RAW Image Files: The Way To HDR Images From A Single Exposure

Massimo Fierro, Tae-Hyoung Lee, Yeong-Ho Ha
School of Electrical Engineering and Computer Science
Kyungpook National University
1370 Sankyuk-Dong, Buk-Gu, Taegu
702-701, Republic of Korea.
Telephone: +82 (0)53-950-5535, Fax: +82 (0)53-957-1194
fierro, katugi, yha @ee.knu.ac.kr

Abstract

HDR image formation and display has been an argument of extreme interest even when digital cameras were not yet consumer products. While recent research in both fields has seen very interesting works, none is really revolutionary, since what goes on behind the scene has been left basically unchanged. In the image formation field in particular, a lot of energy has been spent so to solve the problems that arise when taking multiple exposures: illumination change, camera shake and in-scene movement. In this paper we approach HDR image formation from a different perspective, which tries to solve in one move all the mentioned problems. More specifically, we propose a method that is able to estimate missing exposures for HDR image formation starting from only one under-exposed shot. Estimation is done through artificial neural networks: the development of a mathematical model is a highly desirable, but time consuming task. The results are are very interesting, although not perfect, and suggest that further research might lead to a suitable solution.

Introduction

Since the beginning of the era of high end digital cameras, camera producers strived to give photographers the best working environment possible: they provide advanced software solutions for image manipulation, and, most important, they also allow access to the so called RAW files. Such files, which do not yet have a common format for every producer (although a standard has been proposed by Adobe), contain relatively unprocessed data, possibly taken right after the Analog-to-Digital (AD) conversion process. Consequently, RAW files contain values that reflect the Color Filter Array (CFA) format chosen by the manufacturer, and they also have the same representation precision granted by the AD Converter (ADC). For example, the ADCs in the Canon 10D used in this work have a precision of 12 bits.

These characteristics of RAW files have always been deemed of high importance by photographers, since they allow to perform certain tasks that are impossible with 8-bit, color compressed images. One of such tasks inspired the present work: the ability to change the exposure in order to recover detail present in either the highlight or the shadow regions of the image, which are usually clipped by the tone compression curves of the camera.

Also, in this work we are going to challenge the following sentence, found in the introduction of the High Dynamic Range (HDR) imaging reference text at the time of writing [10]: "... 10 to 12 bits of linear data affords about the same precision as an 8-bit gamma-compressed format, and may therefore still be considered LDR". While this may be true when only considering the numerical precision of the representation, 8-bit Low Dynamic Range (LDR) images produced by image acquisition devices also undergo the process known as Tone Compression, usually performed by means of S-shaped functions. This means that RAW data from a camera is compressed twice, hence it is evident how a LDR image does not have the same expressive power as RAW data (and the reason why RAW files are so valued by photographers)[1]. The importance of this consideration is going to grow as in camera ADCs grow in precision (14-bit and over), and sensors gain wider and wider dynamic ranges..

In the present work, then, we are going to demonstrate how the higher expressive power granted by RAW image files can be exploited in the field of HDR imaging. More specifically we will show how, starting from a single exposure, we can obtain a reasonable approximation of a series of three exposures taken at with a bracketing interval of 2 stops, which can in turn be used to generate HDR content. This work is meant to be a proof of concept to be used as a base for further research.

As far as our knowledge is concerned, our approach is completely different from anything present in literature. Even recent developments on HDR video sensors [9], still need to two different exposures to obtain the final image. An added benefit of our proposal is that we do not need to worry about image registration and ghosting [6], which arise from differences in the position of the camera and/or subjects when taking multiple exposures (although the former can be exploited to perform super-resolution as in [5]).

The paper is organized as follows. We will start by describing digital camera sensors in general, then analyze in detail the sensor of the Canon 10D, the camera that was used for the tests. Next we explain the details of the proposed method for HDR image contents estimation. The subsequent section gives a description of the experiments along with the results, and following we will propose a possible enhancement to the method. The last section is dedicated to the conclusions, and it sums up this work with concluding remarks and prospective work.

If not otherwise specified when naming an image in this work, we refer to a RAW image, i.e. a matrix of monochromatic values representing the CFA components.

Camera sensor analysis

On the sole premise of theory, it is a common thought that, given a well calibrated silicon based sensor (CCD or CMOS), it is supposed to have a linear input/output response [7]. In other words, if we increase the input intensity by two times, so should the output (within the limits of the sensor itself).

Our first task is to verify such belief: if the assumption

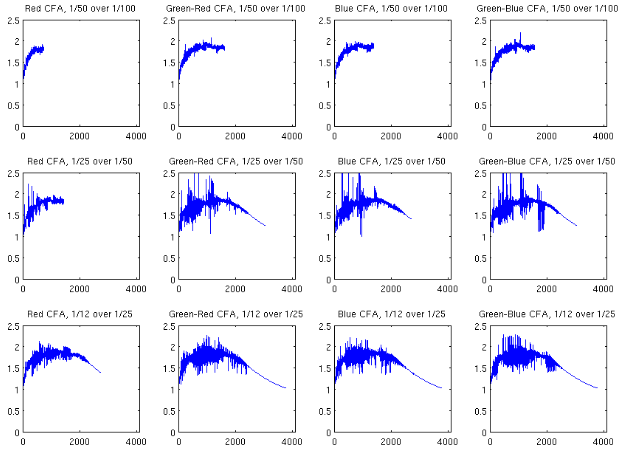


Figure 1. Each row shows the Red, Green-Red, Blue and Green-Blue CFA components for a couple of exposures 1 stop apart. The x-axis represents the input intensity and the y-axis the average ratio for that level.

is verified, the ratio between two exposures of the same scene, taken with a 1 EV latitude difference (double the amount of incoming light), should be constant for most of the output range.

The tests were carried out with a Canon 10D DSLR (6 MPixel, CMOS sensor). We would like to stress that changing the sensor or camera model would require to perform again all of the test in this paper, not to mention network training. The different performance could also lead to different considerations regarding the best starting exposure for the estimation process.

In Fig. 1 we are showing, for each of the Bayer CFA channels, the average ratio between couples of exposures from one of our test scenes, each 1 stop apart. ISO and aperture were kept fixed, while we manually increased the exposure time to obtain the desired latitude difference.

There are some interesting facts that can be distinguished by observing the plot

1. Even using RAW files, increasing the exposure by 1 stop does not produce (on average) a doubling of the output intensity values;
2. It seems that all the channels have approximately the same response;
3. After a certain threshold, the ratio becomes dependent on the observed variable only.

We will further examine the third point in order to understand how it could influence decisions regarding our goal, which is to obtain an HDR image from a single exposure.

Low-Medium signal intensity

Any kind of interference on a signal is usually regarded as unwanted noise by most engineers. When working with image acquisition devices, however, part of the "noise" is built in the system itself: the color sensitivity curves for the different components of the CFA are usually overlapping to some extent.

Furthermore, a phenomenon known as channel cross-talk increases the relationship between adjacent sensing elements. Channel cross-talk is due to stray rays of light, and, although it can be greatly reduced by careful lens design and application of micro-lenses at each of the photosites, it cannot be completely eliminated.

Both of the aforementioned facts explain the high instability of the average ratio with the respect to the channel output level: each photosite output does not depend on the observed signal only. Since such dependency between CFA channels has been effectively exploited to perform demosaicking, it is our opinion that it might be useful for our objective as well.

High signal intensity

We have already stated how, from a certain intensity level on, the ratio becomes free of any kind of noise. This is probably due to the sensors reaching its limits near the saturation region, and the threshold seems to be similar for all the CFA components. Hence, there is really no point in trying to estimate a datum which shows no variability across either CFA components, image coordinates or exposure, for it is not dependent on scene contents anymore: it's easy, but most likely meaningless. This is also the reason that lead us to start the estimation process from fast exposures, although these are more affected by photon shot noise.

Proposed method

Conventional methods for the generation of HDR images require access to different exposures of the same scene. Usually the different exposures are obtained in standard ways, such as a DSLR camera shooting in rapid succession, or a specially devised sensor. Independently of the acquisition method, though, different exposures my show a high variability due to illumination changes, camera shake or movement, and movement in the scene. In order overcome the insurgence of such differences, we propose an approximation method based on neural networks, instead of taking repeated shots with different camera settings. This may not be the ideal solution, but we will show that the error levels are quite contained even when estimating a 4 stops difference.

Our method is illustrated in Fig. 3 and can be summarized as

1. Take the first shot: $I^{\mathrm{AE}-2}$.
2. Estimate I^{AE}.
3. Estimate $I^{\mathrm{AE}+2}$ from the estimate of I^{AE}.

where we indicate with AE the camera measured exposure, and with I^x the image corresponding the the input scene at exposure x. Also a rough algorithmic description for steps 2 and 3 is given in Alg. 1.

Our systems requires 8 distinct neural networks: one for each (CFA channel, exposure difference) couple.

Each network takes as inputs 4 values and produces 1 output value. The inputs are the intensity value for the pixel at position (i, j) in the CFA, the two horizontally adjacent values, and the mean of the CFA channel of which the pixel $I^x(i, j)$ is part. The output is the estimated value of $I^{x+2}(i, j)$. The reason why we pass as input three row-adjacent pixels to the network is that most imaging sensors use the area available to photosites and read-out electronics in such a way that it dictates a higher cross-talk effect in the horizontal direction [11].

In order to establish the network structure we proceeded using cascade training to obtain an estimate of the structure itself. The final network structure is shown in Fig. 4. The activation function for all the layers except the output neuron is the logistic function, and the input is normalized to the $[0, 1]$ range according to the maximum output of the ADC of the camera (4096 for a 12-bit ADC). The output neuron activation function is the linear function.

We trained the 8 neural networks necessary to recursively estimate the full 4 stop difference from the initial shot: the training database (Fig. 6) is made of three indoor shots taken while

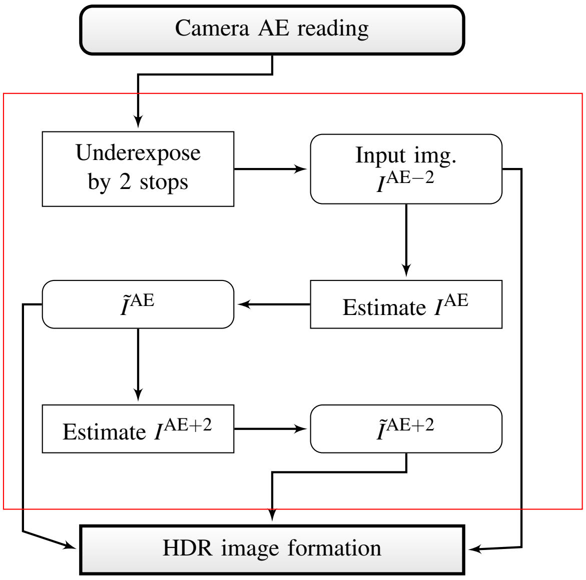

Camera AE reading
Underexpose by 2 stops
Input img. $I^{\mathrm{AE}-2}$
Estimate I^{AE}
$\tilde{I}^{\mathrm{AE}}$
Estimate $I^{\mathrm{AE}+2}$
$\tilde{I}^{\mathrm{AE}+2}$
HDR image formation

Figure 2. Flowchart of the proposed method. The part of the graph enclosed in the red line is unique to our approach.

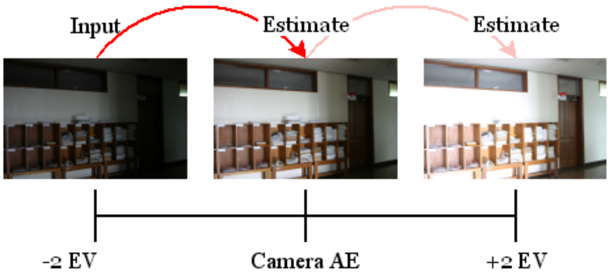


Figure 3. Intuitive view of the proposed method. $I^{\mathrm{AE}-2}$ is the input image. Images are shown in color for ease of visualization, but in the process they contain RAW data.

facing a window, a typical HDR situation, while trying to obtain a sufficient amount of color variety between the shots. We then chose 350k pixels at random from the training set and obtained convergence at $MSE < 0.0001$.

While the database is definitely undersized for extensive training, we point out again that this work shall only serve as a launchpad for future research.

Trying to establish the limits of our work, we also decided to train a set of neural networks for each of the following conditions: RAW camera output with white balance and channel stretching, and RAW output with white balance, channel stretching and denoising. In this case the normalization factor becomes 2^{16}, i.e. the maximum integer representable in a 16-bit TIFF file (the format of choice).

The denoising method we used is that described in [4], while the demosaicing algorithm we implemented is the one of [8]. The white balance algorithm is implemented in the open source utility *dcraw* [2], which features black level estimation and subtraction, and camera white-balance settings readout.

Experiments

After the network training phase, we performed tests on a two set of exposures: one indoor and one outdoor. The test images are shown in Fig. 7 and the results are shown graphically in Fig. 8.

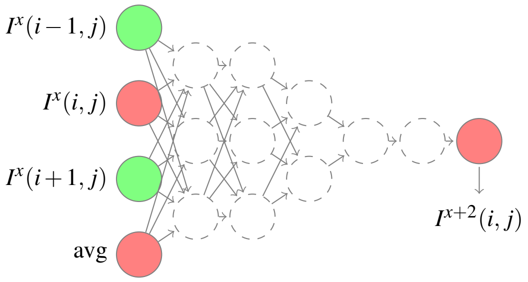


Figure 4. Artificial Neural Network structure. The coloring scheme simulates the application to a Red pixel (hence the green neighbors).

It appears evident how, on the first test image (indoor), the neural networks do not struggle too much estimating the first 2 stop difference. However, the error accumulated in the first step leads to a much larger error in the second estimate. Yet such error is only around 5% on average.

The second test image proves a more challenging: it was shot outdoors and the networks were not trained on average intensity levels such as those present in the last shot of the series. Nonetheless the error is quite contained on the first estimate (roughly on the same levels as for the indoor image), dipping below 15% on the negative scale for the second step.

Many may point out that such an error is way too big for any kind of non academic purposes, and we totally agree, yet it is small enough to serve as an indicator. More specifically it points at the necessity of further work on the topic, and even hints at the chance of formulating a mathematical model powerful enough to supersede the use of neural networks.

In order to explore as much of the problem space as possible we have taken in consideration the output of networks trained data that was generated by applying white balance, channel stretching and denoising.The results are shown in Fig. 10. It is evident how the networks behavior changes drastically, yielding lower average errors both on the positive and negative difference scale.

Another example of the very good performance obtainable by the networks trained on the modified data is given in Fig. 9, where we can compare the results of the tone mapping operator proposed by Drago et. al [3] applied to the original sequence of RAW files and the sequence generated using the described method.

Discussion and future work

Reasoning on the average error shown in Sec. , we hereby propose a modification to the method that should, in theory, lower the error on the estimation of the latitude farthest from the real shot. A schematic of the modified method is shown in Fig. 5.

While estimating a -2 stops difference from any exposure might be too much, given the characteristics of the 10D sensor, we believe that we could safely stop down by one stop using another group of networks. The process then becomes:

1. Take the first shot: $I^{\mathrm{AE}-1}$.
2. Estimate exposures $I^{\mathrm{AE}-2}$ and I^{AE} (± 1 stop).
3. Estimate $I^{\mathrm{AE}+2}$ from the estimate of I^{AE}.

The AE estimate should be less noisy than the before, since there is a latitude difference of only one stop, and consequently the average error for the second estimate should be lower as well. On the other hand we add some noise in the darkest exposure, although it should be consistently lower than that in the first es-

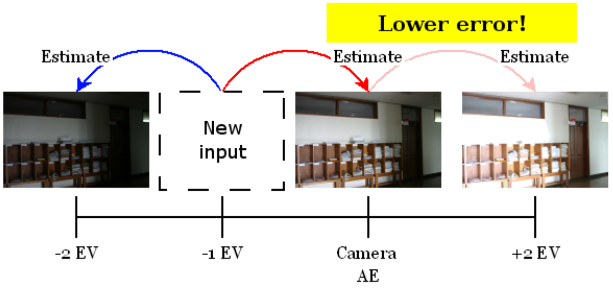


Figure 5. Proposed improvement. Starting from $I^{\mathrm{AE}-1}$, we first estimate a ± 1 stop difference. Then we estimate $I^{\mathrm{AE}+2}$ from I^{AE} as previously seen.

timate for the method originally proposed. Noise should not also be considered an excessive problem, given the excellent denoising techniques that have been developed through the years.

Experiments with the new estimation chain and an extended database are currently being performed.

Conclusions

While this may be only a first step, it is a very important one: we have hereby demonstrated how it is possible to synthesize HDR images from a single shot by estimating the missing data using a set of Artificial Neural Networks. This gets rid of all the problems related to HDR image formation (especially image registration and ghosting), although it introduces the unknown variable of the estimation itself. Since the proposed approach works on a lower level than tone mapping operators, it can be easily combined with any of the methods available. It could also be employed to produce HDR video footage offline, once the method becomes stable enough.

The current estimation errors for a generic image are still too big for real-world scenarios, as can be seen from the experiments we run. Yet, those same experiments yielded some astonishing results which push us into believe that further work is needed in this same direction, and we expect the evolution of our approach to be of practical utility.

Also, a consideration worth doing is that, if the task can be accomplished by using Neural Networks, it is highly likely that the same could be achievable by means of a “standard” framework (given a reasonably accurate mathematical model of the image formation process) which should be easier to implement on imaging hardware.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant, funded by the Korea government (MEST) (No. 20090079859), and by the second phase of the Brain Korea 21 Program in 2009.

References

[1] Digital photography review. Website.

[2] D. Coffin.

[3] F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In P. Brunet and D. Fellner, editors, *EUROGRAPHICS 2003*, volume 22, 2003.

[4] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *Image Processing, IEEE Transactions on*, 17(10):1737 –1754, oct. 2008.

[5] B. Gunturk and M. Gevrekci. High-resolution image reconstruction from multiple differently exposed images. *IEEE Signal Processing Letters*, 13(4):197–200, 2006.

[6] T. Jinno and M. Okuda. Motion blur free hdr image acquisition using multiple exposures. In *Proc. 15th IEEE International Conference on Image Processing ICIP 2008*, pages 1304–1307, 2008.

[7] W.-C. Kao, C.-C. Hsu, L.-Y. Chen, C.-C. Kao, and S.-H. Chen. Integrating image fusion and motion stabilization for capturing still images in high dynamic range scenes. *Journal on Consumer Electronics, IEEE Transactions on*, 52(3):735–741, Aug. 2006.

[8] H. Malvar, L. wei He, and R. Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. volume 3, pages iii – 485–8 vol.3, may 2004.

[9] T. Poonnen, L. Liu, K. Karia, M. Joyner, and J. Zarnowski. A cmos video sensor for high dynamic range (hdr) imaging. In *Proc. 42nd Asilomar Conference on Signals, Systems and Computers*, pages 853–856, 2008.

[10] E. Reinhard, S. Pattanaik, G. Ward, and P. Debevec. *High Dynamic Range Imaging*. The Morgan Kaufman Series in Computer Graphics and Geometric Modeling. Morgan Kaufman (Elsevier), 2nd edition, 2008.

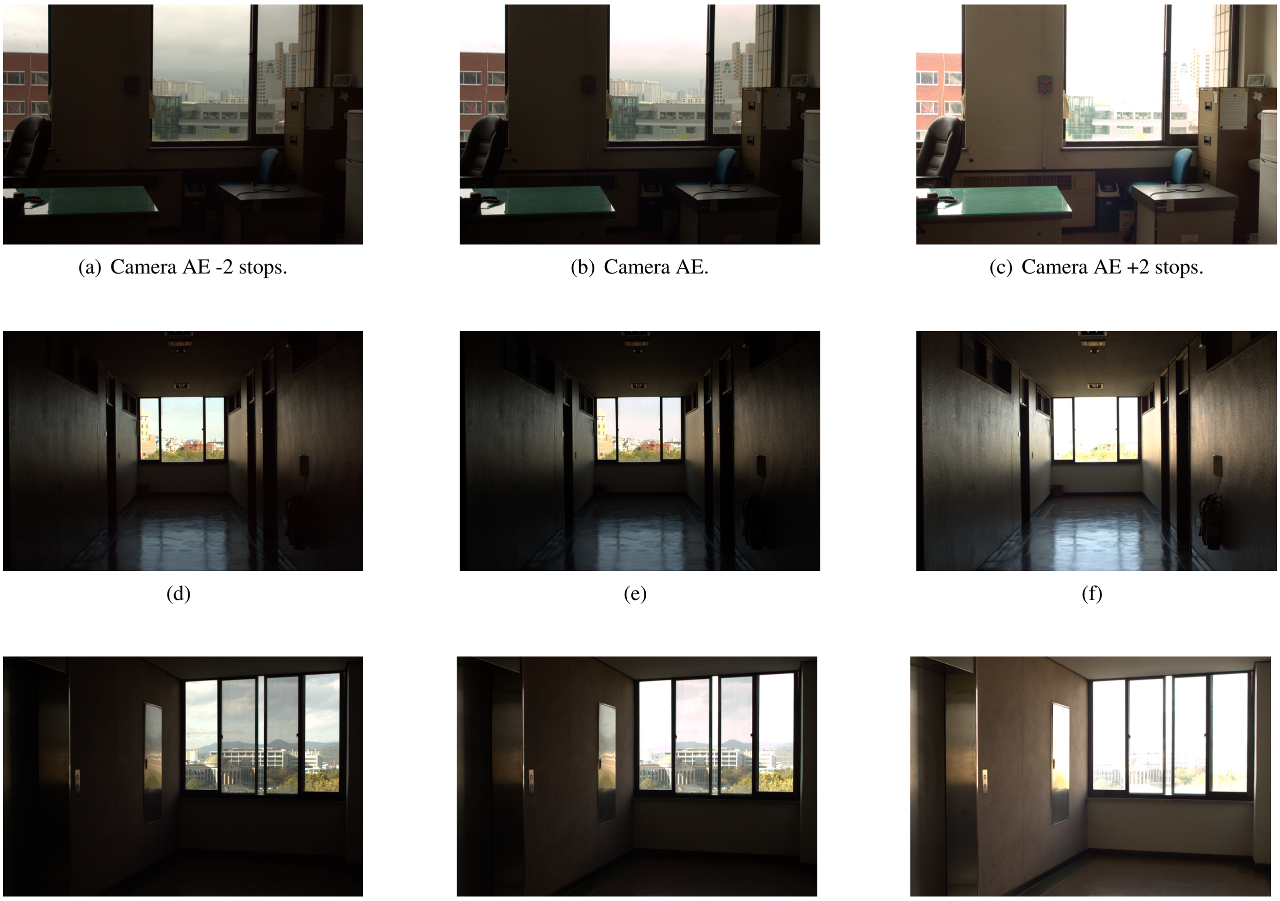
[11] H. Tian, B. Fowler, and A. E. Gamal. Analysis of temporal noise in cmos photodiode active pixel sensor. *IEEE Journal of Solid-State Circuits*, 36(1):92–101, Jan. 2001.

Author Biography

Massimo Fierro has received the Laurea in Informatica at the Dipartimento di Tecnologie dell’Informazione, Universitá di Milano. He is currently a Ph.D. student at the Kyungpook National University, and a member of CILAB under the guidance of prof. Yeong-Ho Ha.

Tae-Hyoung Lee received his BS and MS in Electronic Engineering from Kyungpook Nation University, Taegu, Korea, in 2005 and 2007, respectively. Now he is a Ph. D. candidate in Kyungpook National University. His research interests include display characterization, color management, image quality evaluation, auto exposure in camera, and high dynamic range imaging.

Yeong-Ho Ha received the B. S. and M. S. degrees in Electronic Engineering from Kyungpook National University, Taegu, Korea, in 1976 and 1978, respectively, and Ph. D. degree in Electrical and Computer Engineering from the University of Texas at Austin, Texas, 1985. In March 1986, he joined the Department of Electronics Engineering of Kyungpook National University and is currently a professor. He served as TPC chair, committee member, and organizing committee chair of many international conferences held in IEEE, SPIE, and IS&T and domestic conferences. He served as president and vice president in Korea Society for Imaging Science and Technology (KSIST), and vice president of the Institute of Electronics Engineering of Korea (IEEK). He is a senior member of IEEE and a member of Pattern Recognition Society and Society for IS&T and SPIE. Also he has been a fellow for IS&T since 2009. His main research interests are in color image processing, computer vision, and digital signal and image processing.



(a) Camera AE -2 stops.

(b) Camera AE.

(c) Camera AE +2 stops.

(d)

(e)

(f)

(g)

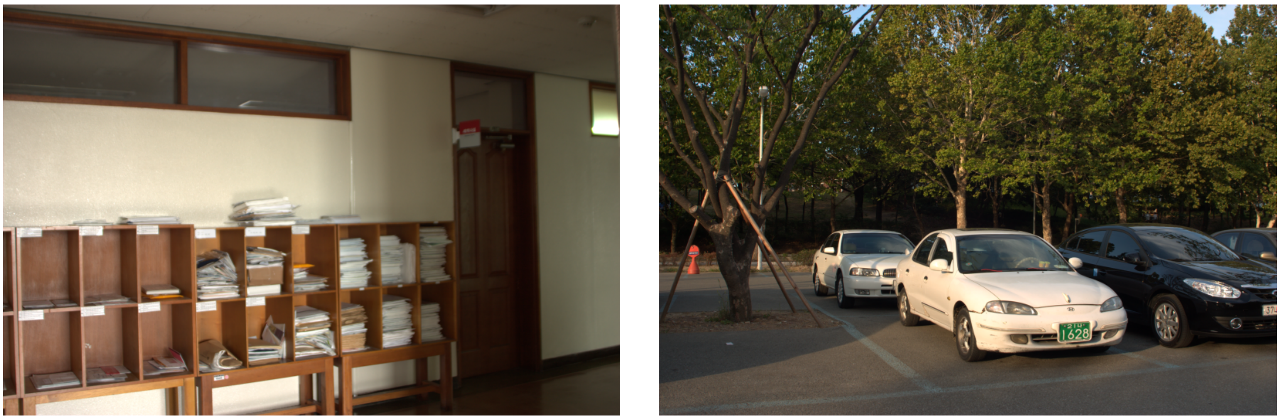
(h)

(i)

Figure 6. *Image database for ANN training. Different exposures of the same scene are shown on the same row: (a, b, c), then (d, e, f) and (g, h, i). All the frames have been taken with F16 aperture and ISO 100 sensibility.*

Algorithm 1 In the listing, cfa(i,j) is a function that outputs the CFA channel that contains at $I(i,j)$, its output is defined as $c \in \{R,GR,B,GB\}$. I_c indicates all the pixels in I that are part of the c channel, while ANN_c is the neural network trained to estimate the next exposure for the c channel.

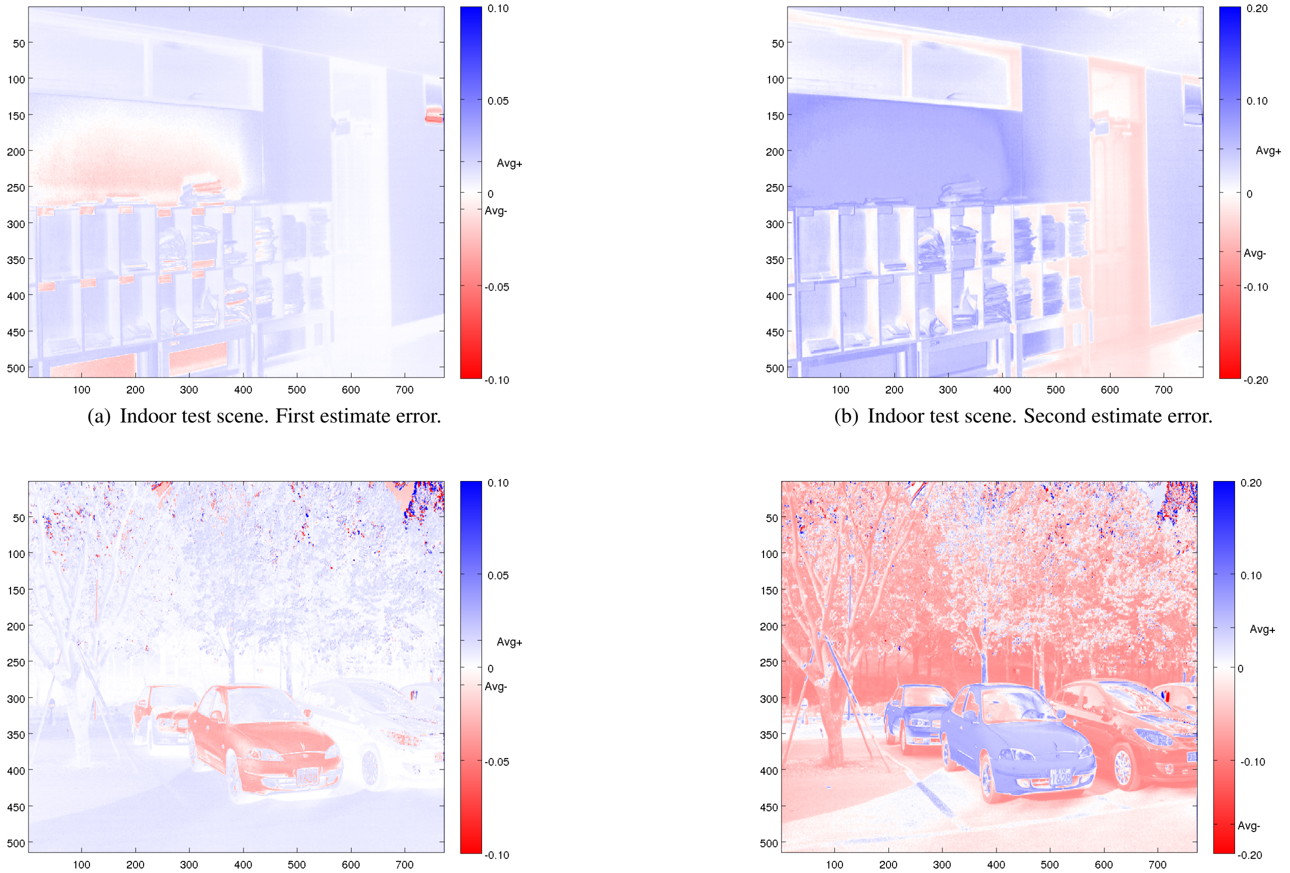
```
$w \leftarrow \mathrm{width}(I)$
$h \leftarrow \mathrm{height}(I)$
for $i = 2$ to $w-1$ do
    for $j = 2$ to $h-1$ do
        $a \leftarrow I(i-1,j)$
        $x \leftarrow I(i,j)$
        $b \leftarrow I(i+1,j)$
        $c \leftarrow \mathrm{cfa}(i,j)$
        $avg \leftarrow < I_c >$
        $O(i,j) \leftarrow \mathrm{ANN_c}(a,x,b,avg)$
    end for
end for
Replace the outer border by repeating.
```



(a) Test scene 1 (indoor).

(b) Test scene 2 (outdoor).

Figure 7. *Test scenes shown at the exposure measured by the camera.*



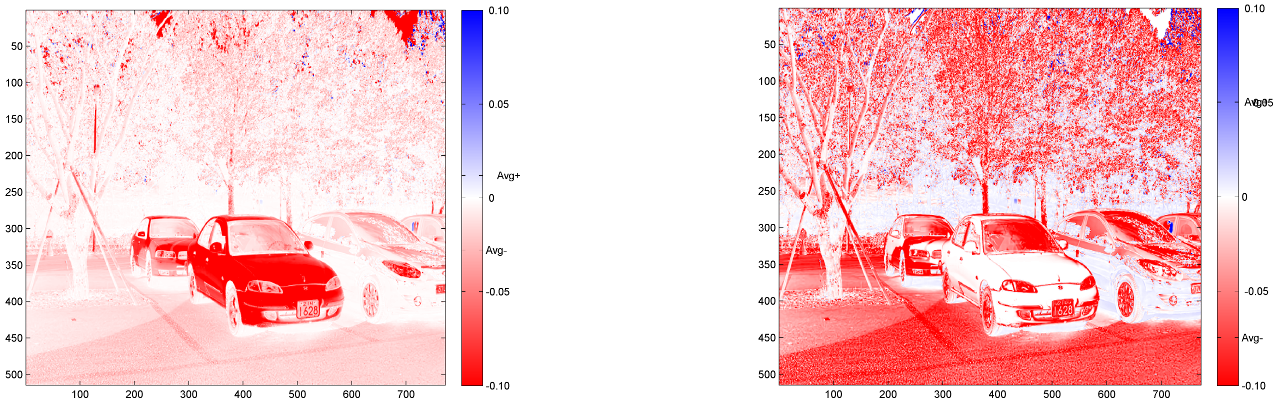
(a) Indoor test scene. First estimate error.

(b) Indoor test scene. Second estimate error.

(c) Outdoor test scene. First estimate error.

(d) Outdoor test scene. Second estimate error.

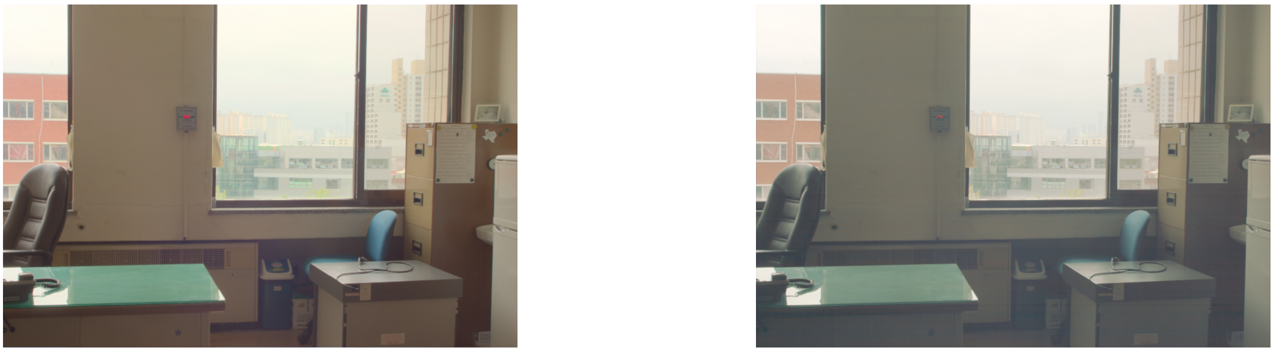
Figure 8. *False color plots of the estimation errors (scale within ±10% for the first estimate and ±20% for the second). Over-estimation is shown in blue, under-estimation in red. The average positive and negative errors are indicated with Avg+ and Avg-, respectively.*



(a) Outdoor test scene. First estimate error using white balanced images for HDR estimation.

(b) Outdoor test scene. Second estimate error using white balanced images for HDR estimation.

Figure 9. *False color plots of the estimation errors (scale within ±10%) using networks trained on white-balanced, channel stretched images. The images should be read in the same way as Fig. 10.*



(a) First database image tonemapped from RAW generated HDR.

(b) First database image tonemapped from estimated HDR.

Figure 10. *Example of tone mapping from the original RAW files compared to the tone mapping produced by one original exposure and two estimates.*