Enabling Stereoscopic High Dynamic Range Video

Elmedin Selmanovic[a,*], Kurt Debattista[a], Thomas Bashford-Rogers[a], Alan Chalmers[a,b]

[a] *WMG, University of Warwick, United Kingdom*
[b] *goHDR Ltd., United Kingdom*

Abstract

Stereoscopic and high dynamic range (HDR) imaging are two methods that enhance video content by respectively improving depth perception and light representation. A large body of research has looked into each of these technologies independently, but very little work has attempted to combine them due to limitations in capture and display; HDR video capture (for a wide range of exposure values over 20 f-stops) is not yet commercially available and few prototype HDR video cameras exist. In this work we propose techniques which facilitate stereoscopic high dynamic range (SHDR) video capture by using an HDR and LDR camera pair. Three methods are proposed: one based on generating the missing HDR frame by warping the existing one using a disparity map; increasing the range of LDR video using a novel expansion operator; and a hybrid of the two where expansion is used for pixels within the LDR range and warping for the rest. Generated videos were compared to the ground truth SHDR video captured using two HDR video cameras. Results show little overall error and demonstrate that the hybrid method produces the least error of the presented methods.

Keywords: Stereoscopy, High Dynamic Rage, Imaging, Stereo Correspondence, Expansion Operators

1. Introduction

One of the major goals of digital imaging is representing an imaged scene to a human observer in the most realistic manner possible. Ultimately, the presented scene should be indistinguishable from reality. Current imaging techniques are able to convey the appearance of a real scene to an extent, but they are still limited in a number of aspects including accurate lighting and depth reproduction. More recent techniques such as stereoscopic imaging and high dynamic range imaging overcome some of these limitations.

The traditional representation of depth, present in paintings, photographs, and television relies only on monoscopic cues and thus does not render a realistic representation of a scene. Stereoscopy improves the situation by providing a significantly improved perception of depth. Both stereoscopic video cameras [1] and stereoscopic displays [2] have recently become available to consumers. Stereoscopy has been shown to provide benefit in a number of tasks including position and distance judgement, identifying objects, spatial manipulation of objects, navigation, and spatial understanding [3]. Application areas of stereoscopy include medicine [4], military [5], entertainment, industrial computer aided design [6] and photogrammetry [7].

HDR imaging is a relatively novel imaging technique which enables capture of all the lighting visible to the human eye and more [8, 9]. Traditional imaging methods are unable to capture such a luminance range and are hence termed low dynamic range (LDR). They contain underexposed and overexposed regions which lack any detail, and data in those regions is lost. This is caused by the limitations of the capturing and display devices. Native HDR video capture is limited to the research community where expensive camera prototypes are being developed [10, 11]. Even traditional viewing benefits from HDR content; methods called tone mappers scale down dynamic range of the HDR content to fit that of LDR displays attempting to preserve as much visual information as possible [12, 8]. HDR imaging has a number of application areas including physically-based rendering, remote sensing, digital photography, image and video editing, entertainment industry, virtual reality, computer vision and security [9].

Stereoscopic high dynamic range (SHDR) imaging brings these two technologies together thereby enabling content with an unprecedented level of realism. It can deliver advantages common to both technologies: improved depth perception and realistic lighting representation of a scene. Currently there is no camera which would natively capture SHDR content, file formats and compression methods for SHDR are lacking and displays which would show SHDR are unavailable.

This paper is concerned with the capture of SHDR videos. The straightforward approach would use two HDR cameras mounted side-by-side. Currently, however, HDR cameras are expensive and bulky so combining two of them is unfeasible. Two views in a stereo pair are similar and there is significant data overlap between them which can be exploited to avoid using two HDR sensors. In this work we propose generating SHDR video from an HDR-LDR video pair. Figure 1 outlines the concept. Selmanovic et al. [13] demonstrated that it is possible to generate SHDR content from an HDR-LDR pair. In a user study they demonstrated that participants could not

*Corresponding author
Email address: elmedin.selmanovic@warwick.ac.uk (Elmedin Selmanovic)

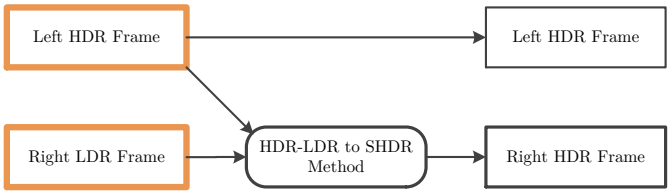


Figure 1: SHDR video is generated from an HDR-LDR video pair. The original HDR video is unmodified and represents one view while the second HDR view is obtained from the LDR video using the proposed methods.

notice any significant difference between an HDR-HDR pair computed from HDR-LDR and a ground truth HDR-HDR reference. However, that work was limited to static images. This paper extends that work to compute SHDR video.

The most successful method presented by Selmanovic et al. [13], based on using stereo correspondence, is extended to generate SHDR video here, and while found to perform well, it does suffer from some issues when extended into the temporal domain. In this paper a further two methods are presented. One uses a novel expansion operator which extends the dynamic range of the LDR image based on the HDR data. The second approach combines the other two methods and exploits the advantages of both. It uses stereo matching for overexposed and underexposed pixels and expands the rest. These three methods were compared to ground truth SHDR videos where both views were captured using an HDR camera. Objective measurements tested temporal and spatial qualities of all the methods over five video sequences.

2. Related Work

Three major concepts need to be considered when developing a technique for generating SHDR video: image and video capture using asymmetric sensors, image plus depth video capture, and stereoscopic high dynamic range imaging.

2.1. Asymmetric Sensor Imaging

The concept of using camera sensors of different qualities and combining them to produce enhanced viewing experience has been employed before. Sawhney et al. [14] proposed a method in which high spatial resolution stereoscopic video (at least 6,000 horizontal pixels) was generated from a pair of videos with asymmetric resolution. One video was of target resolution (e.g. $6,000$ pixels) while the other was typically a quarter of the size. Stereo correspondences between two frames in a pair were calculated. Neighbouring frames helped increase the robustness of the correspondences especially for occluded regions and an alignment map identified mismatches. The high resolution image was then warped using a disparity map to generate a novel view and mismatched pixels were obtained from the upsampled low resolution image. Two video sequences were used to test the approach. The number of misaligned pixels was used as the objective performance measure. Results were given for a single frame from each sequence and less than 1.4% misaligned pixels were reported for both.

Lo et al. [15] tested whether rendering times of stereo images could be decreased by reducing the resolution of one of the views while still preserving the same image appearance. The approach utilised binocular fusion - a process in which the HVS fuses two percepts presented to each eye into a single view. A single rendered scene was used for testing. In the stereoscopic asymmetric test condition, the resolution of one image in the pair was decremented in 10 steps generating 10 novel images while the other was kept at the maximum. Such images were compared to the ground truth (GT) - a stereo image were both views had maximum resolution of 800×800 pixels. Results showed that less than 15% of participants could differentiate between the GT and the image pair with one image at the reduced resolution of 640×640. Once the image was reduced to 320×320, more than 50% of participants could detect the difference.

Bhat et al. [16] suggested an approach for enhancing low quality video using high quality photographs. Correspondence between image and video data was found using multiview stereo, and structure from motion algorithms. Then, the spatial and temporal gradient fields were used to transfer properties of photographs (e.g. spatial resolution, lighting and dynamic range) onto a video. The technique could also be used for object removal, shake reduction and more efficient video editing. However, the method was limited to static scenes only. The authors reported very slow computation speeds (five minutes per single low resolution image), but suggested that they could be improved. Quantitative results were lacking and only one image to illustrate each application was provided.

A large scale multiple sensor approach was proposed by Wilburn et al. [17]. They constructed a matrix of low quality cameras whose outputs were combined to produce high quality video (comparable to the one captured using expensive high end consumer products). Applications of this approach included video of increased resolution, better frame-rate and higher dynamic range. It could also simulate camera motion and large camera aperture. The method could potentially be modified to capture SHDR video, but this was not examined. The system consisted of 100 camera sensors, lenses and processing boards. These were connected and controlled by four PCs. Engineering an entire system may require considerable assembly rendering it impractical for everyday situations. Both compressed and uncompressed video could be stored before processing and the authors reported that two and a half minutes required 2 GB when MPEG compressed.

2.2. Image Plus Depth Capture

Imaging sensors have also been combined with depth sensors allowing for image based rendering of the second view (or multiple views). ZCam [18] measured the time that projected infra-red light took to reflect back to the sensor. This measurement inferred the distance of objects to the camera. Similarly, Kawakita et al. [19] used an infra-red LED array for capturing depth with their HDTV Axi-vision camera at a resolution of more than 920,000 pixels at 30 frames per second. Alternative methods described by Scharstein and Szeliski [20] project a structured light pattern onto a scene. Shapes and distances

of the objects caused distortions in the pattern which was analysed to obtain depth. A limitation of all the mentioned methods is the range and the precision of captured depth data. For example, ZCam had range from 1 to 10 m with resolution of 0.5 cm (for distances of 1 m) while Axi-vision had unreported range with resolution of 1.7 cm (for distance of 2 m).

2.3. Stereoscopic High Dynamic Range Imaging

Lin and Chang [21] suggested a method for creating HDR images using stereo. An image pair was taken at a different exposure levels and combined to generate HDR images using stereo correspondence. Capturing SHDR by modifying the method would required only two LDR cameras making the approach appealing, inexpensive and practical. However, to generate a reliable disparity map required image warping, the number of over- and under-exposed pixels had to be minimised, which limited the potential dynamic range of the generated HDR image. This was shown in the examples provided by the authors.

SHDR imaging was first proposed by Selmanovic et al. [22]. In their work they examined five different methods for compressing SHDR data. All were backwards compatible with traditional and LDR stereo image viewers. Initially, each image in the pair was compressed using JPEG-HDR [23] coding but similar approaches could have been used as well. After that initial step, two of the methods relied on LDR stereo techniques to store images in a side-by-side and half side-by-side fashion. The other two methods used image-based rendering and exploited low frequency, low range and single channel attributes of the disparity map for coding. The final technique relied on motion compensation and produced the best quality per bit rate results.

Selmanovic et al. [13] have explored how to generate static SHDR images from an HDR-LDR camera pair. They proposed two general approaches and tested two methods for each. The first approach was based on expansion operators where the dynamic range of the LDR image was expanded. Parameters used for expansion were set using the information present in the HDR image. Two expansion operators were tested: *linear scaling* [24] and the *expand map* technique [25]. The second approach relied on stereo correspondences. A single exposure was extracted from the HDR image, so its dynamic range matched that of the LDR one. Then pixel correspondences between the two LDR images were found and the disparity map was obtained. A new HDR image was generated by warping the existing HDR image using image-based rendering techniques. Two stereo matching algorithms were tested: *sum of absolute differences (SAD)* [26] and a *correspondence with occlusion via graph cuts (COGC)* technique [27]. In a user study all four techniques were compared to the ground truth (GT)- both images of stereo pair captured using an HDR camera. Results showed that the SAD technique was indistinguishable from GT and the second closest was COGC. Expansion operator techniques performed worse than stereo correspondence ones and took the last two places. Also it was shown that objective and subjective measurements were correlated.

3. LDR to HDR Methods

As discussed previously, the aim of this work is to generate SHDR video from an HDR-LDR stereo video pair. In essence, one HDR view is missing and needs to be reconstructed using the available LDR video as a guidance. The LDR image captures only a subset of the full range, it is scaled and quantised appropriately into an 8-bit range (per channel). Generating an HDR frame from the LDR frame is an ill-posed problem for which an exact solution cannot be found as the required data is missing and can only be estimated.

Overexposed and underexposed regions in the LDR stream are outside the captured range, lack any information, and are difficult to reconstruct on their own. Out-of-range pixels may be recovered from the HDR stream which may contain those missing regions. However, the two streams are not aligned, so mapping between pixels in the left and the right image depends on the depth of the imaged object from the camera.

We propose three techniques for generating SHDR video from an HDR-LDR stereo video pair. Two of them tackle a subset of challenges discussed above. A stereo matching approach utilises the disparity map to generate HDR frames while an approach based on expansion operator is concerned with mapping LDR to HDR values. The last technique combines the merits of both achieving the best reconstruction of HDR data as shown in Section 4.

3.1. Stereo Correspondence (SAD)

The stereo correspondence approach relies on a disparity map to transfer data between HDR and LDR view. Imaged objects are projected to pixels which are horizontally offset in the stereo image pair depending on their distance from the camera. The disparity map provides these offsets and hence connects pixels representing the same 3D point in both views. This allows for the correlation of HDR data from one image with its counterpart in the LDR image making it possible to transfer values.

The calculation of the disparity map is a challenging problem and hundreds of methods have been proposed [28]. In general, they compare differences of pixel intensities between the two views and try to minimise these by offsetting regions in one of the views.

In the study by Selmanovic et al. [13] the best identified method was the *sum of absolute differences (SAD)* method. For each pixel SAD finds the correspondence by looking for the pixel in the other image which generates the smallest absolute difference between the two. In order to reduce ambiguity, a window of neighbouring pixels is used. The error of selecting a specific pixel can be formally expressed as:

$$SAD(x,y) = \sum_{k\in R,G,B} \sum_{(i,j)\in W(x,y)} |I_{k,1}(x+i,y+j) - I_{k,2}(x+d_x+i,y+j)| \quad (1)$$

where $W(x,y)$ are pixel coordinates of a window located at (x,y), $I_{k,l}(x,y)$ are the intensity values of k-th channel of l-th image at (x,y), d_x is a horizontal image disparity, and $SAD(x,y)$ is the value representing the difference between the compared

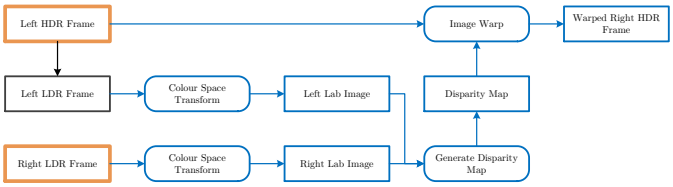


Figure 2: HDR-LDR stereo correspondence pipeline finds spatial matches between HDR and LDR pixels and uses them to guide warping of existing HDR image, thereby generating a novel HDR view.

regions. The disparity d is selected using a winner-takes-all (WTA) technique where pixel generating smallest SAD cost is chosen.

A detailed pipeline for generating an HDR image using the SAD method is shown in Figure 2. Without any loss of generality the left frame is considered HDR and the right is considered LDR. First a single exposure is selected from the HDR frame to match that of the LDR one. This is achieved by minimising the difference between histograms of the existing and extracted LDR image. Values obtained in this step can be transferred to the next frames to speed up the process. Both LDR frames are then transformed to Lab colour space which approximates human vision and aspires to perceptual uniformity. This means that the differences between channels of the left and right frame are related to perceptual differences; such differences are more perceptually accurate than if RGB space were used. Next, the SAD algorithm is used to compute the disparity map between stereo frames. The disparity map then guides image warping [29] of the available HDR image thereby generating a novel view.

The SAD stereo matching algorithm can compute the disparity map in real-time on a standard PC. The technique transfers actual HDR values to the new position in the other view and so avoids intensity quantisation. While the calculated disparity map can be noisy and incorrect offsets can be present, the algorithm always connects pixels which are close in intensity making it particulary efficient for the generation of the novel stereo view. Selmanovic et al. [13] suggest that such approach for stereo matching was why SAD technique outperformed the other methods in their study.

The overexposed and underexposed pixels are also transferred but can end up in the wrong position. As all the values in those regions have the same value (0 or 255) it is not possible to perform accurate matching. This is especially the case for larger regions where, even with increased window size, it may not be possible to find a pixel within the captured range. Disparity maps for such areas contain constant values. The first tested disparity value is selected by the WTA technique as all the others have the identical SAD cost. Details in these regions are present but may be out of place, and may be perceived as being at the incorrect depth (Figure 11, inset B). Another challenge it that of representing view dependent phenomena, such as occlusion (Figure 11, inset A), reflective objects, and specular highlights. Data for those might be missing from one of the views. However, SAD finds perceptually close intensities for those pixels (albeit from the spatially incorrect positions) which can alleviate the problem to an extent. Such mistakes were not perceived due to binocular fusion for static images [13], but they will cause temporal noise for videos as they are not temporally consistent (Figures 9c and 9f). Thus, the main disadvantage of this approach is potential temporal incoherence due to incorrect disparity matches.

3.2. Expansion Operator (EO)

Expansion operators take an LDR image as an input and produce an HDR image as the output. Multiple approaches have been proposed [8]. One of the state-of-the-art operators - Banterle et al.'s [25] inverse tone mapper that was evaluated as the best expansion operator in a user study [30] - did not perform well when converting HDR-LDR image pairs to SHDR images [13]. When expanding the image such operators take a small number of parameters (e.g. three in the case of the tested one [25]) which control overall brightness of the final image and its peak value. While this is convenient method of adjusting the output, the lack of control means that expanded images are less likely to correspond to the actual values of the imaged scene. For example, the peak luminance parameter influences range and brightness of the expanded image, but such a value is frequently a result of noise (when original HDR images are captured). So using some existing HDR image as the means of setting this parameter is unlikely to produce appealing results so user input is required. Expansion operators are not very suitable for reconstruction of the LDR view for SHDR because discrepancies from the original can be large, and not possible to fuse through binocular single vision [13]. However, for the HDR-LDR pair case, it is possible to create an expansion operator by finding a mapping between the original HDR and LDR values by using the HDR as a reference. The problem is similar to the one faced by Mantiuk et al. [31] where HDR video was compressed by using a residual stream together with a tone mapped stream. It was assumed that TM operator was unknown so the correspondence between the HDR and tone mapped values had to be calculated for the decoder. As HDR and LDR pixels are spatially aligned it was possible to put HDR values into the corresponding 256 LDR bins. As this is a many-to-one mapping multiple HDR values would be assigned to single bin. Mantiuk et al. [31] used the arithmetic mean to find a single value. We extend and modify the approach to generate an HDR image from the LDR one. The reconstruction function (RF) which maps LDR to HDR values is calculated as follows. All the HDR values are ordered. Then, an HDR histogram with 256 bins is created to emulate the LDR one, by putting the same number of HDR values into each bin as there are LDR values in that bin. Formally, this is expressed as:

$$RF(c) = \frac{1}{Card(\Omega_c)} \sum_{i=M(c)}^{M(c)+Card(\Omega_c)} c_{\mathrm{hdr}}(i) \qquad (2)$$
$$\text{where } \Omega_c = \{j = 1..N : c_{\mathrm{ldr}}(j) = c\}$$

$c = 0..255$ is an index of a bin Ω_c, $Card(\cdot)$ is the cardinality function which returns the number of elements in the bin, N

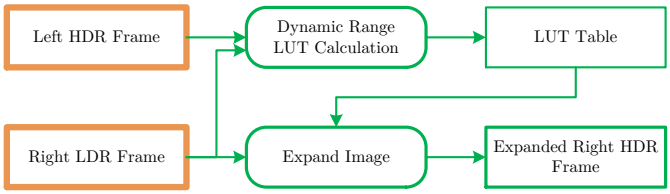


Figure 3: HDR-LDR expansion operator pipeline calculates intensity correspondences between LDR and HDR values and saves them in the look-up table. Expansion is performed by assigning HDR values from the table to the LDR ones.

is the number of pixels in a frame, $c_{\text{ldr}}(j)$ are channel intensity values of the j-th LDR pixel, $M(c) = \sum_0^c Card(\Omega_c)$ is the number of pixels in the previous bins and c_{hdr} are channel intensity values of all HDR pixels sorted in ascending order.

The pipeline to generate an HDR image using this approach is shown in Figure 3. The look-up table (LUT) is calculated using Equation 2. Once the LUT is obtained expansion can be performed quickly in a straightforward manner where each LDR value is assigned a corresponding HDR value from the table. It is possible to re-use the LUT across frames and it can be used to improve temporal quality by filtering.

The proposed method of generating SHDR video from HDR-LDR video stream using expansion operator is quick and can be implemented in real time. Expansion is not view dependent and does not suffer from the same problems stereo matching would in occluded regions. Generated HDR values only depend on captured HDR and LDR streams which are temporally coherent and the method is not expected to introduce flickering. The main drawback of this approach is the lack of a facility to explicitly handle overexposed regions which are of constant, maximum value without any detail. While fusion can also help in those areas differences are frequently high and noticeable (Figure 11, insets B, C, D and E).

3.3. Hybrid Method (HY)

The methods described above both have distinct sets of advantages and drawbacks. Hence, we propose a novel method which combines the two, trying to obtain benefits of both while minimising disadvantages. Effectively, the hybrid method attempts to do well in in-range regions using techniques based on the expansion operator, for example occluded regions are handled more robustly. It also is able to handle out-of-range pixels using methods based on SAD, albeit with a further correction step.

An overview of the technique is displayed in Figure 4. Both an expanded HDR frame and a warped HDR frame are generated. Overexposed and underexposed regions are identified using thresholding of the LDR frame, where a pixel is deemed out of range if the value of one channel is above or below a predefined threshold (e.g. above 250 or below 5). The out-of range pixels are assigned the warped frame data. The rest of the pixels are taken from the expanded HDR frame. The expanded frame is generated using the same approach described above while the

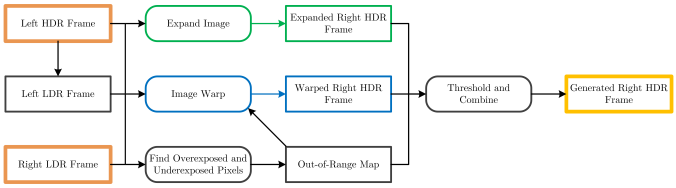


Figure 4: The broad pipeline showing generation of right HDR frame given left HDR frame and right LDR frame. Dynamic range expansion and stereo matching techniques are combined using out-of-range map as a threshold.

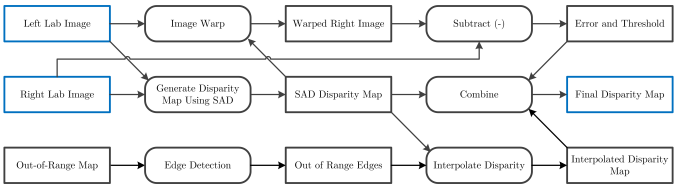


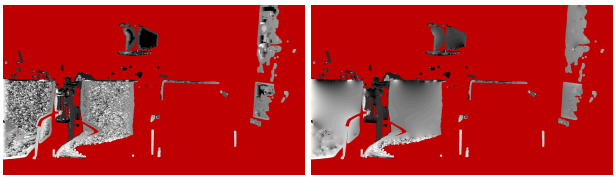
Figure 5: Modified disparity map generations interpolates disparities for the out-of-range region from its neighbours. Disparities generated in this manner that are likely to cause artefacts are replaced by the SAD generated ones.

SAD pipeline is adapted so it reconstructs out of range regions more accurately.

As described in the previous section, overexposed and underexposed regions lack any data and as such cannot be matched. SAD method assigns the first tested disparity value to those regions, which for some cases may be correct. However, it is more likely that out-of-range areas have disparities similar to the neighbouring ones. For this reason, the hybrid method interpolates disparities for overexposed and underexposed regions from well exposed edges. This modified stereo correspondence path is shown in Figure 5.

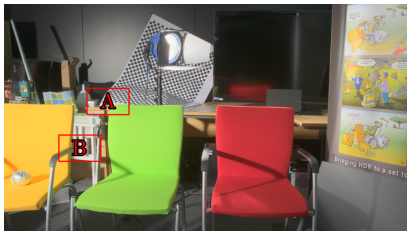
Stereo matching is performed on the images transformed to the Lab colour space in the same manner described above. In addition, a map identifying out-of-range regions (generated by thresholding) is used as an input. The edges of overexposed and underexposed areas are found using any of the edge detection methods. In our implementation we use a morphological edge detection technique [32] because of its speed. Image dilation expands overexposed and underexposed regions in out-of-range map. Edges are found by subtracting the original out-or-range map form the dilated one. The edges are on the outside of the out-of-range regions where robust disparities values are expected. Once edge pixels are identified, smooth interpolation is performed inward. A comparison of the map generated using SAD and the interpolated approach are shown in Figure 6.

During interpolation, foreground objects can influence disparity values of out-of-range background areas and viceversa. This may result in artifacts around such objects as values of foreground objects being transferred to the background, as shown in Figures 7b and 7c. Such artifacts are identified by warping the extracted exposure of the HDR image using the interpolated disparity map and subtracting it from the original LDR image. Differences above the provided threshold are recognised as artifacts. In order to correct for these pixel dis-



(a) SAD Disparities (b) Interpolated Disparities

Figure 6: The disparities for overexposed regions generated using SAD method (a) are less smooth compared to ones obtained using interpolation (b).



(a) GT

(b) A: Artifacts

(c) B: Artifacts

(d) A: Corrected

(e) B: Corrected

Figure 7: Artifacts caused by interpolation are identified and corrected.

parities computed by SAD are used instead. SAD matches are also potentially incorrect as they connect overexposed pixels. However the error in intensity will likely be smaller than the one caused by transferring well-exposed values from the foreground object. Results of this correcting step are shown in Figures 7d and 7e.

4. Results

In order to demonstrate the efficacy of the proposed methods the methods are compared with each other and GT. The way in which the GT videos were obtained is explained next, after which the results of quality evaluation are provided.

4.1. Materials

Ground truth SHDR videos, consisting of HDR-HDR video pairs, had to be obtained in order to enable comparison with the proposed methods. As mentioned in the introduction, camera systems which record two native HDR videos simultaneously do not currently exist and are currently difficult to construct. In order to overcome that challenge we employed three techniques for capturing SHDR video data.

(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5

Figure 8: The example frame from each of the tested SHDR scenes. For illustrative purposes frames are tone-mapped and displayed as anaglyph stereo.

Two static scenes (*Scene 1* and *Scene 2*) were recorded using stop motion by mounting a camera (Canon 1Ds Mark II) on rails and moving it laterally, in small steps (0.5 cm). At each step seven exposures separated by 2 stops were captured and later merged to produce individual SHDR video frames. As the movement was horizontal and orthogonal to the optical axis it was possible to obtain both views. Video for one eye was delayed by 13 frames which corresponded to a camera shift of 6.5 cm - approximation of average interocular distance.

Scene 3 and *Scene 5* were dynamic and recorded using a native HDR video camera [10]. Two takes, one for each eye, were required as only a single camera was available. Object movement in the scene needed to be exactly repeatable between the takes. To this end a high precision robot arm which folded aluminum sheets, and a disco ball that rotated were recorded.

The final scene was computer generated (*Scene 4*) using a virtual stereo camera rig that output HDR images. Tone mapped frames from each of the sequences are shown in Figure 8

All videos were captured and computed in full high definition (1920×1080 pixels). The two dynamic scenes were captured at 30 frames per second (fps), the computer generated one was at 24 fps, while for static scenes it was possible to choose any frame rate. A robust measurement of the dynamic range was obtained by disregarding the top 1 ‰ and bottom 1 ‰ of the values in the frame; this is used to avoid extreme values caused by noise. It varied between the scenes and individual frames, peeking at 16.6 stops for *Scene 4*, and having minimum at 11.1 stops for *Scene 5*. Length also varied between the sequences where *Scene 5* was the longest containing 720 frames and *Scene 4* was shortest with 240 frames. Videos contained regions which would test the limits of the proposed methods including out-of-range areas, view-dependent phenomena, camera movement and object movement. Data for all sequences is summarised in Table 1

4.2. Objective Quality Measurements

Objective measurements were used to evaluate the quality of each method. To estimate the error of the individual frames peak signal to noise ratio (PSNR) was used. It represents the ratio between the maximum possible value of an image (signal) and the power of noise which affects its quality. The mea-

Table 1: Video data for each sequence: the average and maximum dynamic range (in stops) and the total number of frames

Data	Average DR	Maximum DR	Frame No
Scene 1	12.3	14.1	287
Scene 2	13.4	14.8	432
Scene 3	12.3	12.7	368
Scene 4	16.3	16.6	240
Scene 5	12.1	13.1	720

Table 2: Peak Signal-to-Noise Ratio (higher is better)

Method	HY	SAD	EO
Scene 1	51.88	48.77	48.77
Scene 2	45.77	42.26	40.34
Scene 3	48.76	46.66	45.42
Scene 4	54.26	37.29	34.33
Scene 5	59.14	51.97	48.54
Average	51.96	45.39	43.48

surement is logarithmically scaled making it especially suitable for images of high dynamic range, because the HVS system responds to the intensity of light approximately logarithmically [33, 34]. It was also shown to correlate with subjective measures in case of generating SHDR images from HDR-LDR stereo pair [13]. The averaged values for all the scenes and all the methods are shown in Table 2 where higher value represents better quality. Results for individual frames are presented in Figure 12.

As expected, the hybrid (HY) method outperformed the other two achieving the best score for all tested scenes. The SAD technique achieved better results than the EO one for all the scenes. The score difference was greater between HY and SAD than between SAD and EO.

In order to verify the temporal quality we propose a temporal quality (TQ) metric which is inspired by the metric used in the work of Wan et al. [35]. Initially, images are converted to logarithmic space to account for the perception of the HVS and to avoid bias caused by high intensity values present in HDR images. To find temporal differences in a stream two consecutive frames are subtracted. This is performed for both the GT video and the generated video. Temporal error introduced by the generated video is identified by subtracting temporal differences of the generated stream from the GT stream. The error is then weighted by the quality of the reconstructed frame. This increases the inconsistency when there is a larger difference between generated and GT frames. Finally, values are aggregated across all the pixels as shown in Equation 3:

$$TQ(t) = \frac{1}{Card(N)} \sum_{(x,y)\in N} \omega_{(x,y)} |(\Delta I_1(x,y,t) - \Delta I_2(x,y,t))| \quad (3)$$

where t is the frame number, N is the set consisting of all colour channel values for all pixels in a frame, I_1 is the logarithmically scaled GT frame and I_2 is logarithmically scaled generated frame, $\omega_{(x,y)} = |log(I_1(x,y,t)) - log(I_2(x,y,t))| + 1$ is quality weight, $I(x,y,t)$ is the intensity of a pixel at point (x,y) of the frame t and $\Delta I(x,y,t)$ is difference between logarithmically scaled consecutive frames as shown in Equation 4:

$$\Delta I(x,y,t) = \log(I(x,y,t)+1) - \log(I(x,y,t+1)+1) \quad (4)$$

Table 3: Temporal Quality (lower is better)

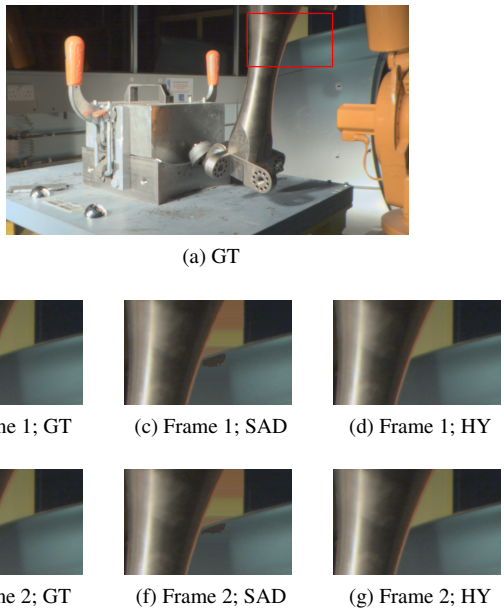
Method	HY	SAD	EO
Scene 1	0.0091	0.0128	0.0156
Scene 2	0.0038	0.0063	0.0072
Scene 3	0.0043	0.0064	0.0074
Scene 4	0.0009	0.0010	0.0014
Scene 5	0.0063	0.0129	0.0110
Average	0.0049	0.0079	0.0085

The summary of TQ values, averaged across the video sequence, are shown in Table 3 where the smaller value represents a better quality. Results for all the frames and all the videos are provided in Figure 13.

Overall, HY method performed best and had the smallest error for all the scenes. SAD technique had better quality than the EO technique for four scenes while EO outperformed SAD for the last scene, which has the smallest average dynamic range. As discussed in Section 3.2, this is expected as the EO method should be, generally speaking, a preferable option to SAD for HDR videos with a lower dynamic range.

Both calculated error metrics only took into account the single generated view. The quality of the existing natively captured HDR view was not taken into the account.

4.3. Qualitative Results



(a) GT

(b) Frame 1; GT (c) Frame 1; SAD (d) Frame 1; HY

(e) Frame 2; GT (f) Frame 2; SAD (g) Frame 2; HY

Figure 9: The SAD method may generate artifacts in the occluded regions which are inconsistent across frames. The HY method recovers these pixels using EO and avoids artifacts.

To complement quantitative metrics, Figure 10 illustrates the different qualities of the methods. The provided example shows

how EO does not manage to reconstruct any details in the out-of-range regions while SAD and HY method appear similar, in general. To show the differences between the two, selected regions are presented in more detail in Figure 11. The inset A shows the lamp leg which is not reconstructed well by the SAD method, due to occlusion. The insets A and D contain occluded areas (along the edges of the chairs), where the SAD method made errors. Due to relying on the EO, HY method is able to preserve the information available in LDR. For the overexposed regions, shown in the insets B, C, D, and E, SAD lacks information required for accurate matches and makes mistakes. The HY method relies on interpolation to obtain disparities from the neighbouring well-exposed pixels and is able to reconstruct these regions successfully. The EO operator, as expected in this case, lacks the required information for reconstruction.

Disparities calculated by SAD methods are noisy in low frequency regions and in occluded areas resulting in artefacts when generating HDR image, as shown in Figures 9c and 9f. HY method achieves temporal consistency by using EO for out-of range pixels (Figures 9d and 9g).

5. Conclusions and Future Work

In this paper we presented a method for capturing SHDR video from HDR-LDR video. Three methods were proposed and as expected the hybrid method outperformed the other two in terms of the scenes shown due to taking advantages of both methods, the expansion method for in-range pixels and the SAD method with a correction step for out-of-range pixels. While the results are good, they are only based on the reconstructed view, so better results would be expected if binocular fusion was taken into account. Since no objective metrics exist to do so, future work will investigate the possibility of a user study similar to the one for static SHDR images [13], to identify if there are further gains in the proposed technique; however, this is not straightforward, as comparing videos with a reference is significantly more complex than comparing static images. Developing a perceptually based metric for measuring spatiotemporal quality of SHDR video is an aim of future work.

This work serves as an enabling method for SHDR video to be adopted without having to await SHDR video capture devices, which may take a while as HDR video is still very much in its infancy.

References

[1] B. Mendiburu, 3D TV and 3D Cinema: Tools and Processes for Creative Stereoscopy, Focal Press, 2011.

[2] H. Urey, K. V. Chellappan, E. Erden, P. Surman, State of the Art in Stereoscopic and Autostereoscopic Displays, Proceedings of the IEEE 99 (4) (2011) 540–555.

[3] J. P. McIntire, P. R. Havig, E. E. Geiselman, What is 3D good for? A review of human performance on stereoscopic 3D displays, in: SPIE Defense, Security, and Sensing, 2012.

[4] D. J. Getty, P. J. Green, Clinical applications for stereoscopic 3-D displays, Journal of the Society for Information Display 15 (6) (2007) 377.

[5] S. Dixon, E. Fitzhugh, D. Aleva, Human Factors Guidelines for Applications of 3D Perspectives: A Literature Review, in: SPIE Defense, Security, and Sensing, Vol. 7327, 2009.

[6] M. E. Brown, J. J. Gallimore, Visualization of three-dimensional structure during computer-aided design, International Journal of Human-Computer Interaction 7 (1) (1995) 37–56.

[7] K. Kraus, Photogrammetry: Geometry from Images and Laser Scans, Walter de Gruyter, 2007.

[8] F. Banterle, A. Artusi, K. Debattista, A. Chalmers, Advanced High Dynamic Range Imaging: Theory and Practice, A K Peters/CRC Press, 2011.

[9] E. Reinhard, W. Heidrich, S. Pattanaik, P. Debevec, G. Ward, K. Myszkowski, High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting, Morgan Kaufmann, 2010.

[10] A. Chalmers, G. Bonnet, F. Banterle, P. Dubla, K. Debattista, A. Artusi, C. Moir, High-dynamic-range video solution, in: ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies, ACM Press, 2009, p. 71.

[11] M. Tocci, C. Kiser, N. Tocci, P. Sen, A Versatile HDR Video Production System, ACM Transactions on Graphics (TOG) 30 (4) (2011) 41–49.

[12] K. Delvin, A. Chalmers, A. Wilkie, W. Purgathofer, Tone reproduction and physically based spectral rendering, Eurographics 2002: State of the Art Reports (2002) 101–123.

[13] E. Selmanovic, K. Debattista, A. Bashford-Rogers, Thomas Chalmers, Generating Stereoscopic HDR Images Using HDR-LDR Image Pairs, ACM Transactions on Applied Perception 10 (1) (2013) 18.

[14] H. Sawhney, Y. Guo, K. Hanna, R. Kumar, S. Adkins, S. Zhou, Hybrid stereo camera: an IBR approach for synthesis of very high resolution stereoscopic image sequences, in: Proceedings of the 28th annual conference on computer graphics and interactive techniques, ACM, 2001, pp. 451–460.

[15] C.-H. Lo, C.-H. Chu, K. Debattista, A. Chalmers, Selective rendering for efficient ray traced stereoscopic images, The Visual Computer 26 (2) (2009) 97–107.

[16] P. Bhat, C. Zitnick, N. Agarwala, M. Agrawala, M. Cohen, B. Curless, S. Kang, Using Photographs to Enhance Videos of a Static Scene, in: Eurographics Symposium on Rendering, 2007, pp. 327—-338.

[17] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, M. Levoy, High Performance Imaging Using Large Camera Arrays, ACM Transactions on Graphics 24 (3) (2005) 765–776.

[18] G. J. Iddan, G. Yahav, Three-dimensional imaging in the studio and elsewhere, in: Proc. SPIE 4298, Three-Dimensional Image Capture and Applications IV, 2001, pp. 48–55.

[19] M. Kawakita, T. Kurita, H. Kikuchi, S. Inoue, HDTV axi-vision camera, in: Proc. of International Broadcasting Conference, no. 8, 2002, pp. 397–404.

[20] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, pp. I–195 – I–202.

[21] H. Lin, W. Chang, High dynamic range imaging for stereoscopic scene representation, in: 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 4305–4308.

[22] E. Selmanovic, K. Debattista, T. Bashford-Rogers, A. Chalmers, Backwards Compatible JPEG Stereoscopic High Dynamic Range Imaging, in: Theory and Practice of Computer Graphics, 2012.

[23] G. Ward, JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG, in: ACM SIGGRAPH 2005 Courses, 2005, p. 8.

[24] A. Akyuz, R. Fleming, B. Riecke, E. Reinhard, H. Bulthoff, Do HDR displays support LDR content?: a psychophysical evaluation, in: ACM SIGGRAPH, ACM, 2007, pp. 38–44.

[25] F. Banterle, P. Ledda, K. Debattista, A. Chalmers, Inverse tone mapping, Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia - GRAPHITE '06 (2006) 349.

[26] B. Cyganek, J. P. Siebert, An Introduction to 3D Computer Vision Techniques and Algorithms, John Wiley & Sons, Ltd, Chichester, UK, 2009.

[27] V. Kolmogorov, R. Zabih, Computing visual correspondence with occlusions using graph cuts, Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 (2001) 508–515.

[28] D. Scharstein, R. Szeliski, R. Zabih, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (1) (2001) 131–140.

[29] C. Fehn, Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV, in: Electronic Imaging 2004,

2004, pp. 93–104.

[30] F. Banterle, P. Ledda, K. Debattista, M. Bloj, A. Artusi, A. Chalmers, A psychophysical evaluation of inverse tone mapping techniques, Computer Graphics Forum 28 (1) (2009) 13–25.

[31] R. Mantiuk, A. Efremov, K. Myszkowski, H.-p. Seidel, Backward compatible high dynamic range MPEG video compression, ACM Transactions on Graphics (TOG) 25 (3) (2006) 713–723.

[32] J. Lee, R. Haralick, L. Shapiro, Morphologic edge detection, Robotics and Automation, IEEE Journal of 3 (2) (1987) 142 – 156.

[33] E. H. Weber, De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae, 1834.

[34] G. T. Fechner, Ueber eine Scheibe zur Erzeugung subjectiver Farben, Annalen der Physik und Chemie 121 (10) (1838) 227–232.

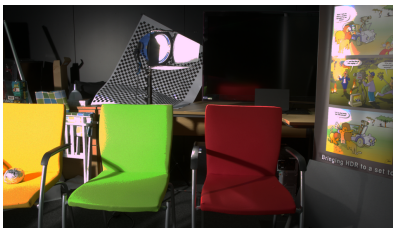
[35] L. Wan, S.-k. Mak, T.-t. Wong, Spatiotemporal Sampling of Dynamic Environment Sequences, Visualization and Computer Graphics, IEEE Transactions on 17 (10) (2011) 1499–1509.

GT

EV: +4; EO

EV: +4; SAD

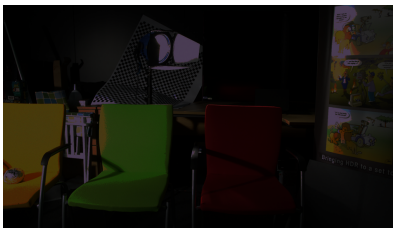
EV: +4; HY



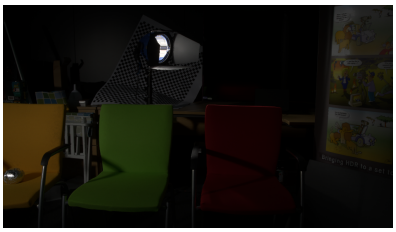
EV: 0; EO

EV: 0; SAD

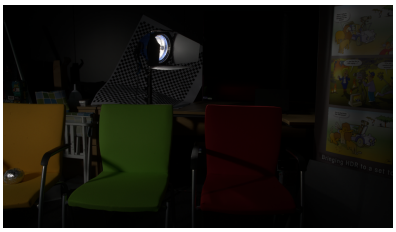
EV: 0; HY



EV: -4; EO

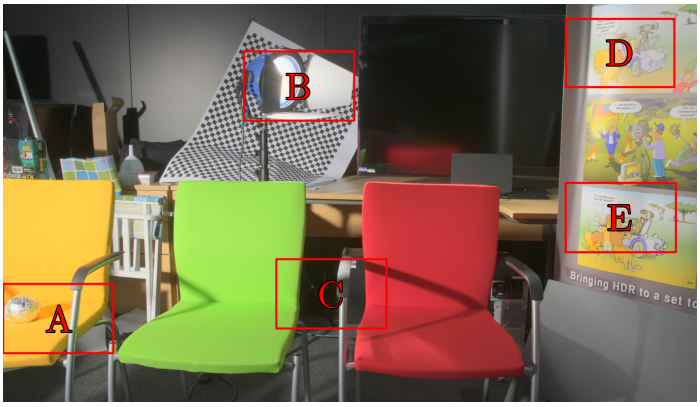


EV: -4; SAD



EV: -4; HY

Figure 10: The reconstructed frame from the SHDR pair for all methods for Scene 1 are presented. GT is tone mapped. For each method three single exposures are selected and shown.



Scene 1

A: GT

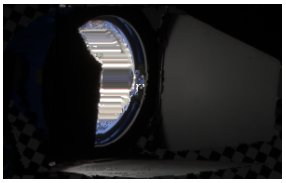
A: EO

A: SAD

A: HY

B: GT

B: EO



B: SAD

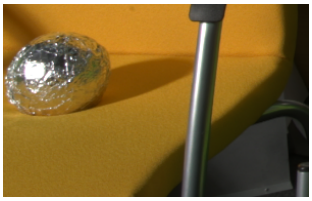
B: HY

C: GT

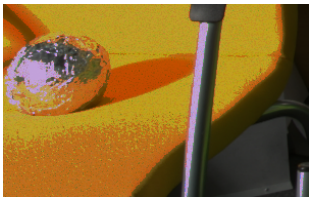
C: EO

C: SAD

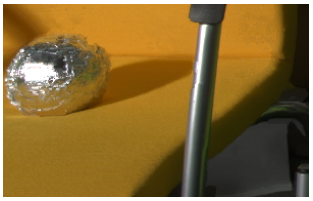
C: HY



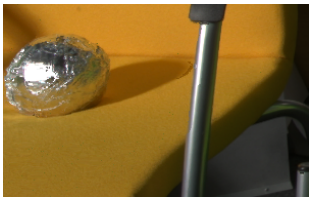
D: GT



D: EO



D: SAD



D: HY

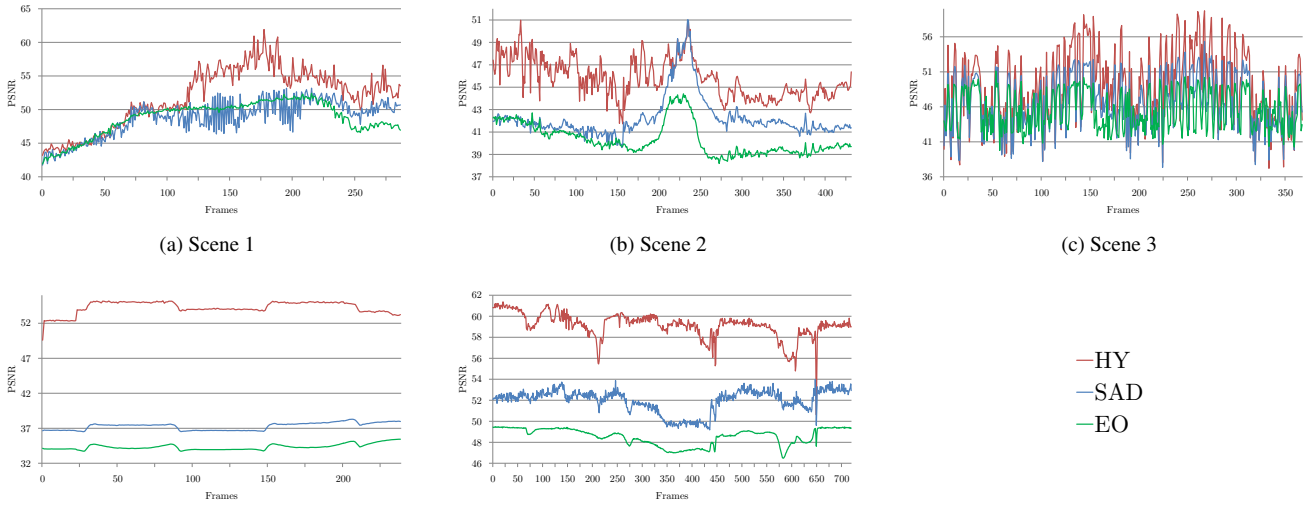
E: GT

E: EO

E: SAD

E: HY

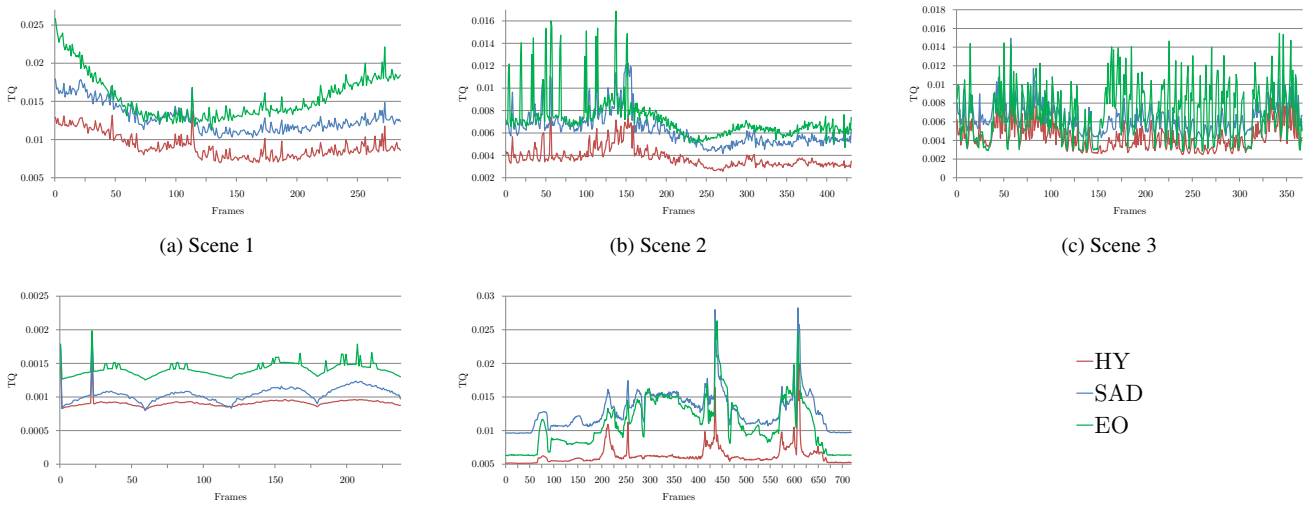
Figure 11: Detailed insets for the reconstructed SHDR frame chosen from Scene 1 showing GT, EO, SAD and HY. All images are shown at the appropriate single exposure.



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Legend

Figure 12: PSNR results for all the scenes an all the frames. Higher is better



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Legend

Figure 13: Temporal quality (TQ) results for all the scenes an all the frames. Lower is better.