HIGH DYNAMIC RANGE IMAGING USING CAMERA ARRAYS

Kalpana Seshadrinathan and Oscar Nestares

Intel Labs, Intel Corporation, Santa Clara, CA - 95054.

ABSTRACT

While High Dynamic Range (HDR) image generation using computational approaches has seen widespread adoption in consumer photography, computational approaches for HDR video creation are only just beginning to emerge. In this paper, we present an algorithm for HDR video generation using short baseline camera arrays (cameras placed close together) with varying exposures. Our main contributions are an algorithm for disparity estimation that specifically targets short baseline camera arrays and an alignment algorithm that minimizes visual artifacts even in the presence of disparity errors. Finally, we present results using our algorithm on real world videos captured using a camera array to demonstrate the success of this application.

Index Terms— High Dynamic Range (HDR) imaging, camera arrays, multi-exposure, fusion, disparity estimation.

1. INTRODUCTION

High Dynamic Range (HDR) imaging attempts to reproduce a greater dynamic range of luminosity than is possible with standard digital imaging or photographic techniques. Several computational approaches have been proposed to increase the dynamic range of a *still image* captured by a camera including time-sequential captures of multiple exposures [1], spatially varying pixel exposures [2] etc. HDR image creation from multiple exposures, in particular, has been studied extensively where the multiple exposures are first aligned and then merged to generate an HDR image [3]. Various alignment techniques (global registration, optical flow) and merging strategies (radiometric calibration and true HDR image creation, tone mapping, exposure fusion, deghosting) have been proposed in the literature to solve this problem [3].

HDR *video* creation using similar approaches has also been studied, although not as extensively [4, 5, 6, 7]. Spatially varying pixel exposures have been commercially deployed in the Sony Xperia Z phone with different exposures for every two lines of pixels. The HTC One phone, Panasonic HC-WX970 consumer camera and HDRx technology in RED's EPIC professional camera capture interleaved frames with high and low exposures, typically at twice the frame rate, to generate HDR video. Multiple synchronized cameras can be used to generate HDR video by assigning different exposures to different cameras and combining the resulting videos [8]. This approach requires additional cameras and synchronized capture, but has the advantage of lack of motion across the different camera images and of recovering artifact-free low dynamic range video from any one of the cameras even in HDR mode. However, the challenge is in dealing with the different perspective of each camera which requires disparity computation for alignment [9]. While image alignment techniques developed for time sequential, multi-exposure HDR can be used, they are suboptimal for camera arrays since they do not exploit information about the camera geometry. Further, while disparity estimation for multi-view stereo (MVS) has been studied extensively in computer vision, assignment of different exposures to individual cameras violates the intensity conservation assumption made by many disparity estimation algorithms [10].

HDR video using camera arrays have been described in a few papers in the literature. One of the first approaches for HDR imaging using large camera arrays did not account for disparity (the scene was far enough away resulting in less than 0.5 pixels of disparity) and could not handle objects close to the camera [8]. HDR image construction from stereo images was studied in [11], but this approach uses only two cameras and fails when one of the images suffers from significant saturation. HDR image creation from multiple views with multiple exposures was described in [12]. MVS is first applied using an exposure invariant similarity statistic on a reference image and the resulting alignment is used to estimate the radiometric response function and map the images to a common radiance space in [12]. MVS is then applied in this space for each input image to generate a per-view depth map, which is then used to align and merge the images. The resulting method is computationally complex and the results presented in the paper focus on depth map generation rather than creation of the merged HDR image.

In this paper, we present an algorithm for HDR video generation using short baseline camera arrays (cameras placed close together) with varying exposures. Our main contributions are an algorithm for disparity estimation that specifically targets short baseline camera arrays and an alignment algorithm that minimizes visual artifacts even in the presence of disparity errors. Our disparity estimation algorithm exploits the fact that the presence and extent of occluded regions are limited when baselines are short and only occur when objects are extremely close to the camera. Therefore, these systems do not require per-view depth maps allowing us to design computationally efficient solutions. We then present an algorithm to align and merge the input videos that minimizes visual artifacts in the output video since even the most sophisticated disparity estimation algorithms can be fragile [13]. Finally, we adopt the popular exposure fusion approach for merging since it offers several advantages such as lack of radiometric calibration, lack of higher bit-depth representations that are expensive in many computer architectures and lack of tonemapping that can alter the appearance of the image [14]. We describe the algorithm for computing disparity from a short baseline camera array in Section 2. We then describe the proposed alignment and merging algorithm in Section 3. Finally, we present results obtained using our proposed algorithm on real video sequences captured using a camera array and comparison to other state-of-the-art methods in Section 4.

2. DISPARITY ESTIMATION

We outline an algorithm to compute disparity that specifically targets short baseline camera arrays with variable exposure assignments. Traditional multi-baseline stereo (MBS) algorithms compute a dis-

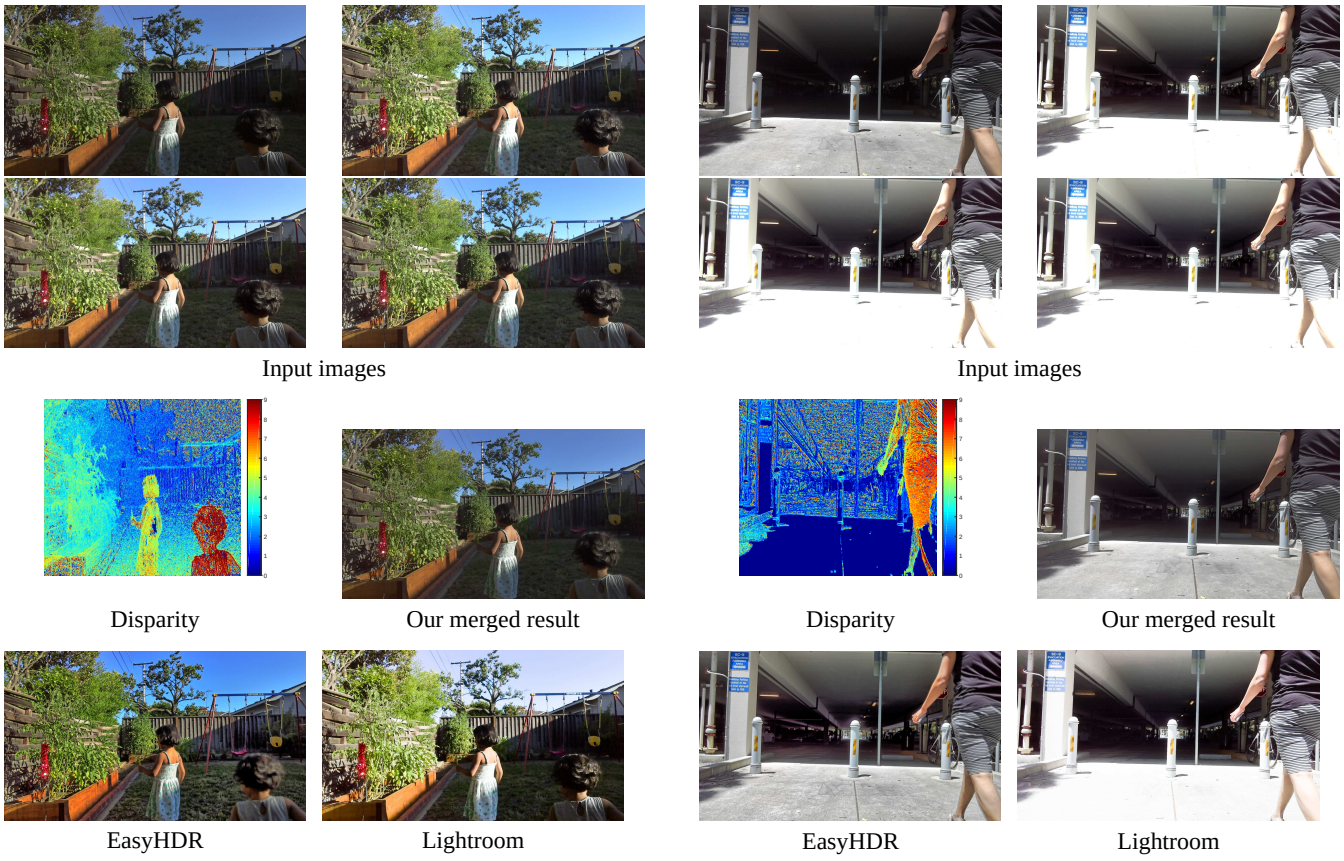

Fig. 1: Example of HDR merging: Input images in the top two rows are captured using different exposure times for each camera that are spaced 1 stop apart (top left is the reference image). Third row shows the disparity map estimated at the finest scale and our merged result. Merged results using EasyHDR and Adobe Lightroom are also shown for comparison.

Fig. 2: Example of HDR merging: Input images in the top two rows are captured using 3 different exposure values spanning approximately 2.5 stops (top left is the reference image). Third row shows the disparity map estimated at the finest scale and our merged result. Merged results using EasyHDR and Adobe Lightroom are also shown for comparison.

parity map for one of the images from the camera array, denoted as the reference image, by defining a matching criterion using all stereo pairs of images involving the reference image [15]. Even if the matching criterion is made robust to exposure variation across the images, this approach fails when the reference camera suffers from saturation since all stereo pairs utilized include the reference. This issue is addressed in [12] by computing a disparity map for every image in the array, which can handle occluded regions that are visible in only some of the cameras. However, for HDR imaging using a short baseline camera array, we argue that it is unnecessary to compute a disparity map for every view or even build a 3D volumetric model of the scene which can be computationally prohibitive [16]. Instead, we propose a method that can accurately estimate disparity for a chosen reference camera, even in the presence of saturation, and exploits the fact that the presence and extent of occluded regions are small in short baseline camera arrays. In particular, when more than two cameras are used, disparity of regions that are saturated in the reference camera can be estimated by matching them in the other cameras and exploiting the known geometric relationship between the cameras. Note that this approach assumes that the image feature is visible in all the cameras, which is typically the case with short baseline camera arrays since they suffer from few occluded regions. With this intuition, we propose utilizing all camera pairs (and not just those involving the reference camera) along with saturation modeling for disparity estimation.

We assume that the cameras are synchronized and geometric calibration (extrinsic and intrinsic) is available for all the cameras in the array. We consider planar camera arrays and images that are rectified such that conjugate epipolar lines are parallel to one of the image axes (horizontal or vertical) [17]. Consider a set of rectified images from a planar camera array and denote them using $I_k(\mathbf{x}, t), k \leq N$, where $I_k(\mathbf{x}, t)$ represents the image from k^{th} camera captured at time instant t. Here, $\mathbf{x} = (x, y)$ represents orthogonal axes in a 3D reference coordinate frame that are aligned with the columns and rows of the images respectively. We do not assume any knowledge of camera response curves or exposure values assigned to each camera. We denote the baseline of camera k using $\mathbf{B}_k = [b_k^x, b_k^y], k \leq N$ and set the baseline of the reference camera B_{ref} to $\mathbf{0}$ without any loss of generality. The choice of the reference may be arbitrary or pre-determined based on the improved quality of one of the cameras in the array (as determined by resolution, signal-to-noise ratio etc.). Let B_{max} represent the longest baseline (horizontal or vertical) in the array and let $\mathbf{R}_k$ represent the baseline ratio for camera k given by $R_k^x = \frac{B_k^x}{B_{\text{max}}}$ and a similar definition for R_k^y. Traditional MBS algo-

rithms attempt to minimize an error function f at a pixel $\mathbf{x}, t$ over a window M for a disparity range $\{d_i, i = 1, 2, \ldots, D\}$ [15].

$$d^*(x,t) = \underset{d_i}{\operatorname{argmin}} \sum_{k \neq ref} \sum_{\mathbf{m} \in M} f\left[I_{\text{ref}}(\mathbf{x}+\mathbf{m}, t), I_k(\mathbf{x}+\mathbf{m}+\mathbf{R}_k d_i, t)\right] \quad (1)$$

We modify this error function to include all camera pairs (not just those involving the reference camera) and to model saturation as follows.

$$d^*(x,t) = \underset{d_i}{\operatorname{argmin}} \sum_{k} \sum_{j>k} \sum_{m \in M} w_{j,k}(\mathbf{x}, t, \mathbf{m}, d_i) \; f\left[I_j(\mathbf{x}+\mathbf{m}+\mathbf{R}_j d_i, t), I_k(\mathbf{x}+\mathbf{m}+\mathbf{R}_k d_i, t)\right] \quad (2)$$

By incorporating all camera pairs into the error function, disparity can be estimated from the remaining cameras even if the reference camera is saturated. Although we include additional camera pairs in the error function, our approach is computationally less expensive than computing a per-view disparity map. With four cameras, our approach performs 6 evaluations of the cost function per-pixel, whereas MBS would require 3 evaluations per pixel for each of the 4 views resulting in twice the computational cost in unoptimized implementations.

The weight $w_{j,k}(\mathbf{x}, t, \mathbf{m}, d_i)$ models saturation as follows:

$$w_{j,k}(\mathbf{x}, t, \mathbf{m}, d_i) = w_j(\mathbf{x}, t, \mathbf{m}, d_i) w_k(\mathbf{x}, t, \mathbf{m}, d_i)$$
$$w_j(\mathbf{x}, t, \mathbf{m}, d_i) = \begin{cases} \frac{I_j(\mathbf{x}+\mathbf{m}+\mathbf{R}_j d_i, t) - L}{T_{\text{L}} - L}, & \text{if } I_j(\mathbf{x}+\mathbf{m}+\mathbf{R}_j d_i, t) < T_{\text{L}} \\ \frac{H - I_j(\mathbf{x}+\mathbf{m}+\mathbf{R}_j d_i, t)}{H - T_{\text{H}}}, & \text{if } I_j(\mathbf{x}+\mathbf{m}+\mathbf{R}_j d_i, t) > T_{\text{H}} \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

Here, T_{L} and T_{H} represent thresholds beyond which the input image is assumed to be too dark or bright and L, H represent the lowest and highest possible pixel values.

We implemented the disparity estimation algorithm in a multi-scale manner using a Gaussian pyramid [18] primarily because a pyramidal decomposition has been shown to enable seamless blending for HDR [14, 19]. Disparity estimation was performed at integer precision on the Gaussian pyramid using a coarse-to-fine approach, where the disparity map at each scale is upscaled and doubled in value for use as the midpoint of the search range at the next finer scale. At the finest scale, disparity estimation is performed directly on the input images. We used a square window of size 7 at all scales utilizing the census transform as the cost function f, which is a fast, local binary pattern based approach that is robust to intensity variations [20]. Cost was aggregated using the sum across all color channels and saturation was modeled by applying Eq. (3) on the grayscale value of each pixel. We also experimented with using sum of absolute differences (SAD) after color matching the input images as the cost function. While the disparity maps generated using this approach were smoother, they require extended precision to store the color matched images due to the variable exposure assignments resulting in higher computational cost of disparity estimation.

Disparity maps generated by our method at the finest scale are shown in Figures 1 and 2. Note that our algorithm is able to estimate disparity even in regions where the reference image is saturated such as the foreground regions of the lawn and the fence in Figure 1. Further, the disparity map is quite noisy in the smooth, textureless regions of the scene which is to be expected since we do not perform any global regularization or utilize adaptive window sizes to aid multi-view matching in these areas. However, these areas do not pose significant issues in HDR merging, despite errors in the disparity map, precisely because they are smooth and textureless. Since our ultimate objective is HDR merging rather than accurate disparity estimation, we are able to utilize the fast and local disparity estimation proposed here along with disparity tolerant alignment to generate good results as shown in Section 4.

EasyHDR Lightroom Our method

Fig. 3: Cropped results: Ghosting artifacts are visible around the head of the girl in EasyHDR. Lightroom is unable to recover the full dynamic range of the scene when deghosting is turned on and merging artifacts are visible in the ceiling where there is a sharp brightness change.

3. ALIGNMENT AND MERGING

We use the disparity map estimated for the reference image to align each image from the camera array to the reference view. Disparity estimation algorithms are error prone in textureless regions, depth discontinuities and around occlusions [10], which need to be accounted for during warping. We first present an alignment algorithm that is tolerant to disparity errors and avoids the need for any deghosting algorithms during merging. While radiometric calibration has been used to generate a true HDR image that can then be tone mapped for display, estimation of the radiometric function can be computationally expensive and varies for each camera [12]. Tone mapping can also be computationally expensive and the visual quality and appearance of tone mapped images can be quite subjective [3]. Exposure fusion which can directly generate an output image without the intermediate HDR image has gained popularity due to its pleasing results and favorable computational complexity and we use this approach for merging the aligned images [14].

Let $I_{\text{ref}}(\mathbf{x}, t)$ denote the reference image and let $d^*(\mathbf{x}, t)$ represent the corresponding disparity map. Although warping can be performed on a per-pixel basis, this can result in visible visual artifacts in the aligned images due to errors in the disparity map. We propose warping using overlapping windows which ameliorates visual artifacts and avoids the need for an explicit deghosting step in the merging. To warp each input image $I_k(\mathbf{x}, t)$ to the reference image, we copy a window M surrounding each pixel in the input image to the warped image (initialized to zero) using the corresponding disparity value.

$$G_k(\mathbf{x}+\mathbf{m}, t) := G_k(\mathbf{x}+\mathbf{m}, t) + I_k[\mathbf{x}+\mathbf{m}+d^*(\mathbf{x}, t)\mathbf{R}_k, t], \mathbf{m} \in M \quad (4)$$

We also keep a count $C(\mathbf{x}, t)$ of the warped pixels (initialized to

zero) so that they can be normalized at the end of this operation.

$$C_k(\mathbf{x}+\mathbf{m},t) := C_k(\mathbf{x}+\mathbf{m},t)+1, \mathbf{m} \in M \quad (5)$$

The warped image can then be generated by normalization: $H_k(\mathbf{x},t) = G_k(\mathbf{x},t)/C_k(\mathbf{x},t)$. Our proposed warping algorithm essentially averages the warped results from a neighborhood around each pixel which minimizes visual artifacts caused by isolated disparity errors. Further, warping is performed in a multi-scale manner in our implementation using a square window of size 7 at all scales which produces seamless alignment. At the coarsest scale, warping is performed on the output of the Gaussian pyramid using the disparity map computed at the coarsest scale. At the remaining scales, warping is performed on the corresponding Laplacian pyramid using the disparity map computed at the corresponding scale.

The aligned pyramidal representations are then merged to generate the output image. Merging is performed on the Gaussian pyramid output at the coarsest scale and the Laplacian pyramid output at all other scales. The resulting Laplacian pyramid is then inverted to obtain the output image. To further minimize artifacts that may occur due to misalignment while merging multiple images, we combine the reference image with just one of the other images at each pixel in the pyramidal representation. The other image is chosen at each pixel location as the one which does not suffer from saturation and has the maximum pyramid coefficient magnitude. We use this choice since higher magnitudes correspond to better sharpness in the Laplacian pyramid representation.

$$k^*(\mathbf{x},t) = \underset{k \neq \text{ref}}{\operatorname{argmax}} |H_k(\mathbf{x},t)| \quad \text{if } H_k(\mathbf{x},t) \geq T_\text{L}, H_k(\mathbf{x},t) \leq T_\text{H} \quad (6)$$

We use a linearly weighted sum at each pixel to merge the warped images with the reference image to improve its dynamic range. The weights at each pixel location are chosen as a function of how well exposed the reference image is and we employ a Gaussian function similar to [14]. The mean of the Gaussian was chosen to be the midpoint of the intensity scale and the smoothness of the function helps minimize visual artifacts in regions with sharp intensity transitions. Reference pixels near this mean are assigned the highest weights and the weights decrease for reference pixels closer to the endpoints (likely saturated).

$$\alpha(\mathbf{x},t) = e^{-\frac{[I_\text{ref}(\mathbf{x},t)-\mu]^2}{2\sigma^2}}$$
$$M(\mathbf{x},t) = \alpha(\mathbf{x},t) I_\text{ref}(\mathbf{x},t) + [1-\alpha(\mathbf{x},t)] H_{k^*}(\mathbf{x},t) \quad (7)$$

The different color channels are merged using the proposed algorithm independently and the resulting pyramidal representations are inverted to obtain the output color image. Weights for merging are computed using the grayscale values of the input images in Eq. (7) and the resulting weight map is decomposed using a Gaussian pyramid to obtain weights at each scale [14].

4. RESULTS

We tested our method using an array of 4 cameras in a 2×2 square configuration with a baseline of 5mm between adjacent cameras. We used the Aptina AR0261 image sensor that can generate images with a resolution of 1920×1080 at 30 frames per second. An FPGA board was used to send a common synchronization pulse to the four cameras to trigger synchronized image or video capture. We geometrically calibrated the camera array using checkerboard patterns and the images were rectified for use as inputs to our algorithm.

Figures 1 and 2 show examples of the application of our algorithm on images captured with varying exposures. The merged image is able to reproduce the dynamic range of the scene better and provide detail in both the dark and bright regions of the scene. Note that there are very few visible artifacts in the merged image due to our proposed alignment and merging algorithms that are tolerant to disparity errors via overlapping windows and multi-scale computation. We also compare our results to those generated by two state-of-the-art HDR merging methods: EasyHDR and Adobe Lightroom. These methods are designed for frame sequential HDR and use generic frame alignment, merging and deghosting that are suboptimal for camera arrays. In particular, several ghosting artifacts are visible in the EasyHDR result with objects that are close to the camera (larger disparity). Deghosting in Adobe Lightroom (set at medium) prevents ghosting at the expense of recovering the full dynamic range of the scene in these regions and also suffers from merging artifacts. We also experimented with turning on the "Autotone" feature in Lightroom, but the resulting images were very noisy in the dark regions of the image. Our method produces visually pleasing results that are also faithful in tones and colors to the original input videos since we employ exposure fusion. Further, our method does not require selection or editing of parameters for deghosting, tonemapping etc. Figure 3 shows cropped regions to better visualize the artifacts. The full resolution images, along with results using Lightroom's tonemapping feature, are available as supplementary material.

We have also tested our method and EasyHDR applied frame by frame on several video sequences captured using the camera array and our method performs better in these cases too. In particular, we found that no temporal artifacts were observed despite the application of our algorithm independently on each frame due to our robust alignment and merging approach. EasyHDR produces visual artifacts when objects are close to the camera and per-frame alignment results in stretching artifacts in the corners of the video. Supplementary material (videos and full resolution images) are available for viewing from: `https://drive.google.com/open?id=0B5YI3IcyN8cfMmdPNGFBTGl3eU0`. We were unable to generate videos using Adobe Lightroom due to the lack of a batch processing feature. We were also unable to compare our method to [12] since the authors were unable to share the code for their method when contacted by email. The C++ implementation of our algorithm takes roughly 4 seconds to merge four frames with resolution 1920×1080 on an Intel(R) Core(TM) i7-4770 CPU @ 3.50GHZ with 32GB memory (RAM). Disparity estimation is the main bottleneck and takes 2.5 seconds, followed by alignment and merging which takes 1.2 seconds.

5. CONCLUSIONS AND FUTURE WORK

We presented an algorithm for generating HDR video from short baseline camera arrays in a computationally efficient manner that is suitable for mobile applications. Camera arrays offer advantages over other methods of generating HDR video and our system also generates a disparity map from variably exposed images that is useful for other applications such as measurement, segmentation etc. In the future, we would like to extend our work to larger baseline camera arrays where more sophisticated methods will be necessary to minimize visual artifacts around occlusion boundaries. We would also like to utilize tracking and other methods to avoid computing disparity at every frame and improve the runtime of our algorithm. Finally, we would like to investigate optimal exposure assignments for different cameras in an array to best capture a given scene.

6. REFERENCES

[1] S. Mann and R. W. Picard, "On Being 'undigital' With Digital Cameras: Extending Dynamic Range By Combining Differently Exposed Pictures," in *Proc. IS&T*, 1995.

[2] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *IEEE Computer Vision and Pattern Recognition*, 2000, vol. 1, pp. 472–479 vol.1.

[3] Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, Morgan Kaufmann Publishers Inc., 2005.

[4] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High dynamic range video," in *ACM SIGGRAPH*, San Diego, California, 2003, pp. 319–325.

[5] S. Mangiat and J. Gibson, "Spatially adaptive filtering for registration artifact removal in hdr video," in *IEEE International Conference on Image Processing*, 2011, pp. 1317–1320.

[6] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B. Goldman, and Pradeep Sen, "Patch-based high dynamic range video," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 1–8, 2013.

[7] Yulia Gryaditskaya, Tania Pouli, Erik Reinhard, Karol Myszkowski, and Hans-Peter Seidel, "Motion aware exposure bracketing for hdr video," in *Proceedings of the 26th Eurographics Symposium on Rendering*, 2015, pp. 119–130.

[8] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance imaging using large camera arrays," in *ACM SIGGRAPH*, 2005, pp. 765–776.

[9] Emanuele Trucco and Alessandro Verri, *Introductory Techniques for 3-D Computer Vision*, Prentice Hall PTR, 1998.

[10] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *IEEE Workshop on Stereo and Multi-Baseline Vision*, 2001, pp. 131–140.

[11] N. Sun, H. Mansour, and R. Ward, "HDR image construction from multi-exposed stereo LDR images," in *IEEE International Conference on Image Processing*, 2010, pp. 2973–2976.

[12] A. Troccoli, S. B. Kang, and S. Seitz, "Multi-view multi-exposure stereo," in *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, 2006, pp. 861–868.

[13] Marc Levoy and Pat Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 1996, pp. 31–42, ACM.

[14] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *Pacific Conference on Computer Graphics and Applications*, 2007, pp. 382–390.

[15] M. Okutomi and T. Kanade, "A multiple-baseline stereo," in *IEEE Computer Vision and Pattern Recognition*, 1991, pp. 63–69.

[16] R. T. Collins, "A space-sweep approach to true multi-image matching," in *IEEE Computer Vision and Pattern Recognition*, 1996, pp. 358–363.

[17] Andrea Fusiello, Emanuele Trucco, and Alessandro Verri, "A compact algorithm for rectification of stereo pairs," *Machine Vision and Applications*, vol. 12, no. 1, pp. 16–22, 2000.

[18] Peter J. Burt and Edward H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Transactions on Graphics*, vol. 2, no. 4, pp. 217–236, 1983.

[19] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Transactions on Graphics*, vol. 31, no. 6, pp. 1–11, 2012.

[20] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on Computer Vision*, 1994, pp. 151–158.