Signal Processing: *Image Communication* 29 (2014) 191–202

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

High dynamic range video reconstruction from a stereo camera setup

Michel Bätz [a,*], Thomas Richter [a], Jens-Uwe Garbas [b], Anton Papst [b], Jürgen Seiler [a], André Kaup [a]

[a] *Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany*

[b] *Fraunhofer Institute for Integrated Circuits, Electronic Imaging, Am Wolfsmantel 33, 91058 Erlangen, Germany*

ARTICLE INFO

Available online 6 September 2013

Keywords:
High dynamic range video
Stereo camera setup
Stereo matching

ABSTRACT

To overcome the dynamic range limitations in images taken with regular consumer cameras, several methods exist for creating high dynamic range (HDR) content. Current low-budget solutions apply a temporal exposure bracketing which is not applicable for dynamic scenes or HDR video. In this article, a framework is presented that utilizes two cameras to realize a spatial exposure bracketing, for which the different exposures are distributed among the cameras. Such a setup allows for HDR images of dynamic scenes and HDR video due to its frame by frame operating principle, but faces challenges in the stereo matching and HDR generation steps. Therefore, the modules in this framework are selected to alleviate these challenges and to properly handle under- and oversaturated regions. In comparison to existing work, the camera response calculation is shifted to an offline process and a masking with a saturation map before the actual HDR generation is proposed. The first aspect enables the use of more complex camera setups with different sensors and provides robust camera responses. The second one makes sure that only necessary pixel values are used from the additional camera view, and thus, reduces errors in the final HDR image. The resulting HDR images are compared with the quality metric HDR-VDP-2 and numerical results are given for the first time. For the Middlebury test images, an average gain of 52 points on a 0-100 mean opinion score is achieved in comparison to temporal exposure bracketing with camera motion. Finally, HDR video results are provided.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The high dynamic range of a real scene cannot be fully captured by regular CCD or CMOS camera sensors. The resulting images are therefore called low dynamic range (LDR) images. However, since the human visual system (HVS) is sensitive to a far wider dynamic range compared to LDR images, techniques for creating HDR content are useful to capture more information of a real scene, and thus, improve image quality. Several methods exist to generate HDR images which can be categorized into three classes according to [1]: Direct acquisition using particular HDR cameras [2], HDR rendering for synthetically created images, and exposure bracketing. The first class has two major drawbacks. The special hardware is very expensive and it cannot capture a dynamic range as large as when using an exposure bracketing technique. The second class

* Corresponding author. Tel.: +49 9131 85 28904; fax: +49 9131 85 28849.
E-mail addresses: baetz@lnt.de (M. Bätz), richter@lnt.de (T. Richter), jens.garbas@iis.fraunhofer.de (J.-U. Garbas), anton.papst@iis.fraunhofer.de (A. Papst), seiler@lnt.de (J. Seiler), kaup@lnt.de (A. Kaup).

0923-5965/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2013.08.016

is only relevant for computer-generated content, and hence, is not applicable for improving images of a real scene. The third class of HDR acquisition is a low-budget solution. In recent years, many consumer cameras started to offer exposure bracketing as a method to create HDR images of static scenes. For exposure bracketing, a series of LDR pictures is taken with different exposure times one after the other in order to estimate the camera response, convert the pixel values into radiance values and then merge them together to generate the HDR output. Two well-known approaches are the algorithm by Debevec and Malik [3] and the method by Robertson et al. [4]. The different exposures can also be realized by using different apertures or neutral density (ND) filters instead of exposure times.

However, this approach is not applicable for capturing dynamic scenes or even HDR video due to the temporal fashion of the exposure bracketing since the scene is changing over time. An example for a static scene without and with camera motion is illustrated in Fig. 1. It can be seen that temporal exposure bracketing completely fails when there is motion. There are approaches which compensate for the motion [5], but they have the drawback not being able to easily apply ND-filters, and hence, yielding artifacts from different motion blur or depth of field.

The goal of this work is therefore to allow for HDR image acquisition of dynamic scenes and HDR video. The proposed framework utilizes a stereo camera setup to realize a spatial exposure bracketing where the differently exposed images are distributed among the two views. Due to this, all necessary information to create an HDR image is available at the same time instant. This stereo setup leads to two challenges. On the one hand, HDR image reconstruction requires perfect alignment of the views after image warping. Therefore, a robust disparity estimation is needed. On the other hand, typical stereo matching algorithms rely on the brightness constancy assumption which does not hold for differently exposed views. Each stage of the proposed framework is built to properly handle the just mentioned challenges and also under-/oversaturated regions. Instead of applying a stereo matching approach, it would also be possible to utilize time-of-flight cameras in order to obtain the disparity maps. However, they can only generate low resolution disparity maps, increase the complexity of the setup, and are still considerably expensive. For testing purposes, datasets from the Middlebury Stereo Vision Page [6] and two self-recorded datasets are used. Besides visual quality evaluation, a suitable quality metric is applied with respect to a single-view HDR reference image.

The key novel contributions are the new one-stage framework compared to the two-stage frameworks in the literature [7,8] and the introduction of a saturation map to reduce errors in the final HDR image. Beyond this, an objective quality assessment, which is directly computed on the HDR images, is done for the first time.

The remainder of this article is structured as follows. The upcoming section presents previous work, whereas Section 3 covers the layout and the details of the framework and its modules. Experimental results are then given in Section 4. Finally, Section 5 concludes this contribution.

2. Previous work

This contribution is mainly inspired by the work in [7,8] where stereo matching is done in two stages and the camera responses are computed from the matches found in the first stage. These camera responses are then used for converting the input images to radiance space and a second stereo matching for disparity refinement is performed. In this work, however, unlike the existing algorithms, the camera responses are not estimated from the stereo images. This step is done only once in an offline step using temporal exposure bracketing on a single-view setup. For that reason, the disparity maps can be computed directly from the radiance space images (RSI) and the need for a second stereo matching stage is removed. In addition, the proposed offline step allows for camera setups with different sensors since each camera can be calibrated separately. Besides the novel one-stage framework, this work also presents the concept of applying a so-called saturation map before the actual HDR generation. In addition to that, objective quality evaluation results are provided for the first time. Furthermore, many ideas for the disparity estimation are taken from the currently

Fig. 1. Temporal exposure bracketing without (top) and with (bottom) camera motion.

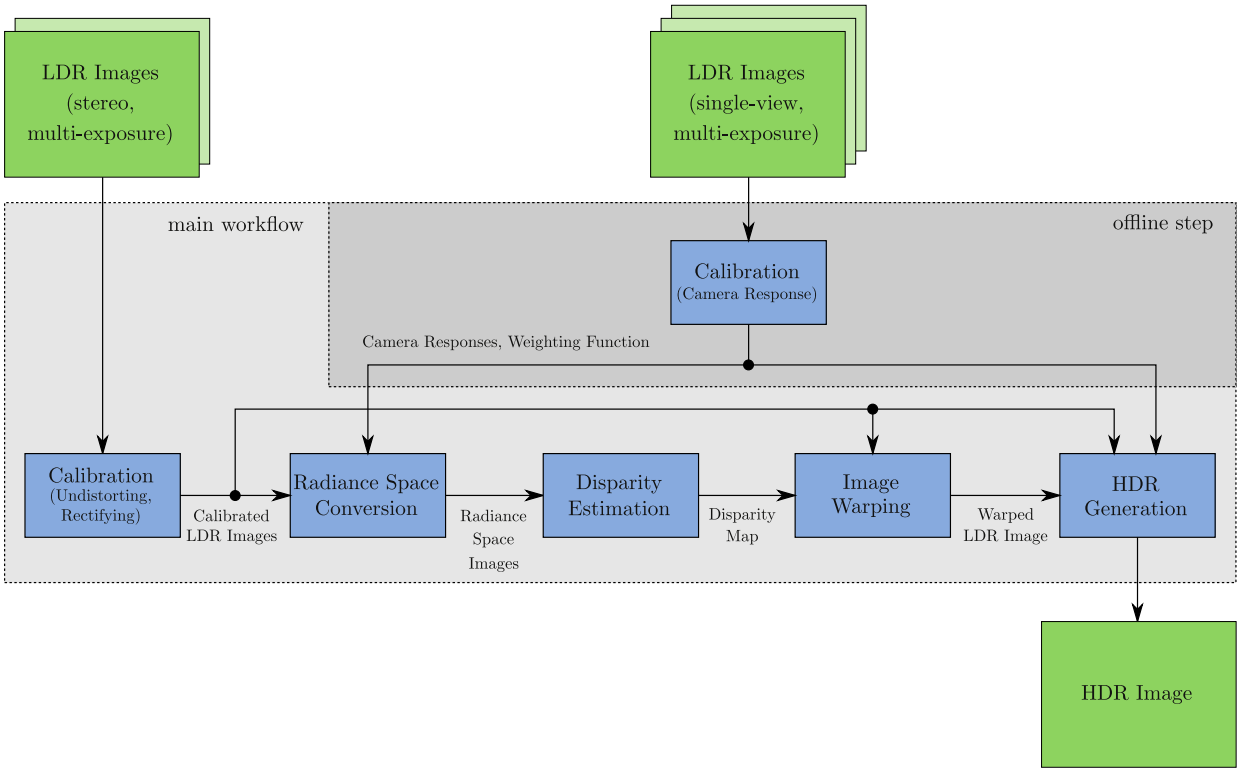

Fig. 2. Proposed stereo HDR imaging framework layout.

second-placed stereo matching algorithm [9] on the Middlebury Stereo Vision Page [6].

3. Proposed stereo high dynamic range imaging framework

Before the particular stages of the proposed stereo high dynamic range (Stereo-HDR) imaging framework are discussed in detail, an overview of the framework is given. A schematic layout in the form of a block diagram is provided in Fig. 2. Basically, the framework can be split into the main workflow and an offline step.

The new offline step is responsible for calculating the camera response curves and weighting functions (see Section 3.4), which are both necessary for the conversion of the LDR images to radiance space. This calibration step has to be done only once. In general, the camera response curves for both cameras the setup have to be calculated. However, when using the same camera hardware for both views it is sufficient to conduct the calibration only for one view. Therefore, multi-exposed LDR images of a single view serve as input for this step. Due to temporal exposure bracketing applied in this offline step, a static scene is required for the camera calibration. This approach avoids errors introduced by disparity estimation and image warping and it also allows for a camera setup with different hardware.

The main workflow comprises six stages and in the case of a video input operates frame by frame. The input for this part consists of the LDR images where each image represents a view taken from a different position with a different exposure. From these images the output HDR image is created.

The first stage is responsible for the geometric calibration of the input images and consists of undistorting and rectifying the input (see Section 3.1). In the next stage, the LDR images are converted to radiance space images. Here, the pixel values of the differently exposed images are brought into a common value space where properly exposed pixels in all images yield approximately the same radiance value. This is done to ease the finding of correct correspondences which is performed by the stereo matching algorithm. For proper conversion, all the side information like camera response curves, exposure times, apertures, ISO speeds and weighting functions is necessary. The third stage deals with the disparity estimation. Since it contains a lot of modules it is described in detail in Section 3.2. The output of this stage is a set of disparity maps. The following stage handles the image warping, for which the calibrated LDR image to be warped and its corresponding disparity map are required. This is discussed more precisely in Section 3.3. In the last stage, the HDR generation is conducted, which requires three different inputs. The first one is the calibrated LDR image of the view that is selected to be enhanced to an HDR image. For best results, this view should contain the least under- or overexposed regions. The other two inputs are the associated warped LDR image and the necessary side information which was already needed for the radiance space conversion. More details on the HDR generation are

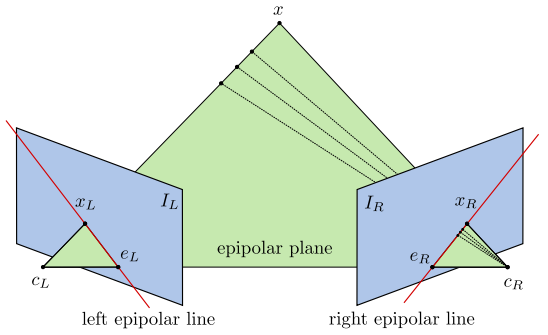

Fig. 3. Epipolar geometry in general.

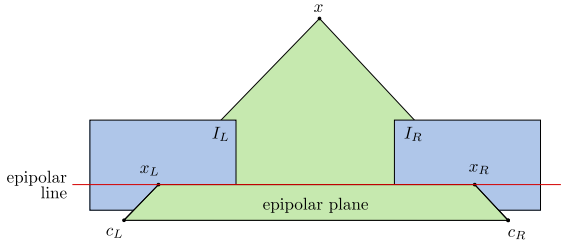

Fig. 4. Epipolar geometry rectified.

given in Section 3.4. Finally, the HDR image is stored in the common file format OpenEXR [10].

In the following section, the geometric camera calibration and its advantages for this setup are described in detail.

3.1. Camera calibration

The reason for applying a rectification step is to make the stereo matching process easier. With the help of epipolar geometry [11] the relations between different camera views of the same scene can be described. Fig. 3 shows an epipolar geometry in a general case, where x is a point of the real scene in 3-D world coordinates and I_L and I_R are the left and right images, respectively. c_L and c_R are the centers of projection of the left and the right camera, respectively. Together with x, they span the epipolar plane. x_L and x_R are the projections of x in the corresponding image. As long as the camera setup is fixed, all epipolar lines go through their corresponding epipole e_L or e_R, respectively.

Given an image point x_L, the distance to the actual point x is ambiguous, and thus, according to the epipolar geometry the corresponding point x_R can be located anywhere on the right epipolar line. In order to find the corresponding pixel in case of unknown epipolar geometry, the whole image I_R has to be searched, whereas only the epipolar line has to be searched for known geometry. Due to the different rotations of this line for each matching point, the search is rather complex. Because of that rectification can be done to simplify this search. For this purpose, the image planes are aligned and their associated optical axes are made parallel. A rectified epipolar geometry is illustrated in Fig. 4. This process causes the left and right epipolar lines to merge, and thus, the search for the proper matching point only has to be done in the same image row, easing the stereo matching. To realize this calibration step, the intrinsic and extrinsic camera parameters have to be known or estimated.

After the rectification of both views, the disparity maps can be computed.

3.2. Disparity estimation

Disparity estimation or stereo matching is the process of finding correspondences between two camera views. The result is a disparity value d which denotes the spatial difference between the matched points. Due to rectification, this is only a scalar value describing a shift in horizontal direction. Disparity estimation can be done in a blockwise or pixelwise fashion. The latter yields dense disparity maps and is therefore applied in this work. In general, disparity estimation can be split into four particular steps. These are cost initialization, cost aggregation, disparity optimization, and disparity refinement. This chain is shown in Fig. 5. Some of these steps can be merged or skipped depending on the algorithm at hand. In the following, the steps are explained in detail.

The first step is the cost initialization. The majority of matching costs rely on the brightness constancy assumption which leads to one of the two major challenges of this work. For that reason, most matching costs are inappropriate for handling images with different exposure. Although

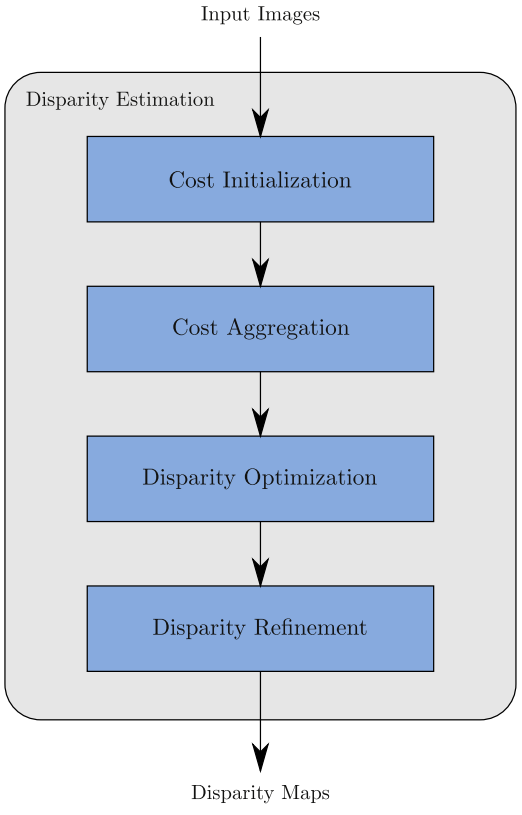

Fig. 5. Disparity estimation chain.

the radiance space images are used for the matching there can be still variations in corresponding values, and thus, a robust matching cost is preferable.

According to [12], matching costs can be grouped into parametric and nonparametric costs and mutual information. Common parametric costs are the sum of absolute differences (SAD), the sampling-insensitive absolute difference of Birchfield and Tomasi (BT) or the normalized cross-correlation (NCC). A variant of the NCC is the zero-mean normalized cross-correlation (ZNCC) which can compensate for both changes in gain and offset inside the correlation window. For that reason, it is a good choice for matching images of different exposure. However, because of the fixed support window size it produces a fattening effect near object boundaries [13]. There is also an adaptive normalized cross-correlation (ANCC) [13] which tries to decrease the fattening effect. In comparison to parametric costs, nonparametric costs are not based on intensity values, but on the local ordering. That is why they can make up for radiometric changes well as long as the ordering is not affected. Common matching costs of this class are the rank filter and the census transform [14]. However, they are only manipulating the input, and hence, are not directly matching costs. For the census transform, the actual matching cost is computed with the aid of the Hamming distance.

According to the extensive results in [12], the census transform and the ZNCC are among the best matching costs when it comes to exposure variations. In their experiments, the census transform outperforms the ZNCC. For our case, though, the ZNCC turned out to deliver slightly better results, and hence, it was chosen for the final setup of this work. Its equation is as follows:

$$C(\mathbf{p},d)=\frac{\sum_{\mathbf{q}\in N_p}(I_L(\mathbf{q})-\bar{I}_L(\mathbf{p}))(I_R(\mathbf{q}-(d,0))-\bar{I}_R(\mathbf{p}-(d,0)))}{\sqrt{\sum_{\mathbf{q}\in N_p}(I_L(\mathbf{q})-\bar{I}_L(\mathbf{p}))^2\sum_{\mathbf{q}\in N_p}(I_R(\mathbf{q}-(d,0))-\bar{I}_R(\mathbf{p}-(d,0)))^2}} \tag{1}$$

where $C(\mathbf{p},d)$ is the matching cost for pixel position $\mathbf{p}$ for disparity level d. I_L and I_R are the left and right images, respectively, while $\bar{I}_L$ and $\bar{I}_R$ are the corresponding mean values for the current neighborhood N_p. $\mathbf{q}$ represents a pixel position inside the local neighborhood. A support window size of 9×9 is utilized in this framework. The result of the ZNCC ranges from -1 to 1 where -1 is the worst and 1 the best matching. In order to make it truly a cost, the signs are swapped afterwards. All types of cost initializations can be done on grayscale or color images. But based on the results in [12,15], color information does not improve the cost initialization step for the ZNCC. For that reason, the matching is done on the luminance images.

No matter which matching cost is chosen, the result is a cost matrix containing the costs for all possible disparities for each pixel. This matrix is also called disparity space image (DSI) [16] and has the form $m\times n\times d_{sr}+1$, where m and n are the image dimensions and d_{sr} is the disparity search range. This DSI is the basis for the cost aggregation step.

This step is primarily done to smooth the DSI, and thus, to get more reliable and robust costs. However, the aggregation should not be done across object boundaries. The cost aggregation can be realized as an averaging over a 2-D support window in each slice separately or as an averaging over a 3-D window. For this work, only cost aggregations with a 2-D support window were taken further into account. There are several approaches like square window, shiftable window, adaptive weight or adaptive window [17]. Since adaptive window approaches can fit the input data very well, a particular adaptive window method, called cross-based aggregation [18,9] is selected in this contribution for cost aggregation purposes. The work in [9] also shows that the improved version of the cross-based aggregation can achieve similar or slightly better results compared to the adaptive weight aggregation. Based on a comparison conducted in [17], the adaptive weight aggregation is said to be one of the best approaches. Furthermore, the cross-based aggregation needs less memory and computation time and creates a support window for each pixel that can be used for later steps as well.

In principle, the cross-based aggregation of [9] is divided into the cross construction and the actual cost aggregation. These two steps are illustrated in Figs. 6 and 7.

In the first step, for each pixel $\mathbf{p}$ a cross is set up, which is highlighted in blue and its four arms are labelled left, right, upper, and bottom arm. In contrast to the actual matching, this cross creation is performed on the RGB input images. Each arm length is calculated with the help of three rules. Given $\mathbf{p}$, the left arm extends up to a pixel $\mathbf{p}_l$ which is the last point satisfying the following three rules:

$$D_c(\mathbf{p}_l,\mathbf{p})<\tau_1 \quad \text{and} \quad D_c(\mathbf{p}_l,\mathbf{p}_l+(1,0))<\tau_1 \tag{2}$$

with

$$D_c(\mathbf{p}_l,\mathbf{p})=\max_{i=R,G,B}|I_i(\mathbf{p}_l)-I_i(\mathbf{p})|$$

where D_c is the maximum color difference of all RGB channels between two pixels and τ_1 is a preset threshold. In the first term of (2), a color comparison with the center pixel $\mathbf{p}$ and in the second term a color comparison with the previous pixel of $\mathbf{p}_l$ is done.

$$D_s(\mathbf{p}_l,\mathbf{p})<l_1 \tag{3}$$

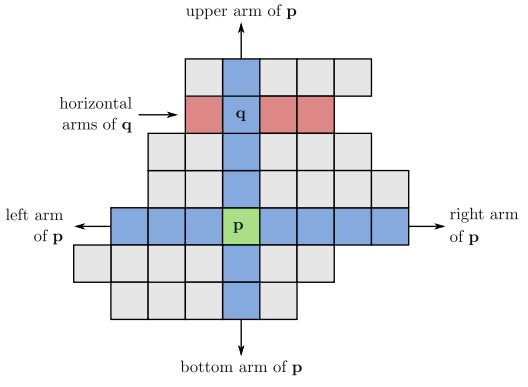

Fig. 6. Cross construction.

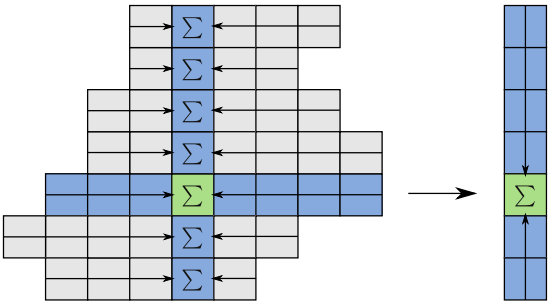

Fig. 7. Cross aggregation.

with

$$D_s(\mathbf{p}_l, \mathbf{p}) = |\mathbf{p}_l - \mathbf{p}|$$

is the second rule where D_s denotes the spatial distance between two pixels and l_1 is a preset threshold. The third rule is given by

$$D_c(\mathbf{p}_l, \mathbf{p}) < \tau_2 \quad \text{if } l_2 < D_s(\mathbf{p}_l, \mathbf{p}) < l_1 \tag{4}$$

with τ_2 and l_2 being stricter threshold values. The four threshold values are set to $l_1 = 34$, $l_2 = 17$, $\tau_1 = 20$, and $\tau_2 = 6$ as proposed in the related paper [9]. The other arms are built analogously. These three rules yield a flexible cross construction which can deal with large textureless regions as well as dark areas or depth discontinuities. Eventually, only four values have to be stored per pixel. In order to create the actual adaptive support window for **p**, all the horizontal arms of the pixels on the vertical arms of **p** are taken. The horizontal arms of **q** are marked in Fig. 6. After having obtained the support region, the cost aggregation can be performed by averaging over all entries. To increase the robustness of this aggregation step, it is done in four passes where the first and the third pass work as explained. For the second and the fourth pass, however, the support region is created by taking all the vertical arms of all pixels on the horizontal arms of **p** into account. Furthermore, to avoid outliers from the other image, a union of the support windows for the current image and the support windows for the other image is used. Afterwards the aggregated DSI can be used for the optimization.

For the disparity optimization, there are local, semi-global or global methods available. On the one hand, local approaches rely only on a local neighborhood for computing the disparities and make implicit smoothness assumptions by cost aggregation. On the other hand, global methods solve an optimization problem where they set up and minimize a global energy function with a data term and an explicit smoothness term. Examples for global optimization approaches are graph-cut, belief-propagation, dynamic programming or scanline optimization [16]. The semiglobal algorithm [19] presents a combined approach. In this contribution a local "winner-take-all" (WTA) approach [16] is selected where the disparity value with the least cost is chosen for each pixel.

The last stage in the disparity estimation chain is the refinement. Here, the initially obtained disparity maps can be improved further by means of various algorithms. In the proposed framework, a cross-checking between the disparity maps, a subpixel enhancement, a median filtering, and a minimum interpolation are applied. In the following, the four applied algorithms are described in detail. To begin with, a cross-checking, also called left-right consistency check [9], is conducted. For that reason, the values in the left-to-right disparity map are compared with their corresponding right-to-left disparity map entries. When the absolute difference of these value pairs exceeds a certain threshold the disparity value is invalidated. This can be expressed with the following equation:

$$|D_L(\mathbf{p}) - D_R(\mathbf{p} - (D_L(\mathbf{p}), 0))| < \tau, \quad \forall \mathbf{p} \tag{5}$$

where D_L and D_R are the left-to-right and the right-to-left disparity maps, respectively. **p** is an arbitrary pixel position. This method ensures that mismatches and occlusions are invalidated. In this contribution invalidated pixels are set to 0 and the tolerance threshold is chosen to be 1 by default.

As a second module, an enhancement towards subpixel accuracy is done. The applied method reduces errors introduced by integer disparity levels and is based on a quadratic polynomial interpolation [9]. It can be formulated as follows:

$$d^* = d - \frac{C(\mathbf{p}, d_+) - C(\mathbf{p}, d_-)}{2(C(\mathbf{p}, d_+) + C(\mathbf{p}, d_-) - 2C(\mathbf{p}, d))} \tag{6}$$

where d^* is the interpolated subpixel-accurate disparity value and d is the current disparity level. $d_+ = d+1$ and $d_- = d-1$, respectively. Finally, the matching cost at pixel **p** and disparity level d is expressed by $C(\mathbf{p}, d)$.

As proposed in the previously mentioned work [9], a 3×3 median filtering is performed in order to smooth the disparity maps. The last refinement step proposed in this contribution is an iterative minimum interpolation with a square window kernel. The decision for such an interpolation step is based on the assumption that occluded pixels most likely come from the background whereas mismatches tend to be located within objects. In each iteration the whole image is processed. Upon finding an invalidated pixel (pixel value 0) a 3×3 neighborhood around it is searched for its minimum entry ignoring all entries being 0. This value is then used to interpolate the current invalidated pixel position. However, the newly interpolated pixels are only employed as a basis for further interpolations starting with the subsequent iteration. With this approach, valid values equally propagate into the holes of the disparity map.

3.3. Image warping

After having obtained the final disparity maps, the auxiliary view can be shifted to the desired main view. This process is called image warping and utilizes the disparity values for each pixel to create a warped image. Basically, there are two possibilities which namely are forward and backward image warping [20]. Occluded areas and newly exposed regions are the two main reasons for errors introduced during the warping process. Both have to be treated in order to get a good result. Unlike disparity estimation, the LDR image is used for warping in this framework instead of the RSI. Forward image warping has

the advantage of implicitly ignoring newly exposed areas but occlusions have to be explicitly dealt with a depth check. Moreover, subpixel-accurate disparity maps can only be used as input when employing an interpolation strategy. On the contrary, backward image warping can directly handle subpixel-accurate disparity maps and implicitly ignores occluded areas. In return newly exposed regions have to be explicitly treated. For this contribution, backward image warping is selected for the final setup. Backward image warping makes use of the target view along with its relating disparity map, passes through them in a pixelwise fashion and searches for the correct pixel value in the image to be warped. Assuming the main view to be the left one and the auxiliary view the right one, a right-to-left warping can be written as

$$I_{\mathrm{wrp}_R}(\mathbf{p}) = I_R(\mathbf{p} + (D_L(\mathbf{p}), 0)) \tag{7}$$

where I_R is the right image, D_L is the left-to-right disparity map and I_{wrp_R} is the warped image from right-to-left. The proposed algorithm applies a bilinear interpolation in order to retrieve values at subpixel positions. Furthermore, warped image positions that correspond to invalidated disparities are set to 0. After the image warping, the HDR generation can be conducted.

3.4. High dynamic range image reconstruction

The HDR reconstruction stage, employed in this framework, consists of two parts. The first part is a novel masking conducted by means of a saturation map. The other part, which is the actual HDR reconstruction step, is based on the algorithm by Robertson et al. [4]. In the following, both parts are explained in detail.

First, the principle of the masking using a saturation map is outlined. Basically, the saturation map is a bit mask representing under- or oversaturated pixels. It is created for the view that is to be enhanced to an HDR image. The exposure for this view should be chosen such that it has the least under- or oversaturated regions. Whenever a single color channel is below or above certain thresholds, e. g., $\theta_u = 5$ and $\theta_o = 250$, a 1 is stored in the saturation map. This behavior can be expressed by the following equation:

$$S(\mathbf{p}) = \begin{cases} 0 & \text{if } \theta_u < I_{CC=\nu}(\mathbf{p}) < \theta_o, \forall \nu \in \{R, G, B\} \\ 1 & \text{else} \end{cases} \tag{8}$$

where $S(\mathbf{p})$ is the newly proposed saturation map, θ_u and θ_o are the thresholds for under- and oversaturation, respectively, and I_{CC} is one color channel of the RGB input image.

The warped image is then multiplied with this mask, and hence, only regions that need a dynamic range extension are taken into account from the warped view for the HDR generation. Here, the idea is that properly exposed regions map correctly to radiance space and therefore, no additional information is required from a differently exposed image. An example for a saturation map can be seen in Fig. 8.

After the masking step, unnecessary pixels in the warped images contain zero entries in all channels which are ignored for the HDR merging due to the weighting factor of 0 for pixel values of 0.

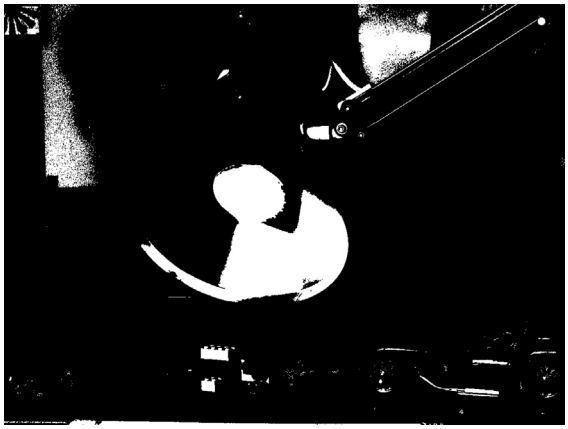

Fig. 8. Saturation map for 'IIS Jumble' where regions with at least one color channel below 5 or above 250 are marked in white.

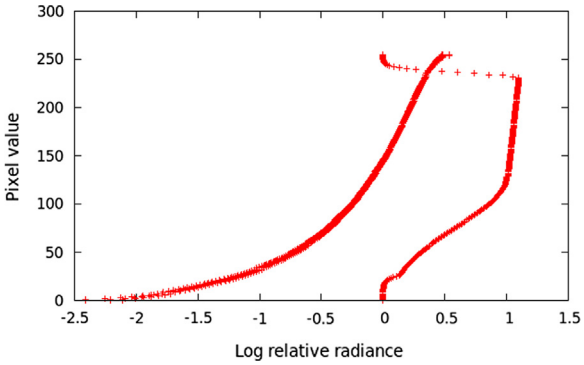

Fig. 9. Mapping of radiance to pixel values created with the help of pfstools [21]. The left curve is actually three functions – one camera response function for each RGB channel. The right curve depicts the composite weighting function with weights between 0 and slightly above 1 for all pixel values.

The second part of the HDR reconstruction is the actual merging step. The formula for combining the LDR images according to [4] is

$$I_{\mathrm{HDR}}(\mathbf{p}) = \frac{\sum_i w(I_i(\mathbf{p}))t_i r(I_i(\mathbf{p}))}{\sum_i w(I_i(\mathbf{p}))t_i^2} \tag{9}$$

where I_{HDR} represents the final HDR image, w is a Gaussian-like weighting function, t_i is the exposure time for image I_i and r denotes the mapping from LDR values to radiance values using the camera responses. The choice for this approach was made because of the good results it achieves and to assure comparability to [21] as it is utilized in this work to compute the camera responses and the single-view HDR reference images. However, it turned out that the software pfstools [21] applies a slightly modified version of the original approach. In addition to the modified equation, the weighting function is also replaced for a composite weighting. This weighting function emphasizes brighter pixel values while still suppressing values at the boundaries of the pixel value range.

The reason for giving brighter values a higher weighting factor compared to darker values is that their SNR is higher. As a consequence, noise in dark regions is suppressed whereas bright intensity values gain a bigger impact. The reason for suppressing values at the boundaries, however, comes from the fact that camera responses are typically most sensitive in the middle of the value range. In other words, the mapping is most accurate near the center of the pixel value range. For better illustration, the camera response curves for the three RGB channels and the composite weighting function of the test dataset 'IIS Jumble' are shown in Fig. 9. The value range of the weighting function is between 0 and slightly above 1.

4. Experimental results

This section is divided into three parts. The first part gives an overview of the simulation setup. The second part is about visual quality evaluation and the last one states some objective quality results.

4.1. Simulation setup

Four different static test datasets and a dynamic test dataset were selected in this contribution. The first three are called 'Aloe', 'Moebius', and 'Art' and are provided by the Middlebury Stereo Vision Page [6,22]. They consist of three different exposures for three different illuminations and a total of seven views in three possible resolutions. Furthermore, ground truth disparity maps, created by a structured light approach [23], are available for the second most left and right views (view1 and view5). All tests in this work were done on those views having a ground truth disparity map. For each set, a fixed illumination was selected and the medium resolutions of the images were taken. The illuminations Illum3, Illum2, and Illum1 were chosen for the datasets 'Aloe', 'Moebius', and 'Art', respectively. The first dataset has a resolution of 641×555 while the other two have a resolution of 695×555. The disparity search ranges were set to 100, 120, and 120 pixels, respectively. Moreover, the different exposure times for 'Aloe' are 125 ms, 500 ms, and 2000 ms, whereas for 'Moebius' and 'Art' they are 250 ms, 1000 ms, and 4000 ms. The fourth dataset, named 'IIS Jumble', was created in the course of this work and is composed of 15 different views arranged in a 5×3 array. Each view provides five different exposures of the scene with a resolution of 2560×1920. The exposure times are 5 ms, 30 ms, 61 ms, 122 ms, and 280 ms. The testing on the fourth dataset was conducted only on view12 and view13 that were downsampled by a factor of 2 in each dimension. After rectification, they were cropped to 1200×900. Finally, the disparity search range was selected to be 100 pixels.

In general, to achieve different exposures for each view the usage of different ND-filters is required in the case of a dynamic scene or video. However, since the first four datasets in this work show static scenes, the utilization of different exposure times was possible in order to simplify the test setup. It should also be mentioned that all HDR images in this contribution are tone-mapped and gamma corrected for displaying with the help of pfstools. The chosen tone mapping operator is the contrast mapping by Mantiuk [24] with 0.5 as contrast scaling factor. As example, the applied left and right views of the 'Art' and 'IIS Jumble' dataset are depicted in Fig. 10.

Aside from these static scene setups, a dynamic scene setup was simulated and an HDR video was created. The dynamic scene 'LMS Pillars' was artificially generated using the open-source 3-D computer graphics software Blender [25]. It contains moving objects with different textures, dark areas, and overexposed spots. The scene is lighted by two distinct light sources and image-based lighting, which employs an HDR environment map from [26]. In order to simulate a stereo camera setup, the scene was rendered with different exposures at two camera positions. The first frame of both the left and right camera together with the tone-mapped HDR sequence is provided in Fig. 11. Unlike the static scenes, the applied tone-mapping operator is the photographic tone reproduction by Reinhard et al. [27]. The complete videos can be found at [28].

For all the tests, the left view is considered to be the target view for HDR reconstruction. Therefore, it is assumed to have the least under- or oversaturated regions. For the Middlebury datasets, the right view was selected to be a brighter version, whereas for the last two datasets a darker version was chosen.

4.2. Visual quality evaluation

In order to evaluate the test results, reference images are required which show the best case. For this setup, it is possible to create two different references: a single-view, and thus, overall best case and a stereo best case which employs cross-checked ground truth disparity maps.

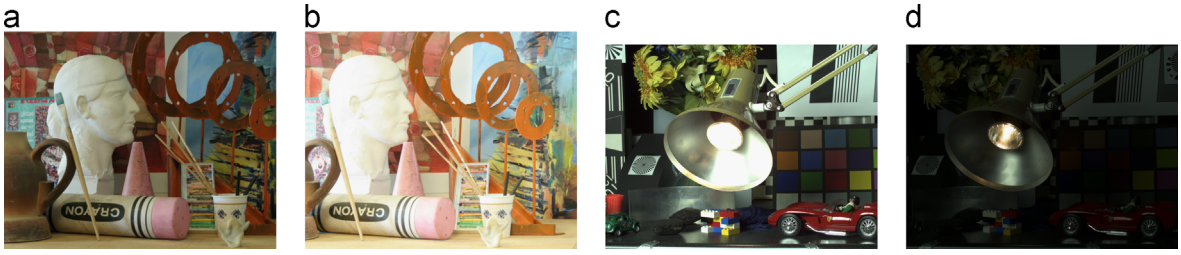

Fig. 10. Stereo, multi-exposure input images depicting 'Art' (a,b) and 'IIS Jumble' (c,d): (a) Left view with normal exposure, (b) right view with long exposure, (c) left view with normal exposure, and (d) right view with short exposure.

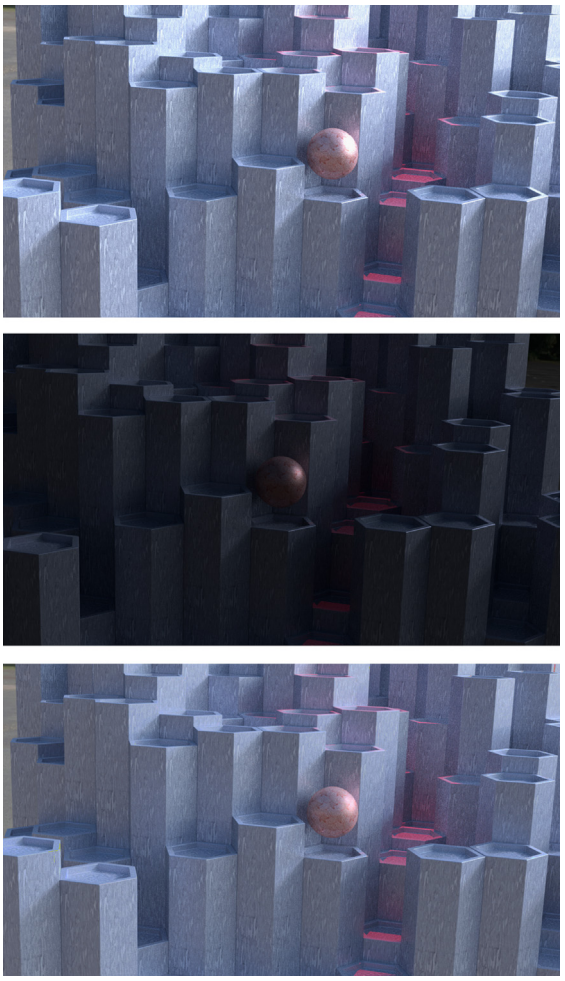

Fig. 11. Dynamic scene 'LMS Pillars' rendered using Blender [25]. From top to bottom: First frame of left camera, right camera, and tone-mapped HDR view. The complete videos can be found at [28].

The single-view reference, however, only works for static scenes as a best case.

Fig. 12(a)–(e) shows the tone-mapped single-view HDR reference images for all four datasets together with a zoomed version of 'Art'. The defective region inside the light bulb in 'IIS Jumble', which is present even in the single-view reference case, is based on an oversaturated area within the shortest exposed LDR input image. Furthermore, Fig. 12(f)–(i) depicts the tone-mapped single-view case acquired during camera movement for the Middlebury datasets. In order to simulate a simple camera motion, a different view is taken for each exposure. For this test setup, the views 1, 3, and 5 are utilized to realize a movement of the camera to the right. It can be seen that regular temporal exposure bracketing is not an option for creating HDR content when there is motion. However, it would be possible to apply a motion estimation approach to the temporal exposure bracketing in order to compensate for the motion [5]. Nonetheless, such an approach cannot easily benefit from ND-filters, and hence, different exposure times or apertures are required, leading to artifacts due to different motion blur or depth of field, respectively.

In addition, Fig. 12(j)–(m) illustrates the tone-mapped stereo HDR reference images. Since 'IIS Jumble' offers no ground truth disparity maps, no stereo reference can be computed. Using the ground truth disparity maps yields much better results, but still suffers from problems at object boundaries. This can be seen in the zoomed version of 'Art'. The results using the proposed framework are shown in Fig. 12(n)–(r) and can even outperform the stereo references.

The masking done with the saturation map has a big impact on the improved results since properly exposed regions do not or only slightly benefit from additional information. However, wrong matches that are taken into account for the HDR creation yield visible errors. As a consequence, only the disparity values at under- or over-saturated positions are of interest. The errors that appear inside the lamp in 'IIS Jumble' come from small erroneous regions within the disparity map which were spread by the minimum interpolation stage.

For all cases, the same camera responses were assumed as calculated in the offline step of the proposed framework. In the next section the visual results from this part shall be confirmed with the help of an objective quality metric.

4.3. Objective quality evaluation

Previous work in [7,8] has only provided results in the form of visual assessment of the tone-mapped HDR images and numerical evaluation of the quality of the created disparity maps. This section therefore serves to build a foundation for comparing the actual HDR images in an objective manner.

For objective quality evaluation of imaging algorithms, quality assessment metrics are very important. Unfortunately, the majority of current metrics like the well-known PSNR cannot handle HDR images properly as they work well only for limited intensity ranges [29]. Therefore, a suitable quality metric that allows for HDR image evaluation had to be chosen for this contribution. For that reason, the visual difference predictor HDR-VDP-2 by Mantiuk et al. [29] was selected. It can predict the quality of the image together with the visibility of errors. While the latter is denoted as a percentage, the image quality is given in the form of a mean opinion score (MOS) ranging from 0 (worst) to 100 (best). The HDR-VDP-2 is a major revision of the HDR-VDP and the VDP and according to [29] it leads to better predictions which are comparable to or better than those of the multi-scale structural similarity (MS-SSIM) [30]. As reference input, the single-view HDR reference images were used and the default settings for the resolution of the displaying device and the distance to the viewer were picked.

Another way to assess the quality of the HDR output would be to first apply a tone mapping operator on them and then employ the commonly used PSNR or more specific metrics based on MS-SSIM as mentioned in [31] or like the TMQI (Tone-Mapped Image Quality Index) [32]. However, the tone mapping step might introduce,

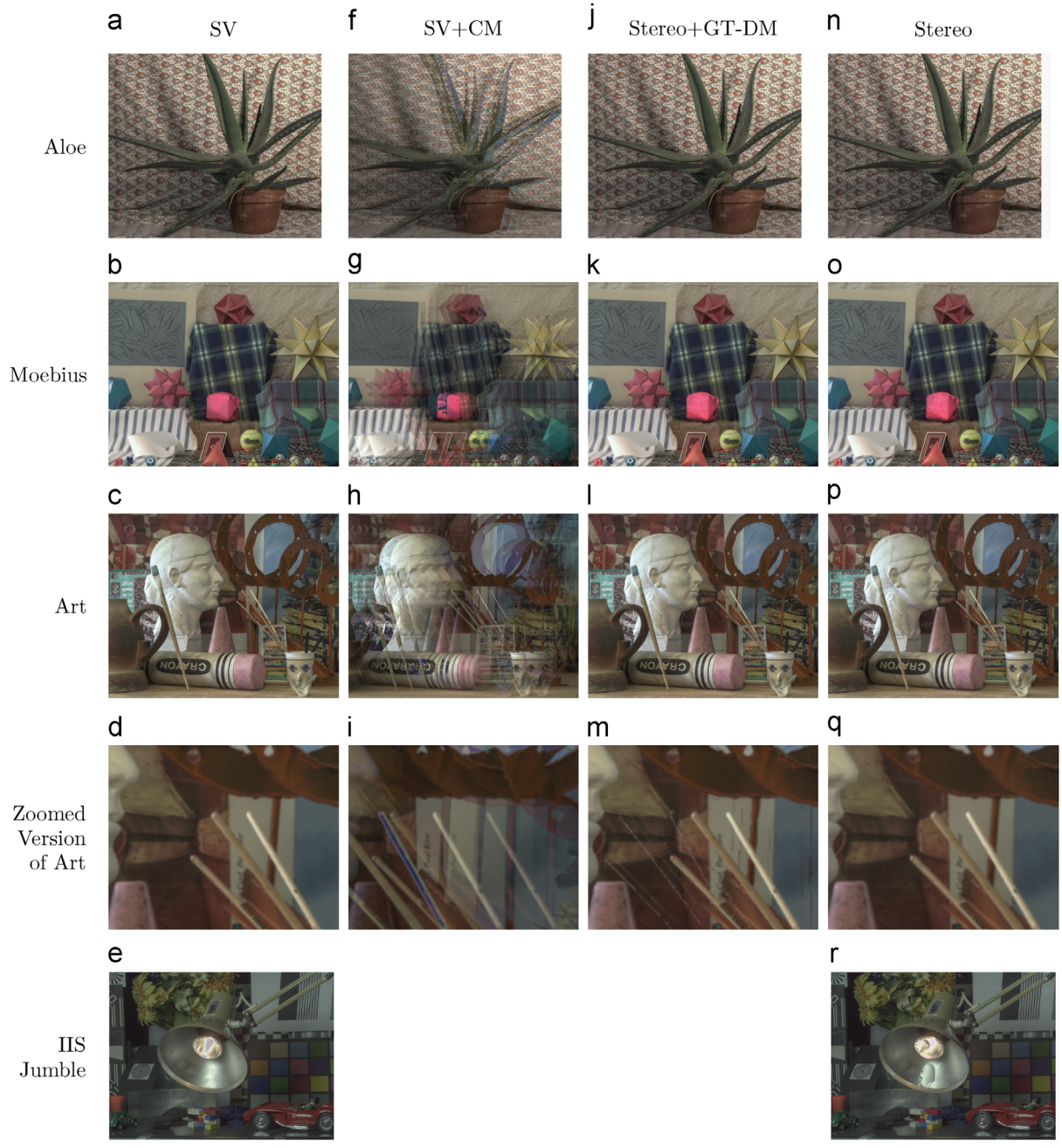

Fig. 12. Tone-mapped HDR images for different test cases and datasets. Datasets from top to bottom: 'Aloe', 'Moebius', 'Art', zoomed version of 'Art', and 'IIS Jumble'. Test cases from left to right: single-view reference (SV), single-view with camera movement (SV+CM), stereo employing cross-checked ground truth disparity maps (Stereo+GT-DM), and proposed stereo setup. No images are given for the camera movement and the ground truth disparity map case for 'IIS Jumble'.

enhance, or attenuate artifacts, and thus, distort the results.

To substantiate the previously obtained visual results, mean opinion scores are predicted with the HDR-VDP-2 and shown in Table 1. It can be seen that the single-view case completely fails when there is a camera movement. For this case, the average MOS drops to 41 points. Furthermore, the proposed refinement including cross-checking, subpixel enhancement, median filtering, minimum interpolation, and saturation masking can even achieve slightly better results compared to simply employing a cross-checked ground truth disparity map. That means that even with ground truth data, errors are introduced during image warping, especially at object boundaries. In addition, the ground truth disparity maps are normally not available anyhow. Since the 'IIS Jumble' dataset provides no ground truth disparities no comparison could be made.

Table 1
Mean opinion scores, where 0 is the worst case and 100 the best, for all four datasets compared to their single-view HDR references. The last column gives average values for the 3 Middlebury datasets. 'SV' denotes single-view and 'GT-DM' means ground truth disparity map. The best values are highlighted.

Predicted MOS		Aloe	Moebius	Art	IIS Jumble	$Avg_{Middlebury}$
	SV with camera motion	38.15	47.44	37.03	–	40.87
	Stereo with GT-DM	92.68	91.79	92.18	–	92.22
No aggregation	No refinement	83.16	87.44	69.91	5.74	80.17
	Proposed refinement	**92.81**	92.13	**92.76**	21.64	92.57
Cross-based aggregation	No refinement	84.69	87.42	69.81	17.93	80.64
	Proposed refinement	92.71	**92.44**	92.67	**45.56**	**92.61**

The MOS values for the Middlebury datasets vary less than those for the self-recorded dataset. This is based on the fact that a properly exposed LDR image of the Middlebury datasets covers a large portion of the dynamic range of the scene. However, because of the bright light bulb in 'IIS Jumble' bigger portions of that image require additional information from the other view.

The reason for two of the Middlebury images having slightly better MOS values for no cost aggregation is also based on the fact that only little information is needed and therefore, filling holes in the disparity map by cost aggregation does not yield a gain. However, for the 'IIS Jumble' dataset the reconstruction of disparity values in large under- or oversaturated regions is necessary. These regions lead to holes in the disparity map which have to be filled by all means in order to achieve good results. Due to the big benefit for datasets with large under- or oversaturated regions, the cross-based aggregation is recommended to be always used together with the proposed refinement steps. On average, the proposed framework achieves a gain of 52 points in the MOS compared to the single-view case with camera motion for the Middlebury datasets. The peak gain is 55.64 points.

5. Conclusion

In this contribution, a stereo high dynamic range imaging framework was presented, which is capable of handling dynamic scenes and video. Beyond that the proposed framework is easily extendable to an arbitrary number of cameras. Nevertheless, this stereo case can be regarded as an important case as it is a good representative for off the shelf low cost solutions. The main idea of this framework is to apply a spatial instead of a temporal exposure bracketing. However, this leads to two challenges. On the one hand, typical stereo matching algorithms rely on the brightness constancy assumption which does not hold for differently exposed views. On the other hand, HDR image reconstruction requires perfectly aligned images. The modules in the proposed framework were selected to alleviate these challenges. First of all, the camera responses are computed in an offline step which allows for robust conversion to radiance space and makes it possible to employ camera setups with different hardware elements. In the main workflow the input images are rectified before performing the disparity estimation. As matching cost for the stereo matching the exposure-invariant ZNCC is applied. Then, instead of simply matching the differently exposed LDR images, the matching is performed on the grayscale radiance space images. Furthermore, a cross-based cost aggregation, a local "winner-take-all" optimization and a set of disparity refinement steps are done. After the disparity estimation, a backward image warping and the presented HDR generation are conducted. The introduction of a newly proposed masking step with the aid of a saturation map during the HDR generation stage leads to improved results. Testing was done on the three Middlebury datasets 'Aloe', 'Moebius', and 'Art' and on the two self-recorded datasets 'IIS Jumble' and 'LMS Pillars'. Besides visual quality assessment, objective quality was evaluated with the help of the quality metric HDR-VDP-2, which yields a predicted mean opinion score. To the best of our knowledge, an objective quality assessment of the HDR images, obtained by a multi-view approach, is done for the first time in this contribution. The proposed framework yields an average gain of 52 points in the MOS compared to single-view HDR with camera motion for the Middlebury datasets.

Due to the large variety of options for disparity estimation, image warping, and HDR reconstruction, there are a lot of possibilities for improvements and future work. A first idea would be to use a semiglobal or global disparity optimization approach instead of the local one. The ZNCC could also be supplied with an adaptive support window to reduce fattening effects. Although the framework at hand is ready for more than two views, all tests and evaluations so far were limited to the stereo case. To further improve image quality, a cancellation of small erroneous regions in the disparity maps could be applied. Research on this multi-view HDR topic is ongoing.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version of http://dx.doi.org/10.1016/j.image.2013.08.016.

References

[1] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics), Morgan Kaufmann, 2005.
[2] J. Guo, S. Sonkusale, A high dynamic range CMOS image sensor for scientific imaging applications, IEEE Sensors Journal 9 (10) (2009) 1209–1218.

[3] P.E. Debevec, J. Malik, Recovering high dynamic range radiance maps from photographs, in: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1997, pp. 369–378.
[4] M. Robertson, S. Borman, R. Stevenson, Dynamic range improvement through multiple exposures, in: Proceedings of International Conference on Image Processing (ICIP), vol. 3, Kobe, Japan, 1999, pp. 159–163.
[5] S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High dynamic range video, ACM Transactions on Graphics 22 (3) (2003) 319–325.
[6] Middlebury Stereo Vision Page, ⟨http://vision.middlebury.edu/stereo/⟩.
[7] N. Sun, H. Mansour, R. Ward, HDR image construction from multi-exposed stereo LDR images, in: Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, China, 2010, pp. 2973–2976.
[8] A. Troccoli, S.B. Kang, S. Seitz, Multi-view multi-exposure stereo, in: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission, Chapel Hill, North Carolina, USA, 2006, pp. 861–868.
[9] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, X. Zhang, On building an accurate stereo matching system on graphics hardware, in: Proceedings of IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 2011, pp. 467–474.
[10] F. Kainz, R. Bogart, Technical Introduction to OpenEXR, Industrial Light & Magic (February 2009).
[11] A. Bovik, Handbook of Image and Video Processing, Academic Press, San Diego, 2000.
[12] H. Hirschmüller, D. Scharstein, Evaluation of stereo matching costs on images with radiometric differences, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (9) (2009) 1582–1599.
[13] Y.S. Heo, K.M. Lee, S.U. Lee, Robust stereo matching using adaptive normalized cross-correlation, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2011) 807–822.
[14] K. Ambrosch, C. Zinner, H. Leopold, A miniature embedded stereo vision system for automotive applications, in: Proceedings of IEEE 26th Convention of Electrical and Electronics Engineers in Israel (IEEEI), Eilat, Israel, 2010, pp. 786–789.
[15] M. Bleyer, S. Chambon, Does color really help in dense stereo matching?, in: Proceedings of International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), Paris, France, 2010, pp. 1–8.
[16] D. Scharstein, R. Szeliski, R. Zabih, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, in: Proceedings of IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV), Kauai, Hawaii, 2001, pp. 131–140.
[17] M. Gong, R. Yang, L. Wang, M. Gong, A performance study on different cost aggregation approaches used in real-time stereo matching, International Journal of Computer Vision 75 (2007) 283–296.
[18] K. Zhang, J. Lu, G. Lafruit, Cross-based local stereo matching using orthogonal integral images, IEEE Transactions on Circuits and Systems for Video Technology 19 (7) (2009) 1073–1079.
[19] H. Hirschmüller, Stereo processing by semiglobal matching and mutual information, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 328–341.
[20] D. Tian, P.-L. Lai, P. Lopez, C. Gomila, View synthesis techniques for 3D video, in: Proceedings of SPIE Optical Engineering+Applications, International Society for Optics and Photonics, San Diego, California, USA, 2009, p. 74430T.
[21] R. Mantiuk, G. Krawczyk, R. Mantiuk, H.-P. Seidel, High dynamic range imaging pipeline: perception-motivated representation of visual content, in: B.E. Rogowitz, T.N. Pappas, S.J. Daly (Eds.), Human Vision and Electronic Imaging XII, Proceedings of SPIE, vol. 6492, San Jose, California, USA, 2007.
[22] H. Hirschmüller, D. Scharstein, Evaluation of cost functions for stereo matching, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, Minnesota, USA, 2007, pp. 1–8.
[23] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, Madison, Wisconsin, USA, 2003, pp. I–195 – I–202.
[24] R. Mantiuk, K. Myszkowski, H.-P. Seidel, A perceptual framework for contrast processing of high dynamic range images, in: Proceedings of the Second Symposium on Applied Perception in Graphics and Visualization (APGV), ACM Press, New York, NY, USA, 2005, pp. 87–94.
[25] Blender, ⟨http://www.blender.org/⟩.
[26] HDRI-Hub.com, ⟨http://www.hdri-hub.com/free-samples/⟩.
[27] E. Reinhard, M. Stark, P. Shirley, J. Ferwerda, Photographic tone reproduction for digital images, ACM Transactions on Graphics 21 (3) (2002) 267–276.
[28] HDR Video: LMS Pillars, ⟨http://www.lms.lnt.de/forschung/arbeitsgebiete/video/vanal/hdrvideo.php⟩.
[29] R. Mantiuk, K.J. Kim, A.G. Rempel, W. Heidrich, HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Transactions on Graphics 30 (4) (2011). 40:1–40:14.
[30] Z. Wang, E. Simoncelli, A. Bovik, Multiscale structural similarity for image quality assessment, in: Proceedings of Conference Record of the 37th Asilomar Conference on Signals, Systems and Computers, vol. 2, Pacific Grove, California, USA, 2003, pp. 1398–1402.
[31] H. Yeganeh, Z. Wang, Objective assessment of tone mapping algorithms, in: 17th IEEE International Conference on Image Processing (ICIP), 2010, pp. 2477–2480.
[32] H. Yeganeh, Z. Wang, Objective quality assessment of tone-mapped images, IEEE Transactions on Image Processing 22 (2) (2013) 657–667.