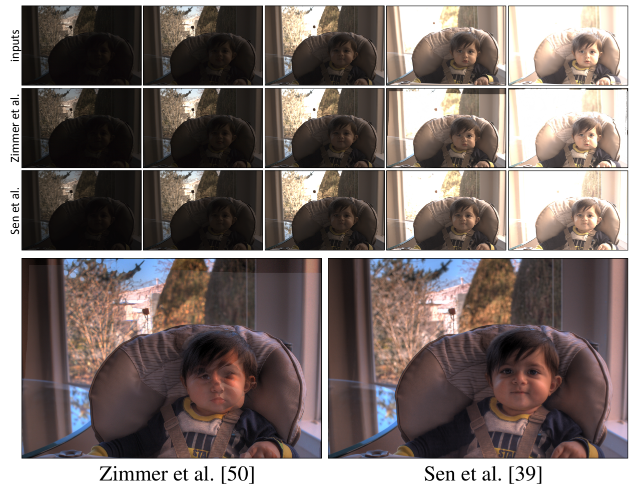


Figure 4. *The input stack in the top row is aligned to the reference (middle image) using the method of Zimmer et al. [50] to produce the sequence shown in the middle row. Because of the complex deformations of the child's face, their optical-flow method cannot align the images properly, resulting in visible artifacts in the final HDR result shown at the bottom. On the other hand, the method of Sen et al. [39] simultaneously solves for the aligned images and the final HDR result, producing aligned images that are properly aligned with the reference yet capture the detail in the different exposures.*

Patch-based optimization methods

Until recently, all HDR deghosting algorithms were either rejection or alignment methods. Then in 2012, Sen et al. [39] introduced a new kind of deghosting algorithm using patch-based synthesis that addresses the problems with both rejection and alignment methods. To do this, they propose a new equation called the *HDR synthesis equation* which embodies the goals of all reference-based HDR reconstruction algorithms: 1) to produce an HDR result that resembles the reference image in the parts where the reference is well exposed, and 2) to add detail wherever the reference is poorly exposed using well-exposed information from other images in the stack . This equation can be written as:

$$\text{Energy}(\widetilde{E}) = \sum_{p \in \text{pixels}} \big[\alpha_{\text{ref}}(p) \cdot \| h(I_{\text{ref}}(p)) - \widetilde{E}(p) \|^2 + (1 - \alpha_{\text{ref}}(p)) \cdot E_{\text{MBDS}}(\widetilde{E} \mid I_1, \ldots, I_N) \big]. \quad (3)$$

The first term states that the desired HDR image $\widetilde{E}$ should be similar to the LDR reference I_{ref} in an L_2 sense, where function $h()$

maps the LDR values of I_{ref} (which are, say, 0 to 255) to the floating-point irradiance domain: $h(I_{\text{ref}}(p)) = f^{-1}(I_{\text{ref}}(p))/t_{\text{ref}}$. This should only be done where the reference is well exposed, which is controlled by the α_{ref} term, a trapezoidal function in the pixel value domain that favors values near the middle of the pixel value range and down-weights values at the extremes.

In the parts of the image where the reference image I_{ref} is poorly exposed (indicated by $1-\alpha_{\text{ref}}$), the algorithm must draw information from the other images in the stack. Since we are assuming that there is motion, this cannot be enforced using an L_2 norm, so instead Sen et al. propose to use a metric derived from the *bidirectional similarity metric* (BDS) of Simakov et al. [41]:

$$\text{BDS}(T \mid S) = \frac{1}{|S|}\sum_{P\in S}\min_{Q\in T} d(P,Q) + \frac{1}{|T|}\sum_{Q\in T}\min_{P\in S} d(Q,P). \quad (4)$$

This original function measures the similarity (or rather the *dissimilarity*) of a pair of images (source S and target T) by computing the patchwise distance between patches P and Q. It is minimized when all patches P in source S are equal to patches Q in target T (given by the first term, known as *completeness*), and all patches Q in target T are identical to patches P in source S (second term, known as *coherence*).

The BDS metric has been used successfully in synthesis applications (Simakov et al. used it for image retargeting) because the completeness term guarantees that all of the source information is found somewhere in the target, and the coherence term ensures that the result does not have objectionable artifacts since those are not found in the source. To apply this to HDR synthesis equation in Eq. 3, Sen et al. [39] first extended BDS to use N source images from the stack in what is called the *multi-image bidirectional similarity metric* (MBDS):

$$\begin{aligned}\text{MBDS}(T \mid S_1,\ldots,S_N) = &\frac{1}{N}\sum_{k=1}^{N}\sum_{P\in S_1,\ldots,S_N} w_k(P)\min_{Q\in T} d(P,Q) + \\ &\frac{1}{|T|}\sum_{Q\in T}\min_{P\in S_1,\ldots,S_N} d(Q,P). \quad (5)\end{aligned}$$

This simply ensures that all patches from the N sources are found in the target and that all patches in the target can be found in one of the sources. The main modification for the HDR application is the weight $w_k(P)$, which ensures that only well-exposed patches are used in the completeness term. In other words, patches that are over- or under-exposed need not be included in the final result.

Finally, we must put MBDS into a form that can be plugged into Eq. 3, since BDS (and hence MBDS) operates on LDR patches. To do this, MBDS of Eq. 5 is applied to all N source images in the input stack $I_1,\ldots,I_N$ by defining an energy function that tries to keep each exposure of the estimated HDR image $\widetilde{E}$ as similar as possible to all input sources adjusted to that exposure:

$$\begin{aligned}E_{\text{MBDS}}(\widetilde{E} \mid I_1,\ldots,I_N) = &\\ &\sum_{k=1}^{N}\text{MBDS}\big(l^k(\widetilde{E}) \mid g_1^k(I_1),\ldots,g_N^k(I_N)\big), \quad (6)\end{aligned}$$

where $l^k()$ is a function that computes the LDR image from the estimated HDR irradiance map $\widetilde{E}$ that resembles the k^{th} exposure, and $g_q^k(I_q)$ maps the q^{th} LDR source to the k^{th} LDR exposure: $g_q^k(I_q) = l^k(h(I_q))$. Essentially, this function ensures that every exposure of the HDR image $l^k(\widetilde{E})$ contains only information found in the exposure-adjusted versions of all N input images so that the final HDR result will not have artifacts, and that well-exposed information from all these images can be found in the final HDR result which will add well-exposed detail from the entire stack. This energy equation is plugged in to the second term of Eq. 3, which combined with the first term satisfies the goals of reference-based HDR reconstruction as specified before.

Figure 5. *Deghosted HDR results from the patch-based optimization of Sen et al. [39] on the (cropped) input LDR images shown.*

To optimize Eq. 3, Sen et al. introduce auxiliary variables $\widetilde{I}_1,\ldots,\widetilde{I}_N$ which represent the *aligned* LDR images and is equivalent to $\widetilde{I}_k = l^k(\widetilde{E})$. They then solve for these aligned LDR images and the HDR result simultaneously by rewriting Eq. 3 as:

$$\begin{aligned}\text{Energy}(\widetilde{E},\widetilde{I}_1,\ldots,\widetilde{I}_N) = \sum_{p\in\text{pixels}}\Big[&\alpha_{\text{ref}}(p)\|h(I_{\text{ref}}(p)) - \widetilde{E}(p)\|^2 \\ &+ (1-\alpha_{\text{ref}}(p))\sum_{k=1}^{N}\text{MBDS}\big(\widetilde{I}_k \mid g_1^k(I_1),\ldots,g_N^k(I_N)\big) \\ &+ (1-\alpha_{\text{ref}}(p))\sum_{k=1}^{N} w_k(p)\|h(\widetilde{I}_k(p)) - \widetilde{E}(p)\|^2\Big]. \quad (7)\end{aligned}$$

where the first term is the same as before, the second term has been modified to do the MBDS on the aligned LDR images $\widetilde{I}_k$, and the last term has been introduced to enforce the relationship between the aligned LDR images and the final HDR result based on the merging process of Eq. 1 (the $w_k(p)$ function simply applies the normalized "triangle" merging weights proposed by Debevec and Malik [6]). Eq. 7 can now be optimized with a two-stage alternating minimization algorithm that solves for the HDR result $\widetilde{E}$ and the aligned LDR images $\widetilde{I}_1,\ldots,\widetilde{I}_N$ simultaneously:

Stage 1: The algorithm first optimizes for the aligned LDR images $\widetilde{I}_1,\ldots,\widetilde{I}_N$ using a bidirectional search-and-vote process [41] accelerated by PatchMatch [2], thereby minimizing the second term in Eq. 7. This adds information into each aligned LDR images from all the other images in the stack in order to handle things like disocclusions that can reveal previously unseen content at that exposure level.

Stage 2: The algorithm optimizes for $\widetilde{E}$ by merging the aligned images $\widetilde{I}_1,\ldots,\widetilde{I}_N$ together with Eq. 1 (minimizing the third term of Eq. 7), and then *injects* the parts of the reference image that are well-exposed into the result (minimizing the first term of Eq. 7). The resulting irradiance map estimate $\widetilde{E}$ is used in the next iteration as the initial target for the search-and-vote process, which forces the aligned images to be aligned with the reference.

As is common for patch-based algorithms [2, 5, 41], the algorithm is performed at multiple scales, starting at the coarsest resolution and gradually working to the finest, to keep the optimization from settling on a local minimum. Once finished, the algorithm returns both the desired HDR image $\widetilde{E}$ as well as the "aligned" images at each exposure $\widetilde{I}_1,\ldots,\widetilde{I}_N$. Results produced with this algorithm is shown in Fig. 5 and throughout the paper.

Hu et al. [19] later proposed a related patch-based HDR reconstruction algorithm which solves for the aligned LDR images $\widetilde{I}_1,\ldots,\widetilde{I}_N$ by minimizing the following energy function:

$$\mathrm{Energy}(\widetilde{I}_k, g^k_{\mathrm{ref}}, \mathbf{u}) = C_r(\widetilde{I}_k, I_{\mathrm{ref}}, g^k_{\mathrm{ref}}) + C_t(\widetilde{I}_k, I_k, \mathbf{u}). \tag{8}$$

where $\mathbf{u}$ is the displacement field that warps image I_i to match the reference. The first term, $C_r()$, enforces that the final aligned image $\widetilde{I}_i$ actually matches the reference:

$$\begin{aligned} C_r(\widetilde{I}_k, I_{\mathrm{ref}}, g^k_{\mathrm{ref}}) = \sum_{p\in\mathrm{pixels}} \Big(& \|\widetilde{I}_k(p) - g^k_{\mathrm{ref}}(I_{\mathrm{ref}}(p))\|^2 + \\ & +\gamma\,\|\nabla\widetilde{Z}_k(p) - \nabla g^k_{\mathrm{ref}}(I_{\mathrm{ref}}(p))\|^2 \Big), \end{aligned} \tag{9}$$

where the reference image is mapped to the k^{th} exposure by function $g^k_{\mathrm{ref}}()$. Unlike Sen et al. [39], this method assumes that the camera response curve is not known so it solves for $g^k_{\mathrm{ref}}()$ on the fly. Since this can introduce errors, they leverage earlier work that successfully matches patches with different exposure levels by comparing both the color *and* the gradient terms [5], which gives them more flexibility during the matching process.

The second term in Eq. 8 enforces that the aligned image $\widetilde{I}_k$ should resemble the original input image I_k after being warped by deformation field $\mathbf{u}$ to try to keep it free of artifacts:

$$\begin{aligned} C_t(\widetilde{I}_k, I_k, \mathbf{u}) = \frac{1}{q}\sum_p \Big(& \|P_{\widetilde{I}}(p) - P_{I_k}(p+\mathbf{u}(p))\|^2 + \\ & +\gamma\,\|\nabla(P_{\widetilde{I}}(p) - \nabla P_{I_k}(p+\mathbf{u}(p))\|^2 \Big), \end{aligned} \tag{10}$$

where q is a normalization factor and $P_I(p)$ represents a patch in image I around pixel p. To minimize Eq. 8, Hu et al. propose a three-stage iterating optimization which, like the previous method of Sen et al. [39], is also performed at multiple scales:

Stage 1: First, the algorithm estimates all $g^k_{\mathrm{ref}}()$ functions by comparing the intensity image histograms [14] at the coarsest level of the pyramid. It also estimates the displacement field $\mathbf{u}$ in Eq. 10 using generalized PatchMatch [3].

Stage 2: The algorithm then refines the aligned estimate $\widetilde{I}_k$ by averaging information between $g^k_{\mathrm{ref}}(I_{\mathrm{ref}}(p))$ and $I_k(p+\mathbf{u}(p))$, using weights that account for how over-exposed or under-exposed the reference would be in each region. Since it has to do this for both the color and gradient domains, the algorithm essentially must solve a screened Poisson equation [5].

Stage 3: Finally, the updated $\widetilde{I}_k$ is used to correct $g^k_{\mathrm{ref}}()$. As the algorithm iterates and moves to finer levels, u is linearly upsampled but $g^k_{\mathrm{ref}}()$ is left the same.

Discussion

Advantages: Patch-based optimization algorithms are fundamentally different than previous rejection or alignment methods. Unlike rejection methods, patch-based methods can bring HDR information from different pixels across the stack and so can handle dynamic HDR content. Furthermore, unlike alignment methods, they can *synthesize* missing information due to occlusions or camera motion, and are able to handle complex deformations to produce better aligned images and therefore better HDR results. In fact, algorithms like that of Sen et al. [39] effectively combine both alignment *and* rejection in the inner loop of the optimization.

This has made patch-based HDR reconstruction methods the most successful HDR deghosting algorithms to date. For example, Tursun et al. [45] conducted a user study to examine HDR quality and found that the methods of Sen et al. [39] and Hu et al. [19] ranked first and second, respectively, with a sizeable margin over other state-of-the-art methods. More recently, Karaduzovic-Hadziabdic et al. [25] also found that the method of Sen et al. [39] outperformed others for most scenes. Finally, while most algorithms require that the image stack be captured only by varying the exposure time between images (changing settings like the aperture affects the depth-of-field, which makes images difficult to align or deghost), Sen et al. [39] showed that their patch-based method is able to handle changes in depth-of-field automatically, thereby enabling longer exposures than could be done by simply changing the exposure time.

Disadvantages: Although patch-based HDR reconstruction algorithms produce high-quality results, they are computationally expensive and require fairly long computation times. For example, Tursun et al. [45] report that the algorithms of Sen et al. [39] and Hu et al. [19] took an average 209.78 and 230.36 seconds, respectively. This algorithmic complexity makes them challenging to port to mobile devices, at least in their current form. Finally, although their quality is superior to all previous methods, these algorithms still produce artifacts, especially in scenes with very complex motion and extreme dynamic range. The Sen et al. [39] algorithm, for example, tends to leave some noise in the dark regions, while the Hu et al. [19] method sometimes overblurs detail or produces artifacts where the reference is poorly exposed.

Learning-based methods

The most recent kind of stack-based, HDR deghosting method is the *learning-based* method proposed by Kalantari and Ramamoorthi [23], which uses deep learning to reconstruct the final HDR image (see Fig. 6). Their approach assumes that the stack is composed of three differently exposed images I_1, I_2, I_3 with the middle image I_2 as the reference. After using optical flow to align the two other images to it, they replace the standard merging step of Eq. 1 with a deep learning network that removes the ghosting artifacts caused by misalignments from the optical flow. Specifically, they propose and analyze three different architectures based on a four-layer convolutional neural network (CNN):

1. The *direct* approach, which takes in the aligned images as input and outputs the HDR result directly.
2. The *weight estimator* (WE), which outputs the three blending weights w_1, w_2, w_3 used in Eq. 1 to compute the final result.
3. The *weight and image estimator* (WIE), which computes refined versions of the two aligned images as well as the weights used to merge them together.

To train their systems, they built a dataset with image stacks of different dynamic HDR scenes. Since they need ground-truth HDR results, they capture a static set (where the subjects are still) as well as a dynamic set (where the subject and camera can move) for every scene. The middle image of the static set is then used as the

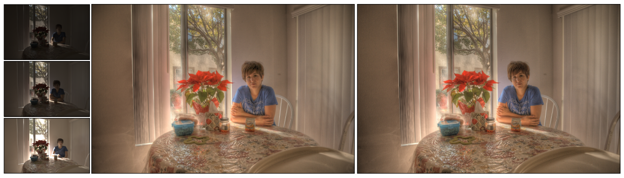


Figure 6. Results of the learning-based method of Kalantari and Ramamoorthi [23] from the image stack shown on the left. For completeness, we show the comparison against the patch-based synthesis algorithm of Sen et al. [39]. Images courtesy of Kalantari and Ramamoorthi [23].

middle image of the dynamic set to create a new stack is used as input to the network. The training itself is done using tonemapped HDR images in the loss function, which they found worked better than using the HDR ground truth directly. The authors report that while the WIE network produces the lowest numerical error, detail is better preserved with the WE architecture.

Discussion

Advantages: The results of this learning-based approach are impressive, although the difference with the results of other state-of-the-art algorithms, such as patch-based synthesis [39], is relatively small. However, the biggest advantage of this algorithm is its speed: the authors report that it takes only 30 seconds to produce an image. For this reason, it is highly likely that future HDR reconstruction algorithms, especially for mobile devices, will be learning based.

Disadvantages: As with all learning algorithms, this one needs a large dataset to learn to work robustly for a diverse set of scenes. As such, it is difficult to know what scenes it will not work for, or even what the failure cases will be. Furthermore, since their loss function tests against the tonemapped ground truth, they need a differentiable tonemapping function to perform backpropagation. This severely limits the kinds of tonemappers that can be used. In this work they use the μ-law function, which is a common range compressor for audio processing but not for tone mapping as it produces results that appear "washed out" and of low contrast.

Extensions to HDR video

The stack-based deghosting algorithms presented so far are normally used to capture still images, but some of them can be modified to reconstruct high-dynamic range video captured with alternating exposures. In this application, the camera takes a sequence of differently-exposed images, usually in an alternating pattern of two or three exposures, which the algorithm must use to produce an HDR sequence of frames. Clearly, the reference-free rejection methods cannot be used for this application, since they do not produce an image that adheres to a real image.

The first to tackle such a problem was Kang et al. [24], whose method using optical flow was described earlier in this paper. Later, Mangiat and Gibson [28, 29] proposed to overcome the problems with optical flow by using a block-based motion estimation approach, filtered to remove block boundary artifacts. Currently, the state-of-the-art approach for stack-based HDR video is the algorithm of Kalantari et al. [22], which extends the patch-based algorithm of Sen et al. [39] to make the HDR video streams temporally coherent (see Fig. 7). To do this, they modify the HDR image synthesis equation shown in Eq. 3 to perform a bidirectional similarity between adjacent frames to maintain temporal coherence. They also use optical flow during the optimization to constrain the patch-based search.

Figure 7. Stack-based acquisition for HDR video. The top row shows the captured frames with three alternating exposures. The bottom row shows the HDR video result reconstructed by the patch-based algorithm of Kalantari et al. [22].

Future directions and conclusion

Although HDR imaging has improved tremendously in the past 20+ years, there is still much work to be done. Deghosting algorithms could still be improved, especially for scenes with very complex motion and extreme dynamic range. Exploring other machine learning approaches, or combinations of machine learning and patch-based synthesis, seems like a promising way to do this. Furthermore, since machine learning has been shown to be very successful in image synthesis applications, researchers have started to explore the idea of single-image HDR where the HDR content is hallucinated [8]. Along with work on new sensors and camera technology, this would enable broader dynamic-range capture with each image and would reduce the need for stack-based approaches.

In conclusion, we have presented the basics of stack-based HDR imaging and discussed four different approaches to perform the deghosting that occurs in dynamic scenes: rejection, alignment, patch-based synthesis, and learning-based methods. Of these, patch-based synthesis approaches provide high-quality results but are expensive to compute, while machine-learning methods seem to offer a good combination of quality and speed.

Acknowledgments

The author's work presented was sponsored in part by National Science Foundation IIS grants #08-45396, 13-21168, and 16-19376. We also thank researchers in HDR imaging for publishing their work and helping to make this a vibrant area of research.

References

[1] A. Badki, N. K. Kalantari, and P. Sen, "Robust radiometric calibration for dynamic scenes in the wild," in *IEEE ICCP*, Apr. 2015.
[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24:1–24:11, Jul. 2009.
[3] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized PatchMatch correspondence algorithm," in *ECCV*, Sep. 2010, pp. 29–43.
[4] L. Bogoni, "Extending dynamic range of monochrome and color images through fusion," in *IEEE ICPR*, 2000, pp. 3007–3016.
[5] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 82:1–82:10, Jul. 2012.

[6] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *ACM SIGGRAPH*, 1997.
[7] A. Eden, M. Uyttendaele, and R. Szeliski, "Seamless image stitching of scenes with large motions and exposure differences," in *IEEE CVPR*, vol. 2, 2006, pp. 2498–2505.
[8] G. Eilertsen, J. Kronander, G. Denes, R. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 178:1–178:15, Nov. 2017.
[9] O. Gallo, A. Troccoli, J. Hu, K. Pulli, and J. Kautz, "Locally non-rigid registration for mobile HDR photography," in *IEEE CVPR Workshops*, Jun. 2015, pp. 48–55.
[10] O. Gallo, N. Gelfand, W.-C. Chen, M. Tico, and K. Pulli, "Artifact-free high dynamic range imaging," in *IEEE ICCP*, 2009.
[11] O. Gallo and P. Sen, "Stack-based algorithms for HDR capture and reconstruction," in *High Dynamic Range Video*, F. Dufaux, P. L. Callet, R. Mantiuk, and M. Mrak, Eds. London: Academic Press, 2016, ch. 3, pp. 85–119.
[12] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free HDR reconstruction," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 201:1–201:10, Nov. 2013.
[13] T. Grosch, "Fast and robust high dynamic range image generation with camera and object movement," in *International Symposium on Vision, Modeling and Visualization*, 2006, pp. 277–284.
[14] M. D. Grossberg and S. K. Nayar, "Determining the camera response from images: What is knowable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 11, pp. 1455–1467, Nov 2003.
[15] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski, "Non-rigid dense correspondence with applications for image enhancement," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 70:1–70:10, Jul. 2011.
[16] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 192:1–192:12, Nov. 2016.
[17] Y. Heo, K. M. Lee, S. U. Lee, Y. Moon, and J. Cha, "Ghost-free high dynamic range imaging," in *ACCV*, vol. 4, 2010, pp. 486–500.
[18] J. Hu, O. Gallo, and K. Pulli, "Exposure stacks of live scenes with hand-held cameras," in *ECCV*, 2012.
[19] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *IEEE CVPR*, 2013.
[20] K. Jacobs, C. Loscos, and G. Ward, "Automatic high-dynamic-range image generation for dynamic scenes," *IEEE Comput. Graph. Appl. Mag.*, vol. 28, no. 2, pp. 84–93, Apr 2008.
[21] T. Jinno and M. Okuda, "Motion blur free HDR image acquisition using multiple exposures," in *IEEE ICIP*, Oct. 2008, pp. 1304–1307.
[22] N. K. Kalantari, E. Shechtman, C. Barnes, S. Darabi, D. B. Goldman, and P. Sen, "Patch-based high dynamic range video," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 202:1–202:8, Nov. 2013.
[23] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144:1–144:12, Jul. 2017.
[24] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High dynamic range video," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 319–325, Jul. 2003.
[25] K. Karauzovi-Hadiabdi, J. Hasi Telalovi, and R. K. Mantiuk, "Assessment of multi-exposure HDR image deghosting methods," *Comput. Graph.*, vol. 63, no. C, pp. 1–17, Apr. 2017.
[26] E. A. Khan, A. O. Akyüz, and E. Reinhard, "Ghost removal in high-dynamic-range images," in *IEEE ICIP*, 2006.
[27] B. C. Madden, "Extended intensity range imaging," University of Pennsylvania, Tech. Rep., 1993.
[28] S. Mangiat and J. Gibson, "High dynamic range video with ghost removal," in *SPIE*, 2010.
[29] ——, "Spatially adaptive filtering for registration artifact removal in HDR video," in *IEEE ICIP*, Sep. 2011.
[30] S. Mann and R. Picard, "Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures." in *Society for Imaging Science and Technology (IS&T)*, 1995, pp. 442–448.
[31] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *Pacific Conference on Computer Graphics and Applications*, 2007, pp. 382–390.
[32] T.-H. Min, R.-H. Park, and S. Chang, "Histogram based ghost removal in high dynamic range images," in *IEEE ICME*, 2009, pp. 530–533.
[33] S. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *IEEE CVPR*, 2000, pp. 472 – 479.
[34] T.-H. Oh, J.-Y. Lee, and I. S. Kweon, "Robust high dynamic range imaging by rank minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1219–1232, Jun. 2015.
[35] F. Pece and J. Kautz, "Bitmap movement detection: HDR for dynamic scenes," in *The Conference on Visual Media Production (CVMP)*, Nov. 2010, pp. 1–8.
[36] S. Raman and S. Chaudhuri, "Reconstruction of high contrast images for dynamic scenes," *The Visual Computer*, vol. 27, no. 12, pp. 1099–1114, 2011.
[37] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.
[38] P. Sen and C. Aguerrebere, "Practical high dynamic range imaging of everyday scenes: Photographing the world as we see it with our own eyes," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 36–44, Sep. 2016.
[39] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 203:1–203:11, Nov. 2012.
[40] D. Sidibe, W. Puech, and O. Strauss, "Ghost detection and removal in high dynamic range images," in *European Signal Processing Conference (EUSIPCO)*, Aug. 2009, pp. 2240–2244.
[41] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *CVPR*, Jun. 2008, pp. 1–8.
[42] A. Srikantha and D. Sidibé, "Ghost detection and removal for high dynamic range images: Recent advances," *Image Commun.*, vol. 27, no. 6, pp. 650–662, Jul. 2012.
[43] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 41:1–41:10, Jul. 2011.
[44] A. Tomaszewska and R. Mantiuk, "Image registration for multi-exposure high dynamic range image acquisition," in *Intl. Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2007.
[45] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "The state of the art in HDR deghosting: A survey and evaluation," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 683–707, May 2015.
[46] G. Ward, "Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures," *Journal of Graphics Tools*, vol. 8, no. 2, pp. 17–30, 2003.
[47] S. Wu, S. Xie, S. Rahardja, and Z. Li, "A robust and fast anti-ghosting algorithm for high dynamic range imaging," in *IEEE ICIP*, Sep. 2010, pp. 397–400.
[48] L. Zhang, A. Deshpande, and X. Chen, "Denoising vs. deblurring: HDR imaging techniques using moving cameras," in *IEEE CVPR*, Jun. 2010, pp. 522–529.
[49] W. Zhang and W.-K. Cham, "Reference-guided exposure fusion in dynamic scenes," *Journal of Visual Communication and Image Representation*, vol. 23, no. 3, pp. 467–475, Apr. 2012.
[50] H. Zimmer, A. Bruhn, and J. Weickert, "Freehand HDR imaging of moving scenes with simultaneous resolution enhancement," in *Eurographics*, vol. 30, no. 2, Apr. 2011, pp. 405–414.

Author Biography

Pradeep Sen is an Associate Professor in the Department of Electrical and Computer Engineering at the University of California, Santa Barbara. He received his B.S. from Purdue University and his M.S. and Ph.D. from Stanford University. His core research is in the areas of computer graphics, computational image processing, and computer vision. He is the recipient of an NSF CAREER award and a senior member of IEEE.