Expanding dynamic range in a single-shot image through a sparse grid of low exposure pixels

Leon Eisemann; Stuttgart Media University (HdM); Stuttgart, Germany
Jan Froehlich; Stuttgart Media University (HdM); Stuttgart, Germany
Axel Hartz; Stuttgart Media University (HdM); Stuttgart, Germany
Johannes Maucher; Stuttgart Media University (HdM); Stuttgart, Germany

Abstract

Camera sensors are physically restricted in the amount of luminance which can be captured at once. To achieve a higher dynamic range, multiple exposures are typically combined. This method comes with several disadvantages, like temporal or alignment aliasing. Hence, we propose a method to preserve high luminance information in a single-shot image. By introducing a grid of highlight preserving pixels, which equals 1% of the total amount of pixels, we are able to sustain information directly in-camera for later processing. To provide evidence, that this number of pixels is enough for gaining additional dynamic range, we use a U-Net for reconstruction. For training, we make use of the HDR+ dataset, which we augment to simulate our proposed grid. We demonstrate that our approach can preserve high luminance information, which can be used for a visually convincing reconstruction, close to the ground truth.

Introduction

Current camera sensors are physically restricted in the amount of photons, which can be captured by the full well capacity of the individual sensor elements. This results in an overall limited luminance range for the whole sensor, which is usually controlled by aperture and electronic gain. Especially in the field of HDR creation this limitation is problematic, as the dynamic range of most sensors is smaller than the dynamic range of the scene to capture. To overcome this limitation several solutions were introduced. Firstly HDR-Burst, where a certain amount of images with different increasing exposures are shot in a short timespan to reduce the amount of photons per exposure and hence gain a high virtual full well capacity for the final exposure, which is one of the most commonly used methods [1]. These images are then aligned and processed to combine the contained information [1]. Similarly methods include RED Digital Cinemas HDRx, which combines one normal and one short exposure frame to gain additional information in the highlights [2] and DualISO, where the dynamic range is extended through different AD converter or denoising algorithms to get a low-noise signal to preserve high- and lowlight information [3]. Another approach is to combine two camera sensors, as proposed by Tocci et al. [4] and applied by Froehlich et al. [5] who used a stacked camera system, one for low- and one for high luminance, to preserve a high dynamic range, which is then later combined in post-production. Although these techniques show impressive results, they come with several disadvantages, like temporal aliasing, the need to align the individual images/video and longer processing times or enormous costs and camera system size increase, in case of Froehlich et al. [5]. On the other side of the spectrum are technologies, for example the work of Hasinoff et al. [6] which propose a Hybrid Dynamic Range Autoencoder for predicting HDR Images based on LDR input, or the work of Banterle et al. [7], which used inverse ToneMapping to achieve the same goal.

Methods

To solve some of the mentioned disadvantages, we propose a different method with the aim to preserve more information in high luminance image areas for later processing and enable sensors to gain a higher dynamic range. Similar to other approaches we also adapt different exposures/sensor sensitivity to preserve high luminance information. In contrast to related techniques we propose a system, that does not concatenate the sensor temporarily or physically for capturing more information. Instead our method is based on an \( k \times k \) grid of pixels, as shown in figure 1, which are altered on a sensor to be less light sensitive and as a result highlight preserving. As we also tried a 45° rotation of the grid, it was therefore calculated as

\[
g_{\text{rot}} = k_{\text{rot}} \times k_{\text{rot}} + (1.5 \ast k_{\text{rot}}) \times (1.5 \ast k_{\text{rot}})
\]

with \( k_{\text{rot}} = k \ast \sqrt{2} \) and \( k_{\text{rot}}, k \in \mathbb{N} \) (1).

Due to technical limitations \( k_{\text{rot}} \) was rounded to \( \mathbb{N} \). With \( k = 10 \), which is mainly used in this paper, only every 100th pixel is modified, which equals 1% of the total amount of pixels. This value for \( k \) was chosen as it was found in internal studies at HdM, that images with an amount of 1% replaced dead pix-
Additionally, U-Nets allow us to offer an automated procedure to perform feature engineering while training the model automatically has to perform a feature engineering while training the decoder part is trained to reconstruct again the missing information back. As U-Nets have shown impressive full well capacity, we attempt to gain additional dynamic range. One of the key challenges for a learning based highlight reconstruction with a decent resolution is to gather a large enough dataset. Especially in our case, the resolution was a keypoint to construction with a decent resolution is to gather a large enough dataset, which contains a sufficient amount of information in high luminance regions, the HDR+ Dataset by Hasinoff et al. was chosen. The dataset consists of 3640 images total, which were created through exposure stacking and were then tonemapped into sRGB domain. As the number of images is not sufficient enough for training a good generalizing network, the data was additionally augmented. Therefore the images were transformed and flipped, with reflection of the content at the image border to prevent blank spaces. As it was important to keep a realistic imaging, no color-shift or noise were applied in the augmentation. The simulation of LE-Pixels was then applied onto the augmented data so the grid keeps a static position while the scenery changes, as it would be in a real world camera.

For testing this hypothesis and to provide empirical evidence, that a small amount of low exposure pixels can boost the performance of highlight reconstruction and thus extend the dynamic range of sensors, we augment images to simulate a small subset of information from LE-pixels solely purpose is to gain additional dynamic range. Moreover this method can be easily applied to sensors, as there is no significant change to the pixel design necessary. The required grid structure could be applied by adding a new filter layer, by changing the shape of the micro-lenses or by covering parts of the pixel. This allows manufacturers to deploy this method in a fast and easy way and deliver the reconstruction logic later on, as algorithmic approach for their cameras over software updates or in the proposed way as SDK addition.

For testing this hypothesis and to provide empirical evidence, that a small amount of low exposure pixels can boost the performance of highlight reconstruction and thus extend the dynamic range of sensors, we augment images to simulate a small subset of information from LE-pixels solely purpose is to gain additional dynamic range. Moreover this method can be easily applied to sensors, as there is no significant change to the pixel design necessary. The required grid structure could be applied by adding a new filter layer, by changing the shape of the micro-lenses or by covering parts of the pixel. This allows manufacturers to deploy this method in a fast and easy way and deliver the reconstruction logic later on, as algorithmic approach for their cameras over software updates or in the proposed way as SDK addition.

Figure 2. Comparison of the ground truth in figure (a) against the augmented network input, with our proposed low exposure pixel grid, rotated 45° (b). In (b) all values above 0.5 therefore derive from our low exposure grid. For our proof of concept, the images were saved as compressed jpegs and then reloaded at a later point to calculate the histograms.

Figure 3. Magnification of reconstructions without (b) and with (c) skip-connections. The autoencoder is able to reconstruct highlevel information, but without the skip-connections detail information in higher frequency regions is missing and prediction is blurred.
For creation of the artificial dataset, all images were processed in sRGB Domain with the following function:

\[
\text{img}^{\text{dualiso}} = \text{mask} \times \text{img}^{\text{org}} + (1 - \text{mask}) \times \text{img}^{\text{clip}} \\
\text{with} \quad \text{img}^{\text{clip}} = \min(\beta \times \text{img}^{\text{org}}) \quad \text{and} \quad \beta = 0.5
\]

Consequently the non LE-pixels \(\text{img}^{\text{clip}}\) where clipped at half their electronic values (EV), as shown in figure 2. We are aware of the fact that pixels under different exposures show different noise levels, as it is dependent on the signal level 1. Since this is not our focus in this proof of concept work and autoencoders have shown impressive result in denoising challenges, we therefore do not apply any additional noise 7, 9.

**Network Structure**

As our focus was to provide a universal approach of recovering the preserved information, we propose a generic U-Net for this process. Referring to Eilertsen et al. 7, our design does not make use of a fully connected layer for the latent representation and instead uses a multichannel low resolution representation of the input data. Therefore this fully convolutional network (FCN) approach is resolution independent, as long as the input dimension is a multiple of the encoder downscaling factor 9. Since the latent representation of the network is defined as \(\frac{\text{width}^{\text{input}}}{16} \times \frac{\text{height}^{\text{input}}}{16} \times 512\) the resolution must be a multiple of 16. The down-conversion to the latent representation inside an U-Net means that high resolution information is lost and not usable in the decoder, wherefore predictions are lacking them 2. To overcome this limitations skip-connections are used in U-Nets, which transfer information from the encoder into the decoder directly 10. Our particular network structure adds skip-connections between layers with the same spatial resolution in both the encoder and decoder 11. To achieve this, the output of the encoder layer is concatenated along the feature axis of the decoder layer 10. For a given layer in the encoder with \(W \times H \times K\) the resulting decoder layer has the shape \(W \times H \times 2K\). These additional feature maps are then reduced by the decoder in the next convolution step 10, 11. An example of the impact of the introduced skip-connections is displayed in figure 3. Adding the information transfer enables the network for a better reconstruction of high frequency information. Our final structure, as shown in figure 4, is mostly inspired by the work of Mansar 11 and Ronneberger et al. 10 as this showed good performance in image denoising and high flexibility between use cases 11, 10. Our ambition is to provide evidence and give a universal approach, thus we do not introduce special domain transfer or use case specific changes to the final network in contrast to Eilertsen et al. 7 or Park et al. 12. Furthermore, to present a convenient and replicable system, we used the keras functional API with tensorflow as back-end.

The first results of the network contained excessive artefacts from the LE-Grid, especially in the highlights of the image, as seen in figure 5. This originates from a combination of mean squared error (MSE) and max pooling, as the LE-Pixels inherit the highest values, causing them to overweight in the max pooling layers. In addition, the LE-Pixels inherit the unchanged original values and as a consequence have no contribution to the MSE. To overcome this problem we temporarily introduced a specialized linear interpolation layer to mask the LE-Pixels in the network output. As this did not produce the desired results, as can be seen in the following section and figure 6, this layer has been discarded and is no longer utilized in the final notebook.

**Loss Function**

Despite the main goal of restoring information in high luminance areas we decided against a cost function that is formulated in linear quantization domain. One of the reasons was that the available training data contained huge variation above our defined clipping point, which in linear domain would have led to an unsteady cost estimation 9. Additionally it would result

---

1 The network is back-end independent, except the loss function which depends on tensorflow for SSIM.
in an underestimation in lower range luminance [9]. To provide a general cost function for reconstruction we initially used MSE between the network output and the unaltered original image. In the process of experimentation it became apparent that the network predicted inaccurate color temperatures in the highlights of the images, while using MSE as loss function. Especially in skies the network resulted in a warmer color than the original image. In addition even after intensive experimentation with training parameters, the network still had visible problems with the grid structure, which was still shown in the output. Despite introducing a linear interpolation layer, the grid was noticeable in homogeneous areas and the reconstruction of image structures was lagging behind. To put more weight on visual perception as well as stabilizing color reconstruction, the loss function was altered to a combination of multi-scale structural similarity (MS-SSIM) and mean absolute error (MAE/L1) as recommended by Zhao et al. [13].

The new loss $L_{\text{mix}}$ was then calculated with

$$L_{\text{mix}} = \alpha \times (1 - L_{\text{MS-SSIM}}) + (1 - \alpha) \times L_{\text{L1}}$$

with $\alpha = 0.84$ as proposed by the original paper [13]. As shown in figure 5, the changed loss function is capable of removing the grid artefacts in the output and stabilizing the reconstruction of color temperatures.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Input</th>
<th>Baseline rec. w/o LE-Pixels</th>
<th>LE with LE-Pixels</th>
<th>LE 45° rotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>19.3359</td>
<td>31.0810</td>
<td>31.2140</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.3816</td>
<td>3.4907</td>
<td>3.6427</td>
<td></td>
</tr>
<tr>
<td>SSIM</td>
<td>0.8968</td>
<td>0.9488</td>
<td>0.9541</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0639</td>
<td>0.0315</td>
<td>0.0304</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.00159</td>
<td>0.0026</td>
<td>0.00010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0019</td>
<td>0.0018</td>
<td>0.0007</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Comparison of different metrics evaluated over the validation set. Metrics where calculated on per image basis against ground truth, which where then averaged.

### Training

Training of the networks was performed with the ADAM optimizer, with a learning rate of 1e-4. The networks where trained for 10 epoches using a mini-batchsize of 2 due to limitations in processing power. Training with this settings takes around 6 hours on a NVIDIA GTX 1080. For a more flexible training process the keras built-in, ReduceLROnPlateau was used for adaptive learning-rate reduction when necessary. The reconstruction interference time is about 180ms, which makes the U-Net approach currently to slow for real-time applications. As mentioned ear-

Figure 6. Comparison of some of the methods mentioned in this paper, except for baseline (b) all methods could use the additional information to restore huge amount of the structure in the clipped regions.

Figure 7. Zoom-ins of reconstructions of complex light situation. As the baseline (b) is not able to estimate the structure, as well as the luminance. Our method (c) in contrast can not only estimate the correct brightness of the highlights, but also reconstruct the structures close to the ground truth (d).
Preserving and reconstructing highlights is an important and challenging task in the development of better imagery. To give another method to preserve high luminance information in camera, we present a \( k \times k \) grid of highlight preserving pixels to sustain information directly in camera for later processing. To provide evidence that a small number of pixels is enough to gain additional dynamic range, we use a fully convolutional autoencoder for reconstruction, as one possibility of a fully automated process. The functionality, quality and drawbacks of the method are demonstrated through a number of examples.

**References**


IS&T International Symposium on
Electronic Imaging
SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!

JOIN US AT THE NEXT EI!

- SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
- INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org