HDR IMAGE CONSTRUCTION FROM MULTI-EXPOSED STEREO LDR IMAGES

Ning Sun, Hassan Mansour, Rabab Ward

University of British Columbia
Department of Electrical and Computer Engineering
{nings, hassanm, rababw}@ece.ubc.ca

ABSTRACT

In this paper, we present an algorithm that generates high dynamic range (HDR) images from multi-exposed low dynamic range (LDR) stereo images. The vast majority of cameras in the market only capture a limited dynamic range of a scene. Our algorithm first computes the disparity map between the stereo images. The disparity map is used to compute the camera response function which in turn results in the scene radiance maps. A refinement step for the disparity map is then applied to eliminate edge artifacts in the final HDR image. Existing methods generate HDR images of good quality for still or slow motion scenes, but give defects when the motion is fast. Our algorithm can deal with images taken during fast motion scenes and tolerate saturation and radiometric changes better than other stereo matching algorithms.

Index Terms— High dynamic range imaging, stereo matching.

1. INTRODUCTION

Typical CCD or CMOS sensors can only capture between three and four orders of magnitude of light intensity, whereas human eyes are sensitive to around five orders of magnitude simultaneously, far exceeding the dynamic range that can be instantaneously captured by these sensors. High dynamic range (HDR) imaging provides the capacity to represent the wider dynamic range of natural scenes to which the human visual system (HVS) is sensitive in digital form. However, existing sensor technology has not caught up to the demands of HDR imaging. Few studios have so far managed to develop HDR cameras, however, their solutions are expensive and require a long time to capture the full dynamic range. Therefore, there is a need for low cost solutions that can generate HDR content using only low dynamic range (LDR) cameras.

In recent years, several approaches have been developed to produce scenes with expanded dynamic ranges using LDR images. One approach is to compute the inverse tone mapping curve from a given LDR image and use this curve to stretch the dynamic range of the LDR image [1]. The limitation of such an approach is that it cannot recover information lost in saturated or coarsely quantized regions of the LDR image. Other techniques capture multiple images of a static scene at different exposures from a single camera and combine them to form the HDR image [2–4]. The static scene requirement can be removed by setting up sensors that have spatially varying pixel exposures [5]. However, this setup increases the cost of such cameras and it reduces the effective resolution of the resulting HDR image.

Another approach involves subjecting the frames in a video sequence to different exposures then using motion vectors to match objects in each frame and combining the multi-exposed objects to expand the dynamic range [6]. However, this approach is computationally expensive and can lead to significant artifacts in high motion scenes. Recently, adaptive normalized cross-correlation (ANCC) was proposed to deal with illumination and camera variations [7]. It transforms the R, G, B channels to the log space to eliminate the the effect of lighting difference. However, ANCC fails in those image areas that are saturated [7].

This work has been supported by NSERC.

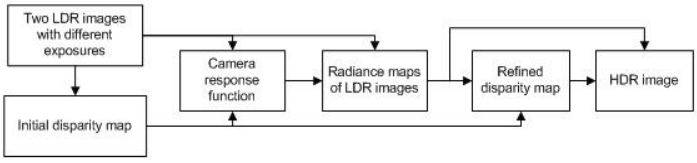

Fig. 1. Our proposed scheme for HDR construction

In this paper, we are interested in a multi-exposure stereo camera setup to HDR image generation. Our approach is inspired by the work in [8] and can be summarized using the following stages:

1. Multi-exposed stereo images are captured using identical cameras placed adjacent to each other on a horizontal line.
2. Stereo matching is then used to find a disparity map that matches each pixel in one image to the corresponding pixel in another image.
3. A subset of the matched pixels is used to generate the camera response function which in turn is used to generate the scene radiance map for each view with an expanded dynamic range.
4. The disparity map is refined by performing a second stereo matching stage using the radiance maps.

Our approach improves on that in [8] through the disparity refinement stage. Consequently, the resulting HDR images exhibit fewer artifacts and encode a wider dynamic range than existing techniques. Fig. 1 illustrates a block diagram of our proposed scheme.

The remainder of this paper is organized as follows. Section 2 presents our proposed stereo matching algorithm that generates an initial disparity map and consequently the camera response function. In section 3 we propose a disparity refinement algorithm which enhances the stereo matching and presents an error HDR composite image. Finally, we present our experiment results in section 4 and draw our conclusions in section 5.

2. STEREO MATCHING

2.1. Overview

The quality of the constructed HDR image depends primarily on the success of the stereo matching scheme used. Stereo matching is a

Fig. 2. Multi-exposed input LDR images: Dolls (top), Arts (below).

field in computer vision that has matured over the last few decades. There are numerous algorithms which perform well on images of the same illumination and exposure. However, most of these algorithms fail on images with large radiometric variations as a result of changes in exposure and lighting [9] such as the images in Fig. 2.

Stereo matching algorithms share a common assumption that the disparity map between two rectified images can be modeled as a Markov random field (MRF). The matching problem is then posed as a Bayesian labeling problem in which the optimal labels f (or pixel disparities in our case) are the values that minimize an energy functional $E(f)$ [10]. The energy functional emerges from the maximum a posteriori (MAP) objective composed of a pixel dissimilarity term $E_d(f)$ and a smoothness term $E_s(f)$ which correspond to the MRF likelihood and the MRF prior, respectively. The best disparity map f^* is therefore obtained by solving the following:

$$f^* = \arg\min_{f \in \mathcal{F}} E_d(f) + E_s(f, N), \tag{1}$$

where $\mathcal{F}$ is the set of feasible disparities from which f is chosen, and N defines a neighborhood window. In what follows, we discuss how the imaging model and the differences in lighting and exposure affects our choice of the energy terms E_d and E_s.

2.2. Imaging model

Imaging models are used to determine the scene radiance from the measured pixel data. Different imaging models have been presented in the literature. In this paper, we introduce two models: the gamma correction model and the polynomial model. We first use the gamma correction model to find an initial disparity estimate and then to find the polynomial camera response function. Next we estimate the scene radiance from the polynomial model to refine our initial stereo matching estimate.

2.2.1. Gamma correction

Every camera measures and quantizes an estimate of the scene radiance R. In a stereo setup, assuming the captured scene is Lambertian, the radiance should be the same in both images and should only be subject to a lateral shift in pixel locations. The image intensities recorded by a camera can be modeled as scaled gamma corrections of the scene radiance R. Therefore, we characterize the imaging model by the following expressions [3]:

$$I_l = R^{\gamma}, \quad I_r = (eR)^{\gamma}. \tag{2}$$

where I_l and I_r are the left and right image intensities, e is the exposure ratio between the left and right images, and γ is the correction factor employed by the camera response curve. However, in really, the radiance received by the two cameras is not exactly the same. We use cost functions described below which are robust to such differences.

2.2.2. Polynomial camera response

In [3], different camera response functions were compared and it was shown that the response curve can be modeled by an n^{th} order polynomial function of the measured pixel values I. The study also showed that it is sufficient to use $n \leq 4$ to build an accurate model. In the stereo matching setup, only left and right image pixels that have the same disparity values (valid pixels) are used to find the camera response function. The polynomial coefficients c_n are then found by minimizing the following cost:

$$J(c_n) = \sum_{p \in \mathcal{P}} \left[\sum_n c_n I_l^n(p) - e \sum_n c_n I_r^n(p) \right]^2 \tag{3}$$

where $\mathcal{P}$ is the set of valid pixels, c_n are the polynomial coefficients, and e is the exposure ratio between the two images.

2.3. Computing the disparity map

The disparity map f characterizes the lateral displacement by an integer number of pixels of an object in the left image compared to its position in the right image. We minimize the energy function $E(f)$ defined in (1) to compute it. However, we must first define the dissimilarity term $E_d(f)$ and the smoothness term $E_s(f, N)$.

2.3.1. Pixel dissimilarity

We choose the normalized cross correlation (NCC) as the pixel similarity measure. In [8,9], it is shown that NCC is the best cost function to cope with exposure variations. For a pixel p and corresponding disparity f_p, NCC is given by the following expression:

$$NCC(p, fp) = \frac{\sum_{q \in W(p)} w_l w_r \tilde{I}_l(q) \tilde{I}_r(q + f_p)}{\sqrt{|w_l \tilde{I}_l(p)|^2} \sqrt{|w_r \tilde{I}_r(p + f_p)|^2}}, \tag{4}$$

where $f_p \in \mathcal{F}$ is the disparity of pixel p, w_l and w_r are bilateral weights defined over a window $W(p)$ centered at p in the left image and $(p + f_p)$ in the right image respectively, and $\tilde{I}_l$ and $\tilde{I}_r$ are functions of the left and right image pixel values which we define below. The bilateral weights for a pixel t in a window $W(p)$ are given by the following expression:

$$w(t) = \exp\left[-\frac{\|p - t\|^2}{2\sigma_d^2} - \frac{\|I^{'}(t) - I^{'}(p)\|^2}{2\sigma_s^2} \right], \tag{5}$$

where σ_s and σ_r are the respective space and range smoothing parameters, and $I^{'} = logI = \gamma loge + \gamma logR$ is the log space pixel intensity. Operating on the log space of the image removes the effect of exposure from the bilateral weights.

The NCC is effective at finding similarities in highly textured surfaces. Therefore, we subtract the low frequency image components before performing the similarity matching. The functions $\tilde{I}$ are then chosen so that:

$$\tilde{I}_l = I_l^{'} - \frac{\sum_{t \in W(p)} w(t) I_l^{'}}{\sum_{t \in W(p)} w(t)} = \gamma \left[\log R - \frac{\sum_{t \in W(p)} w(t) \log R}{\sum_{t \in W(p)} w(t)} \right]. \tag{6}$$

Similarly,

$$\tilde{I}_r = \gamma \left[(\log e + \log R) - \frac{\sum_{t \in W(p)} w(t)(\log e + \log R)}{\sum_{t \in W(p)} w(t)} \right]$$
$$= \gamma \left[\log R - \frac{\sum_{t \in W(p)} w(t) \log R}{\sum_{t \in W(p)} w(t)} \right]. \quad (7)$$

Equations (6) and (7) show that the NCC when applied to $\tilde{I}$ is unaffected by γ and e. The dissimilarity term can then be expressed as follows:

$$E_d(f) = \sum_p D_p(f_p) = \sum_p \left(1 - NCC(p, f_p)\right). \quad (8)$$

2.3.2. Disparity smoothness

The disparity map is assumed to be smooth within solid objects since these objects should have a constant lateral shift between the left and right images. Therefore, we express the smoothness term $E_s(f, N)$ in terms of a total variation function $V_{p,q}$ regularized by weights $\lambda(p, q)$ which are calculated using the perceptually uniform CIELab color space.

Denote by $q \in N(p)$ the pixel indices that fall within a neighborhood window N centered at pixel p. The variation term $V_{p,q}$ is expressed as follows:

$$V_{p,q}(f_p, f_q) = \min(|f_p - f_q|^2, V_{\max}), \quad (9)$$

where $V_{\max}$ is the maximum upper bound, and the regularizing parameter $\lambda(p, q)$ is given by:

$$\lambda(p,q) = \exp\left[- \frac{\|p-q\|^2}{2\sigma_s^2} - \frac{\|I_L(p) - I_L(q)\|^2}{2\sigma_r^2} - \frac{\|I_a(p) - I_a(q)\|^2}{2\sigma_r^2} - \frac{\|I_b(p) - I_b(q)\|^2}{2\sigma_r^2} \right], \quad (10)$$

where I_L, I_a, I_b are the CIELab color space components.

Since the CIELab components are perceptually uniform within an object, $\lambda(p, q)$ ensures that smoothness is imposed within an object and is disregarded at object boundaries. Instead of segmenting images using computationally intensive algorithms, this grouping can be coded using pre-calculated bilateral weights of local support areas [11]. The final smoothness term is expressed as follows:

$$E_s(f, N) = \sum_p \sum_{q \in N(p)} \lambda(p, q) V_{p,q}. \quad (11)$$

2.3.3. Initial disparity and camera response

The energy function given in (1) is then minimized using the graph cut algorithm [12, 13] to produce the initial disparity estimate. This disparity map contains errors mainly in over-exposed and under-exposed regions of the images. Therefore, we calculate two disparity maps for each of the left and right images and cross validate the resulting maps. The pixels that are matched in the two disparity maps are treated as the valid disparities. The remaining pixels are marked as erroneous and represented by black pixels for further correction. Fig. 3 shows the initial disparity map obtained after matching the left and right disparities for two images: Arts and Dolls.

The matched pixels in the two disparity maps are considered as valid disparity values to compute the camera response function using the algorithm in [3]. This camera response function is modeled by a polynomial function. The coefficients c_n are found by minimizing the cost function given by (3).

3. ERROR CORRECTION

After finding the coefficients of the polynomial camera response function, the left and right images are converted to the radiance space $\tilde{R}$ in which the corresponding pixels should have the same value. We use the radiance maps to correct the erroneous pixels identified in the initial disparity map by interpolating between the valid disparities.

We formulate this interpolation problem as another minimization of an energy function (1), but with a different pixel dissimilarity cost $E_d(f)$. Let $\tilde{f}_p$ be the initial disparity estimate of pixel p. For valid pixels in the initial disparity map:

$$D_p(f_p) = \begin{cases} 0, & \text{if } f_p = \tilde{f}_p \\ K, & \text{if } f_p \neq \tilde{f}_p \end{cases}, \quad (12)$$

where K is a large number.

For the erroneous pixels in the initial disparity map:

$$D_p(f_p) = \|\tilde{R}_l(p) - \tilde{R}_r(p + f_p)\| + C_p(f_p, W(p), \tilde{R}_l, \tilde{R}_r), \quad (13)$$

where $C_p(f_p, W(p), \tilde{R}_l, \tilde{R}_r)$ is a cost function that calculates the Hamming distance between pixels p and $p + f_p$ after applying the Census transform [14] over windows $W(p), W(p + f_p)$ in the left and right radiance estimates $\tilde{R}_l, \tilde{R}_r$, respectively.

Notice that the new dissimilarity cost function D_p for erroneous pixels is composed of two norms. The first norm ensures a smooth transition across object boundaries in the radiance map, while the second norm ensures that pixel locations are accurately matched. However, strict disparity matching can cause edge artifacts in the final HDR image which result from occlusions in the stereo setup. Therefore, we enforce the smooth transitioning in the radiance map to remove any possible artifacts that may arise. Finally, in order to speed up the minimization process, we bound the search range of feasible disparity values by the minimum $f_{v,\min}$ and maximum $f_{v,\max}$ valid disparity values found in the initial disparity estimate, such that $f : f_{v,\min} \leq f \leq f_{v,\max}$. Once the depth map and the left and right radiance maps are computed, the value of a pixel in the HDR image is calculated as a weighted average of the corresponding pixels in the two LDR images [3].

4. EXPERIMENTAL RESULTS

We tested our algorithm using stereo images provided by Middlebury College [15]. In this paper, due to space limitation, we only present results for the Arts and Dolls, shown in Fig. 2. In our experiments, the size of the window in the cost function and the neighborhood in the smooth term are (5×5) pixels. The standard deviation σ_s and σ_r in the bilateral weights are 2.6 and 14.0 when computing weighted NCC and 2.4 and 16.0 when calculating $V_{p,q}$. The values of the variables are determined to be optimal for majority of LDR images after several experiments.

Fig. 3 shows the reference disparity maps and the final disparity maps. If the images contain scattered saturated regions of small areas, such as in Dolls, the disparity maps obtained by our algorithm follow closely the reference map. If there are large saturated regions such as in Arts, the final disparity map has discernable difference

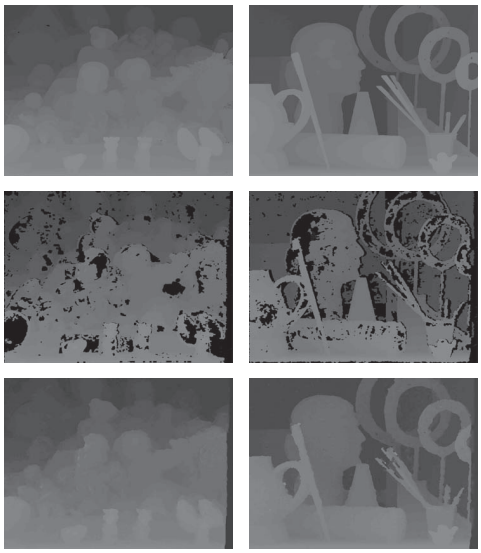

Fig. 3. The first row shows the reference disparity maps. The second and third row shows the initial and final disparity maps.

Fig. 4. Tone-mapped reconstructed HDR images of arts and dolls.

from the reference. However, the errors have little effect on the final HDR images. The tone-mapped HDR images shown in Fig. 4 are obtained by applying tone-mapping operator in [16] Compared to the disparity maps for Arts shown in [8] and [9], the disparity map we computed has less error and better smoothness. The root mean square error (RMSE) and percentage of invalid pixels in our calculated disparity maps are presented in Table 1.

5. CONCLUSION

In this paper, we presented an algorithm that calculates the disparity map of two differently exposed LDR images to generate HDR images. Compared to existing methods, our algorithm can better cope with changes in exposure and can deal with the existence of saturated regions in images. Moreover, our algorithm can be used with fast motion scenes since the proposed setup captures images with different exposures at the same instance, no temporal adjustment is required. Every pair of frames can be treated as images of a static scene and use our algorithm to generate the HDR image.

6. REFERENCES

Table 1. The root mean square error and percentage of invalid pixels in the final disparity maps

	Exposure Ratio	RMSE	Error Percentage
Statue	4	0.9943	8.23
	16	0.976	8.82
Dolls	4	0.8454	4.77
	16	0.8591	5.58
Clothes	4	1.5459	7.43
	16	1.1556	8.15
Baby	4	1.432	9.42
	16	1.4642	10.13

[1] F. Banterle, P. Ledda, K. Debattista, A. Chalmers, and M. Bloj, "A framework for inverse tone mapping," *The visual Computer: International Journal of Computer Graphics*, vol. 23, pp. 467–478, 2007.

[2] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of SIGGRAPH 97*, August 1997, pp. 369–378.

[3] T. Mitsunaga and S. K. Nayar, "Radiometric self calibration," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 1999, pp. 374–380.

[4] R. A. Varkonyi-Koczy, R. Hashimoto, S. Balogh, and S. Y., "Gradient based synthesized multiple exposure time hdr image," in *Instrumentation and Measurement Techonology Conference Proceedings*, May 2007, pp. 1–6.

[5] T. Mitsunaga and S. K. Nayar, "High dynamic range imaging: spatially varying pixel exposures," *ACM Transaction On Graphics*, vol. 3, pp. 267–276, 2002.

[6] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High dynamic range video," in *International Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2003*, 2003, pp. 319–325.

[7] Y. S. Heo, K. M. Lee, and S. U. Lee, "Illumination and camera invariant stereo matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

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

[9] H. Hirschnuller and D. Scharstein, "Evaluation of cost function for stereo matching," in *Proceedings of Computer Vision and Pattern Recognition*, 2007.

[10] S. Z. Li, "Markov random field models in computer vision," in *ECCV '94: Proceedings of the Third European Conference-Volume II on Computer Vision*. London, UK: Springer-Verlag, 1994, pp. 361–370.

[11] R. Brockers, "Cooperative stereo matching with color-based adaptive local support," in *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*, vol. 5702, 2009.

[12] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[13] *http://www.adastral.ucl.ac.uk/ vladkolm/software.html*.

[14] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondance," *Proceedings of ECCV*, pp. 131–158, 1994.

[15] *http://cat.middlebury.edu/stereo/data.html*.

[16] S. P. P. D. E. Reinhard, G. Ward, *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting*. Morgan Kaufmann, 2005.