22^{nd} *Computer Vision Winter Workshop*
Nicole M. Artner, Ines Janusch, Walter G. Kropatsch (eds.)
Retz, Austria, February 6–8, 2017

Towards Dense SLAM with High Dynamic Range Colors

Sergey V. Alexandrov, Johann Prankl, Michael Zillich and Markus Vincze
TU Wien
Vision4Robotics Group, ACIN
{alexandrov, prankl, zillich, vincze}@acin.tuwien.ac.at

Abstract. *Modern dense visual SLAM systems produce high-fidelity geometric models, yet the quality of their textures lags behind. In part, the problem pertains to naive handling of colors. The RGB triplets from images are averaged straight into the model, ignoring nonlinearity of the image color space, vignetting effects, and variations of exposure time between different frames.*

In this paper we propose extensions to a surfel-based dense SLAM framework that enable more faithful capture of scene appearance. We adjust the representation to increase the dynamic range of colors, radiometrically rectify images to work in linear color space, and explicitly handle saturated pixels. Differently to the prior work in HDR-aware SLAM, we advocate turning off the automatic exposure function of the camera and incorporate a custom controller in the SLAM loop. We demonstrate improvements in texture quality compared to LDR systems, and show that self-directed exposure time control yields more complete and consistent color models.

1. Introduction

Visual SLAM (Simultaneous Localization and Mapping) has been a major research topic in robotics for decades [2]. The advent of cheap RGB-D cameras fueled interest in the family of methods that perform dense reconstruction of the underlying geometry. Newcombe *et al*. [14] introduced the idea of tracking a camera against the growing surface model using direct geometric alignment of all input data. In their system the map was represented by a flat voxel grid of limited spatial resolution and size. Follow-up works removed this restriction [3, 15], added loop closure detection [19], and online global optimization of the pose graph [4]. Other map representations, such as surfel-based [8, 20] and keyframe-based [13], were explored as well.

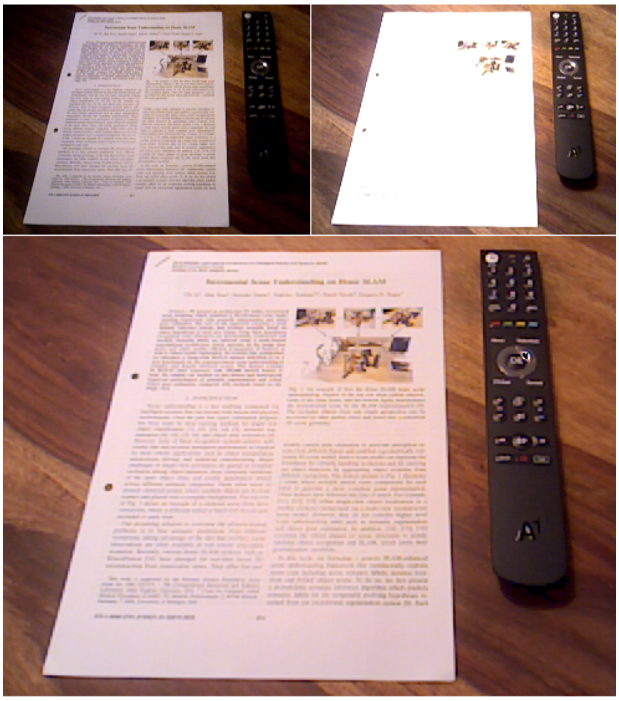
Figure 1. High Dynamic Range Imaging. The scene has light and dark objects; it is not possible to properly expose both simultaneously. Top row: camera images taken with small and large shutter times. Bottom: composite high dynamic range image (tonemapped).

Traditionally, the focus has been on obtaining globally consistent, high-fidelity geometric reconstructions. The color appearance is often neglected; dense SLAM systems either ignore input color data, or fix the sensor exposure time and adopt a simplistic method of averaging image pixel intensities.

There are several problems associated with the latter approach. Firstly, it implies that image pixel intensities directly reflect the apparent color of the scene points. This assumption does not hold due to nonlinear transformations in the process of image formation. In our recent work we demonstrated how appropriate camera calibration allows to rectify input images and resolve this issue [1]. The second major problem has to do with the limited dynamic range of the camera sensors. Only a small range of illumination intensities found in the real world can be captured and represented using conventional 24-bit RGB colors. Intensities outside this range result in under- and overexposed pixels that convey no information about the apparent color of the scene point. Thus, the details in shadows and highlights are lost.

The problem of low dynamic range (LDR) of the camera sensors is central for the photography community. Typically, it is addressed by selecting a shutter time that allows to properly expose the areas of interest in the scene. However, this fails when a scene has a lot of inherent contrast (see Figure 1). To deal with such situations the high dynamic range (HDR) imaging method was developed [18]. It involves taking multiple LDR images at different exposure times, making sure that all areas of the scene are properly exposed in at least one image. Then the images are converted into a linear color space and combined into a radiance map of extended dynamic range.

Recently, two works were presented that implement a dense SLAM pipeline and recover scene colors in HDR [11, 12]. Both rely on the automatic exposure (AE) function of the camera to obtain differently exposed images of the scene. This has two implications: exposure times have to be estimated at per-frame basis, thus incurring computational effort and drift in the long run. Further, AE is shortsighted in that it has no global awareness and is designed to optimize exposure time for the current frame. Consequently, parts of the scene may remain always saturated and thus without valid color.

In this paper we address the problem of capturing and representing the colors in a scene in high dynamic range. Equipped with a consumer-grade RGB-D camera, we strive to obtain a 3D model with accurate colors, without losing any detail in highlights and shadows. In contrast to the prior work, we disable the built-in AE function and design our own controller. The benefit is two-fold: exposure time of every image is known and does not need to be estimated. Secondly, we leverage the reconstructed model to make educated decisions regarding which exposure time should be used next to gain new color information. To the best of our knowledge, this is the first implementation of a custom exposure time controller and the first work to extend a surfel-based dense SLAM framework to handle colors in high dynamic range.

2. Preliminaries and related work

2.1. Radiometric image formation

Image formation is a complex process that involves nonlinear transformations. Scene points emit rays in the direction of the camera. The camera shutter is opened for a certain period of time to allow the light to pass through the lens system. The energy received on the image plane is then converted into an electrical signal and quantized into pixel intensities.

For a scene point $\mathbf{x} \in \mathbb{R}^3$ the intensity of the corresponding pixel $\mathbf{u}$ in the image space domain $\Omega \subset \mathbb{N}^2$ is given by:

$$I(\mathbf{u}) = f\left(tV(\mathbf{u})\mathcal{L}(\mathbf{x})\right), \qquad (1)$$

where $f : \mathbb{R} \to \{0, \ldots, 255\}$ is the radiometric response function of the camera, t is the exposure time, $V : \Omega \to \mathbb{R}$ is the optical response of the lens system (vignetting), and $\mathcal{L} : \mathbb{R}^3 \to \mathbb{R}$ defines mapping between scene points and their radiances.

The function $f(\cdot)$ maps energy received at a pixel well into a quantized 8-bit intensity value. Due to the limitations in the sensor technology, energies outside a certain valid range are mapped to either minimum (0) or maximum (255) intensity values. Such pixels are said to be under- or overexposed and provide only an upper or a lower bound of the received energy.

Vignetting response $V(\cdot)$ maps image locations to attenuation factors. It is often assumed to be radially symmetric and is modeled with a polynomial function [9]. However, recently it was demonstrated that the vignetting response in consumer RGB-D cameras is better modeled with a nonparametric per-pixel map of attenuation factors [1].

Both radiometric and vignetting responses can be precalibrated [1, 6]. Figure 2 shows recovered responses for the blue channel of an Asus Xtion Live Pro camera. The responses in other color channels are similar, but not identical.

2.2. HDR capture

The goal of HDR capture is to recover a radiance image of a scene in its full dynamic range. Based on (1), pixels in I can be converted into a radiance image L:

$$L(\mathbf{u}) = \frac{f^{-1}\left(I(\mathbf{u})\right)}{tV(\mathbf{u})}. \qquad (2)$$

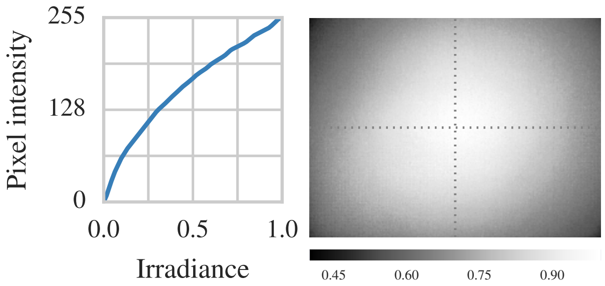

Figure 2. Radiometric calibration of an Asus Xtion Live Pro camera (blue color channel). Left: camera response function, recovered up to an unknown scale factor. Right: vignetting response as a map of pixel attenuation factors.

On its own, this conversion does not increase the dynamic range; the radiances of saturated pixels remain unknown. However, it brings the values into a linear space at absolute scale. Thus, this conversion allows to combine this image with others taken at different exposure times and having different effective range.

The basic procedure of HDR capture involves taking a set of LDR images $\{I_i\}$ at n different exposure times. These images are converted into radiance images $\{L_i\}$ and are combined using:

$$\bar{L}(\mathbf{u}) = \frac{\sum_{i=1}^{n} w(I_i(\mathbf{u})) L_i(\mathbf{u})}{\sum_{i=1}^{n} w(I_i(\mathbf{u}))}, \tag{3}$$

where w is a confidence weight that depends on the pixel measurement. In the early work of Debevec and Malic [5] a hat function was used. Later, Kirk and Andersen [10] characterized noise properties of several other weighting schemes. They concluded that the variance-based weighting gives best lower bound on signal-to-noise ratio. Hasinoff *et al.* [7] investigated the problem of selecting exposure times and gains for noise-optimal HDR capture.

Conventional HDR methods assume that each pixel $\mathbf{u}$ represents the same scene point $\mathbf{x}$ across different images. Effectively, this means that both the camera and the scene should be static. In more recent works, attempts are made to relax this requirement and allow capturing without tripod [22], or to tolerate moving objects in the scene [16].

2.3. HDR mapping

The ideas from the HDR imaging area were applied in the context of 3D reconstruction with RGB-D cameras. Zhang *et al.* [21] presented an offline method to obtain globally optimal HDR textures for a reconstructed 3D model. They formulated a nonlinear optimization problem, where per-image exposures and point radiances are the unknowns.

Motivated by augmented reality applications, particularly insertion of reflective objects with shadows into a video stream, Meilland *et al.* [12] proposed a dense SLAM system that recovers HDR colors. They represent the 3D scene model as a graph of super-resolved HDR keyframes, each of which is a result of fusion of a set of LDR images. Camera tracking is performed through direct alignment with geometric and photometric error terms. The latter includes relative exposure time, estimated jointly with camera transform. They model the camera response with the gamma function and ignore the vignetting effects.

Recently, Li *et al.* [11] described a different approach to mapping with HDR colors, where they extend a volumetric SLAM framework. In their formulation exposure compensation is decoupled from tracking. The alignment problem is cast in the normalized radiance space that is independent of exposure time. Once the new camera pose is estimated, the exposure time change is determined as a weighted average of radiance ratios between corresponding pixels. Finally, the radiance map is scaled using the estimated exposure time and is fused into the global volumetric representation.

Our system is similar to the latter two in that it performs online 3D reconstruction with HDR colors. The difference lies in that we use a surfel-based representation and control the exposure time of the camera based on the current state of the map.

3. System overview

We employ an architecture typical for real-time dense SLAM systems, where camera tracking is alternated with mapping. The flow diagram is presented in Figure 3. Input color images I from an RGB-D camera are radiometrically rectified to obtain radiance maps L. Together with the depth maps they are used to estimate the current camera pose within the map. This is done through direct alignment with the virtual radiance and depth maps predicted at the previous pose of the camera. The rectified input data is then fused into the existing model M using the estimated camera pose. Next we perform view prediction of the updated model from the estimated camera pose. Finally, the predicted saturation map S is used by the exposure time controller to select the shutter time for the next frame.

This pipeline is based on the work of Keller *et al.* [8]. The main difference and contribution of this paper consists in (a) extension of various pipeline

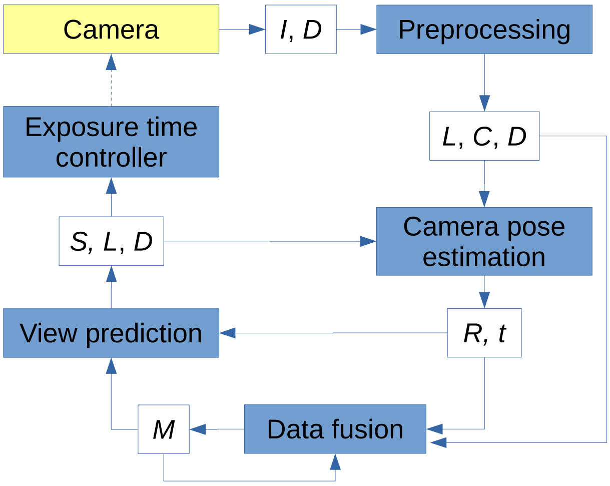

Figure 3. Flow diagram of the proposed SLAM system. Filled boxes represent pipeline processing stages, empty boxes represent data structures, and arrows represent data and control flow.

stages to deal with HDR colors and (b) introduction of an exposure time controller in the loop. The next two sections cover these topics.

4. Mapping with HDR colors

4.1. Preprocessing

The camera delivers noisy LDR images of the scene. Before fusing them into the map, we perform radiometric rectification using (2) to bring them into a linear color space at the same scale as the map. The camera response function and vignetting effects are precalibrated [1], and the exposure time t is known for every frame.

4.2. Map representation

The map is represented by a set of surfels $\mathcal{M}$. Each surfel has the following fields: position $\mathbf{p} \in \mathbb{R}^3$, normal $\mathbf{n} \in \mathbb{R}^3$, weight $w \in \mathbb{R}$, radius $r \in \mathbb{R}$, timestamp t, and HDR color $\mathbf{c} \in \mathbb{N}^4$.

The first element $\mathbf{c}_1$ of an HDR color stores a confidence value. Positive confidence indicates that the color is *valid*, *i.e.* the surfel was observed through a non-saturated pixel at least once. In this case the remaining three elements contain per-channel radiances. Zero confidence indicates that the color is *invalid*, *i.e.* up to now the surfel was always observed through under- or overexposed pixels. In this case $\mathbf{c}_2$ and $\mathbf{c}_3$ store the longest and the shortest exposure times at which the surfel was observed so far, and $\mathbf{c}_4$ is a binary flag indicating whether the observed pixels where under- or overexposed.

Each of the fields of an HDR color is stored in a 16-bit integer, requiring 64 bits in total. Note that *e.g.* ElasticFusion implementation[1] allocates 64 bits for surfel color (although only 24 bits are used). Therefore, the conversion to HDR can be achieved without changing of the memory footprint.

4.3. Camera pose estimation

Camera pose update is estimated through direct registration of the current input frame with rendered view of the model as seen from the previous camera pose. We minimize a joint cost function consisting of geometric and photometric residuals, exactly as described by Whelan *et al.* [20]. Note that in our case the input images are radiometrically rectified and are at scale with the map. Therefore, even though the formulation is the same, our registration minimizes error in the HDR space.

4.4. Data fusion

We perform projective association of the input data with the existing surfels. The computed correspondences are used to update positions, normals, weights, radii, and timestamps of the surfels following the rules described by Keller *et al.* [8]. We introduce an additional rule to handle the fusion of colors.

Consider a pixel $\mathbf{u}$ which is associated with a surfel $\mathcal{M}^s$ having color $\mathbf{c}$. We distinguish between four cases according to the validity of the colors being fused. When both surfel color and input color are invalid, we update the minimum and maximum exposure times:

$$\mathbf{c}_2 \leftarrow \min\left(\mathbf{c}_2, t\right), \quad \mathbf{c}_3 \leftarrow \max\left(\mathbf{c}_3, t\right). \tag{4}$$

When the surfel color is invalid, and input color is valid, the former is overwritten. When the surfel color is valid, and the input color is invalid, nothing is done. Finally, if both colors are valid, then a per-channel weighted average is computed.

4.5. View prediction

After every map update the view of the model $\mathcal{M}$, as seen from the current camera pose, is predicted. We implemented an OpenGL pipeline with a simple surface splatting technique, where each surfel is rendered as an opaque hexagon [17]. The fragment shader is configured to output to multiple textures: depth map $\hat{D}$, radiance map $\hat{L}$, and saturation map

[1] `https://github.com/mp3guy/ElasticFusion`

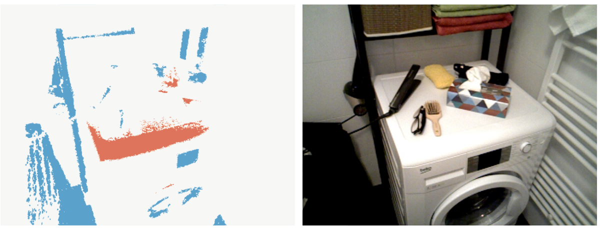

Figure 4. Left: predicted saturation map. Red pixels correspond to the surfels with invalid overexposed colors, blue pixels correspond to the surfels with invalid underexposed colors. Right: color image of the scene.

$\hat{S}$. The latter two are induced by the surfel colors according to the following rules.

For a valid color $\mathbf{c}$ the values stored in $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$ are copied into the radiance map, and the saturation is zero. Contrary, for an invalid color the radiance is set to zero, whereas the saturation is given by:

$$s(\mathbf{c}) = \begin{cases} \mathbf{c}_2, & \text{if } \mathbf{c}_4 = 0 \\ -\mathbf{c}_3, & \text{otherwise.} \end{cases} \tag{5}$$

For the surfels that have been always observed through overexposed pixels, the saturation value will be positive and equal to the shortest exposure time. For the surfels that were always underexposed, the value will be equal to the longest exposure time, with a negative sign. An example of a predicted saturation map is shown in Figure 4a. The saturation map is used to control camera exposure time as detailed in Section 5.

5. Exposure time control

Consumer RGB-D cameras often have built-in AE function. The particular algorithm is vendor-dependent, but generally it tries to adjust exposure time to strike a balance between under- and overexposing the scene. Such control is adequate to achieve as-good-as-possible exposure in a single image, however for the purposes of spatially extended 3D reconstruction it is suboptimal.

Our system analyses the current state of the reconstructed model and decides which exposure time will yield most information gain in the next frame. For this purpose we utilize the saturation map rendered in the view prediction step.

The non-zero pixels of the saturation map correspond to the model surfels that are visible from the current camera pose, but do not have a valid color. By properly adjusting exposure time we may obtain valid color information for these surfels in subsequent frames. We use the following intuition to formulate the control rules:

1. Exposure time should be varied smoothly to avoid sudden massive changes in the image appearance;
2. Saturated pixels close to the center of the image should have more influence because they are more likely to stay in the field of view in the subsequent frames;
3. Exposure controller should have hard limits on the allowed exposure times, because too large exposure times lead to high degree of motion blur, which significantly deteriorates the quality of the color models.

The consequence of the first rule is that the control decision boils down to choosing whether to increase, decrease, or keep the same exposure time. We apply a function $\omega(\cdot)$ to each pixel in the saturation map and sum up the results. The sign of the sum indicates whether to increase or decrease the exposure time.

The function $\omega(\cdot)$ is defined as follows:

$$\omega(s) = \exp\left(\frac{1}{d\left(s,t\right)}\right)\exp\left(\frac{-\gamma^2}{2\sigma^2}\right), \tag{6}$$

where $d\left(\cdot,\cdot\right)$ is the difference between saturation value and current exposure time, γ is the normalized radial distance of the pixel, and σ is a weighting factor. The first part gives more influence to pixels that require less change in exposure time. The second part gives more influence to the pixels close to the center, as stipulated by the second rule.

6. Experimental evaluation

We performed a number of small-scale reconstructions with our system and the state-of-the-art LDR system of Whelan *et al.* [20] to demonstrate the benefits of using high dynamic range colors and the custom exposure controller.

In the first experiment we recorded an RGB-D sequence with fixed exposure time. Figure 5 demonstrates a birds-eye view of LDR and HDR reconstructions. Our texture is more smooth and consistent, especially in bright (white paper) and dark (dark parts of the table) texture regions.

In the second experiment we performed two scans of an office table, one with fixed exposure time, and

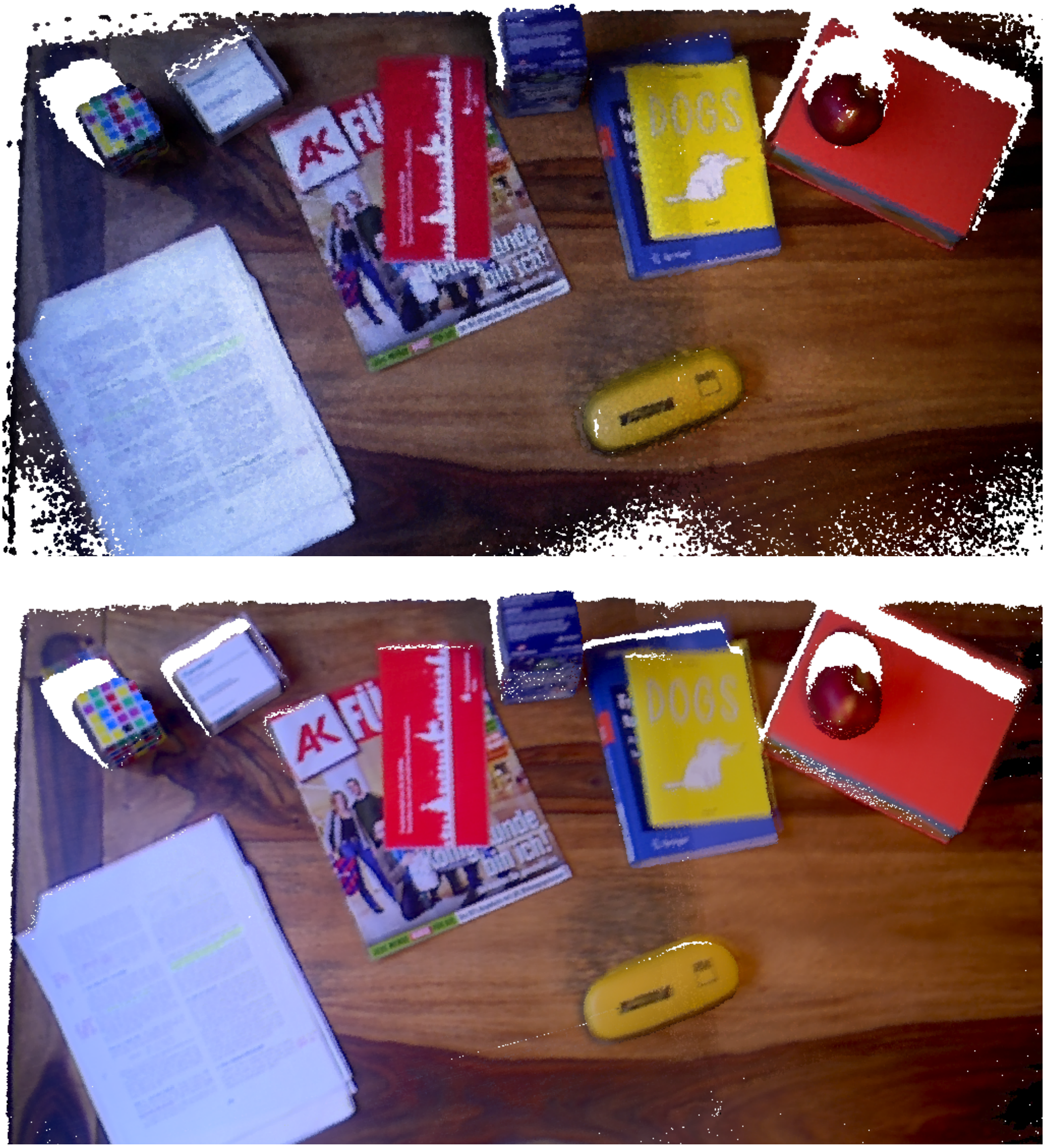

Figure 5. Top: LDR reconstruction obtained using ElasticFusion [20] with disabled camera AE and fixed exposure time. Bottom: HDR reconstruction obtained using our system from the same data sequence.

one with our exposure controller enabled. Figure 6 presents LDR reconstruction using the first sequence and HDR reconstruction using the second sequence. As before, HDR reconstruction has more smooth and consistent textures. Furthermore, the dark objects in the scene (keyboard and phone) have been properly exposed and more details are preserved.

In the third experiment we performed two scans of a washing machine, one with fixed exposure time, and one with our exposure controller enabled. Both sequences were used to produce HDR reconstructions with our system. Figure 4 presents the visual appearance of the reconstructions and the confidence values associated with the surfels. Clearly, the

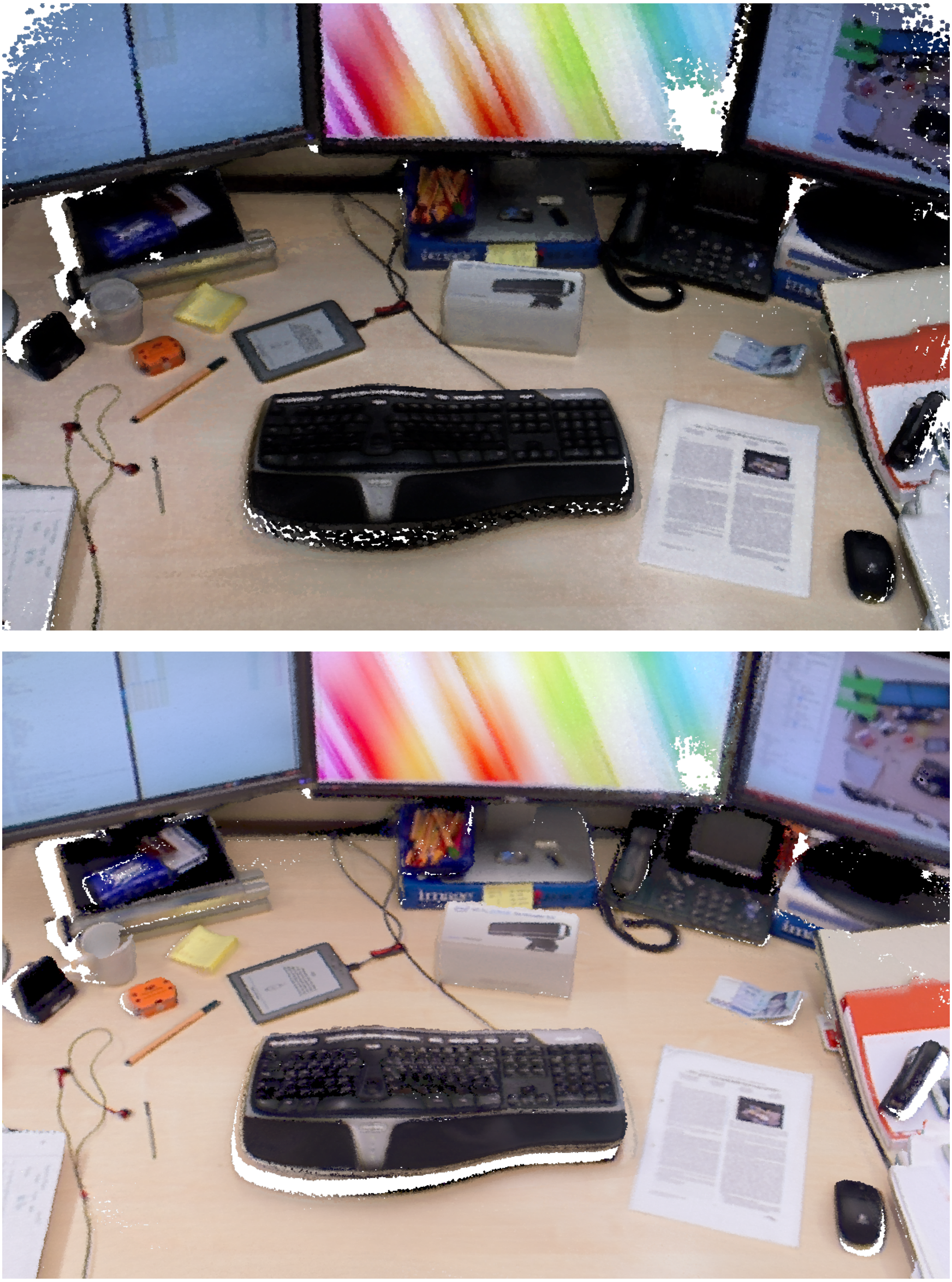

Figure 6. Top: LDR reconstruction obtained using ElasticFusion [20] with disabled camera AE and fixed exposure time. Bottom: HDR reconstruction obtained using our system with the custom exposure controller enabled.

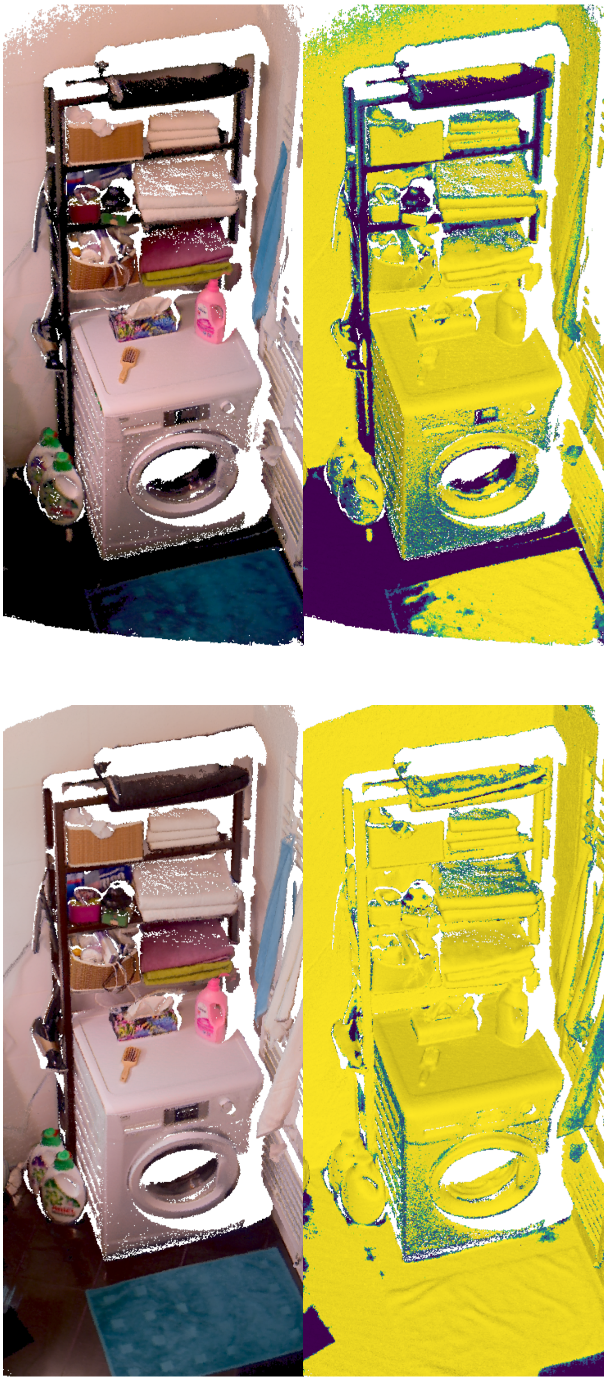

Figure 7. HDR reconstructions obtained using our system, without (top row) and with (bottom row) the custom exposure controller. Left column shows the visual appearance of the reconstruction, right column shows the color confidence of the surfels (color-coded, yellow means confident, dark violet means invalid color).

darker areas in the bottom and the shelves were not properly exposed in the sequence with fixed shutter time. Conversely, in the second sequence the exposure controller made sure that all parts of the scene were properly exposed.

7. Conclusions and future work

In this contribution we presented a study of HDR mapping with consumer RGB-D cameras. Extensions to the standard surfel-based SLAM system were described that lead to improved texture quality.

In our current implementation a number of design decisions (such as exposure control rules) were made without proper empirical evaluation. The challenge, however, is to define suitable evaluation metrics.

It is interesting to evaluate the influence of improved color model and error minimization in HDR color space on the tracking performance.

Since our method is active, it is challenging to perform a fair comparison with ”passive” HDR systems that use cameras’ built-in AE function. The same RGB-D sequence can not be used, and it is almost impossible to reproduce the same trajectory without involved robotic setup. One solution would be use a synthetic dataset, where multiple frames with different exposure times may be rendered for each camera pose in the trajectory.

Reconstruction with HDR colors is a relatively new area, and there are no benchmark datasets. Engel *et al.* [6] recently published a dataset for monocular odometry with radiometrical calibration of the camera. In a similar spirit, it may be useful to collect and publish a RGB-D dataset with such calibration.

Acknowledgments

The work presented in this paper has been funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 600623 (”STRANDS”) and No. 610532 (”SQUIRREL”).

References

[1] S. V. Alexandrov, J. Prankl, M. Zillich, and M. Vincze. Calibration and Correction of Vignetting Effects with an Application to 3D Mapping. In *Proc. of IROS*, 2016. 2, 4

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard. Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age. *arXiv preprint*, jun 2016. 1

[3] J. Chen, D. Bautembach, and S. Izadi. Scalable Real-time Volumetric Surface Reconstruction. *ACM Transactions on Graphics*, 32(4):1, 2013. 1

[4] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt. BundleFusion: Real-time Globally

Consistent 3D Reconstruction using On-the-fly Surface Re-integration. *arXiv preprint*, 2016. 1

[5] P. Debevec and J. Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *Proc. of SIGGRAPH*, pages 1–10, 1997. 3

[6] J. Engel, V. Usenko, and D. Cremers. A Photometrically Calibrated Benchmark For Monocular Visual Odometry. *arXiv preprint*, jul 2016. 2, 8

[7] S. W. Hasinoff, F. Durand, and W. T. Freeman. Noise-optimal Capture for High Dynamic Range Photography. In *Proc. of CVPR*, 2010. 3

[8] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb. Real-time 3D Reconstruction in Dynamic Scenes using Point-based Fusion. In *Proc. of 3DV*, pages 1–8, 2013. 1, 3, 4

[9] S. J. Kim and M. Pollefeys. Robust Radiometric Calibration and Vignetting Correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):562–576, 2008. 2

[10] K. Kirk and H. J. Andersen. Noise Characterization of Weighting Schemes for Combination of Multiple Exposures. In *Proc. of BMVC*, 2006. 3

[11] S. Li, A. Handa, Y. Zhang, and A. Calway. HDRFusion: HDR SLAM using a low-cost auto-exposure RGB-D sensor. *arXiv preprint*, apr 2016. 2, 3

[12] M. Meilland, C. Barat, and A. Comport. 3D High Dynamic Range Dense Visual SLAM and its Application to Real-time Object Re-lighting. In *Proc. of ISMAR*, pages 143–152, 2013. 2, 3

[13] M. Meilland and A. I. Comport. On Unifying Key-frame and Voxel-based Dense Visual SLAM at Large Scales. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3677–3683. IEEE, nov 2013. 1

[14] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time Dense Surface Mapping and Tracking. In *Proc. of ISMAR*, 2011. 1

[15] M. Niessner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale Using Voxel Hashing. *ACM Transactions on Graphics*, 32(6):169:1—-169:11, 2013. 1

[16] T. H. Oh, J. Y. Lee, Y. W. Tai, and I. S. Kweon. Robust High Dynamic Range Imaging by Rank Minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, jun 2015. 3

[17] R. F. Salas-Moreno. *Dense Semantic SLAM*. PhD thesis, 2014. 4

[18] P. Sen and C. Aguerrebere. Practical High Dynamic Range Imaging of Everyday Scenes: Photographing the World as We See It with Our Own Eyes. *IEEE Signal Processing Magazine*, sep 2016. 2

[19] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald. Real-time Large-scale Dense RGB-D SLAM with Volumetric Fusion. *The International Journal of Robotics Research*, 34(4-5):598–626, 2014. 1

[20] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison. ElasticFusion: Dense SLAM Without A Pose Graph. In *Proc. of RSS*, 2015. 1, 4, 5, 6, 7

[21] E. Zhang, M. F. Cohen, and B. Curless. Emptying, Refurnishing, and Relighting Indoor Spaces. *Proc. of SIGGRAPH Asia*, 35(6), 2016. 3

[22] H. Zimmer, A. Bruhn, and J. Weickert. Freehand HDR Imaging of Moving Scenes with Simultaneous Resolution Enhancement. In *Computer Graphics Forum*, volume 30, pages 405–414, 2011. 3