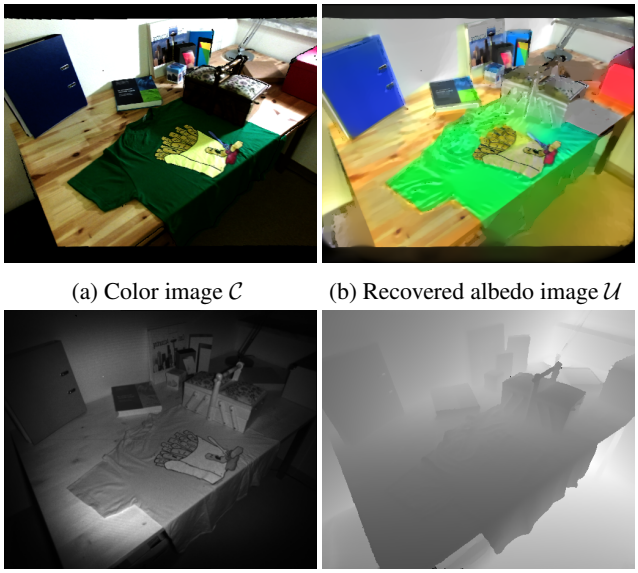
Towards Illumination-invariant 3D Reconstruction using ToF RGB-D Cameras

Christian Kerl[1], Mohamed Souiai[1], Jürgen Sturm[2], and Daniel Cremers[1]

[1]Technische Universität München, {christian.kerl,mohamed.souiai,cremers}@in.tum.de
[2]metaio GmbH, juergen.sturm@metaio.com

Abstract

*Creating textured 3D scans of indoor environments has experienced a large boost with the advent of cheap commodity depth sensors. However, the quality of the acquired 3D models is often impaired by color seams in the reconstruction due to varying illumination (*e.g.*, shadows or highlights) and object surfaces whose brightness and color vary with the viewpoint of the camera. In this paper, we propose a direct and simple method to estimate the pure albedo of the texture, which allows us to remove illumination effects from IR and color images. Our approach first computes the illumination-independent albedo in the IR domain, which we subsequently transfer to the color albedo. As shadows and highlights lead to over- and underexposed image regions with little or no color information, we apply an advanced optimization scheme to infer color information in the color albedo from neighboring image regions. We demonstrate the applicability of our approach to various real-world scenes.*



(a) Color image $\mathcal{C}$ (b) Recovered albedo image $\mathcal{U}$

(c) Infrared image $\mathcal{I}$ (d) Depth image $\mathcal{Z}$

Figure 1: Complex real world illumination introduces artifacts in the 3D reconstruction of textured indoor scenes. In this paper, we propose a method that allows to recover an illumination independent albedo texture. (a) input color image,(b) estimated color albedo image, (c) input infrared and (d) depth image from ToF camera.

1. Introduction

Novel depth sensors have recently led to many novel algorithms for the 3D reconstruction of objects, persons, and indoor environments. KinectFusion [15] demonstrated that 3D reconstruction was feasible in real-time on commodity hardware. Shortly after, the original approach was extended to allow for texture estimation [17]. However, most approaches are prone to color seams in the reconstruction due to variations in illumination. As these seams are partially the result of auto-exposure on the camera, their influence can be reduced by jointly estimating shutter and gain from the input images [12]. However, most natural scenes exhibit shadows and highlights due to non-uniform scene illumination which then also becomes part of the texture of the 3D model. For many applications, like object classification, augmented reality, and image-based remodelling it is desirable to remove these illumination effects completely from the 3D reconstruction. An illumination independent model facilitates scene re-lighting and the addition or removal of objects without interfering shadows.

When the light sources in a scene are unknown, the removal of shades is computationally highly involved as next to the 3D geometry of the room also the positions of all light sources (including indirect light) have to be estimated. Additionally, it is impossible to tell effects caused by lights and shadows, and material properties apart. Therefore, previous approaches rely on strong priors [1]. The problem can be simplified if the scene geometry is known [5, 18].

We suggest to use even more information obtained by

modern time-of-flight (ToF) cameras such as the Kinect One from Microsoft. These provide next to the depth and color image a so-called *active image*: This image only measures the light emitted by the modulated light source of the ToF camera, *i.e.*, all ambient light of other light sources is already filtered out.

In this paper, we propose an approach to infer the infrared and color albedo from data of the Microsoft Kinect One sensor. To this end, we characterize the noise of this sensor both for the intesity images and the depth measurements. In sum, the contributions of this paper are:

1. an illumination model for the IR light source of the Kinect One,
2. a noise model for the infrared image and the depth image of the Kinect One,
3. a method to estimate the IR albedo from infrared and depth images,
4. a method to register and fuse multiple image frames into a single keyframe to improve the quality of depth maps and the IR albedo,
5. a method to transfer the IR albedo to color images,
6. a qualitative evaluation on several indoor scenes.

The result of our approach is an illumination-independent estimate of the albedo of a scene, *i.e.*, where shadows and highlights from all light sources have been removed. With this work, our goal is to pave the way for illumination-free 3D reconstructions of indoor spaces.

2. Related Work

Scene illumination and shading models have been intensively studied in computer vision and computer graphics [14]. Early work by Yu *et al*. [19] shows that it is possible to accurately recover surface properties, if the scene geometry, light sources and camera positions are known. Work by Jachnik *et al*. [9] shows that it is possible to recover surface properties and ambient illumination, if the geometry is known and sufficient observations are obtained.

Another direction of research is concerned with a similar problem called intrinsic image decomposition, which tries to invert the image formation process and estimate depth, illumination, and surface albedo [1]. As this is an ill-posed problem strong priors and involved optimization algorithms are required. Subsequently, other approaches tried to simplify the problem by including scene geometry from RGB-D cameras [1, 5]. Recently, Chen *et al*. proposed in addition to model shadows explicitly [18]. Similarly, we argue to include even more information provided by ToF RGB-D cameras.

ToF cameras have been studied for a while and applied to different research problems. Foix *et al*. [7] give an overview of the measurement principle of ToF cameras and common errors observed in their depth measurements. Kolb *et al*. [11] provide a survey about the application areas. The authors in [2, 6] use a similar shading model for active infrared light sources, but apply it to geometry refinement of ToF and Kinect depth images.

Salamati *et al*. [16] propose to capture a color and an infrared image to detect and remove shadows. However they use no active infrared illumination, but capture the ambient infrared light.

In Section 3 we describe the camera system, the image formation model and the noise characteristics. Section 4 details how we use the infrared camera to estimate the camera motion and estimate an illumination free scene model in the infrared spectrum. Afterwards Section 5 shows how to employ the information from the infrared model to separate illumination and color information in the visible light spectrum. Finally Section 6 gives a qualitative evaluation of our approach on real world data.

3. ToF RGB-D Camera

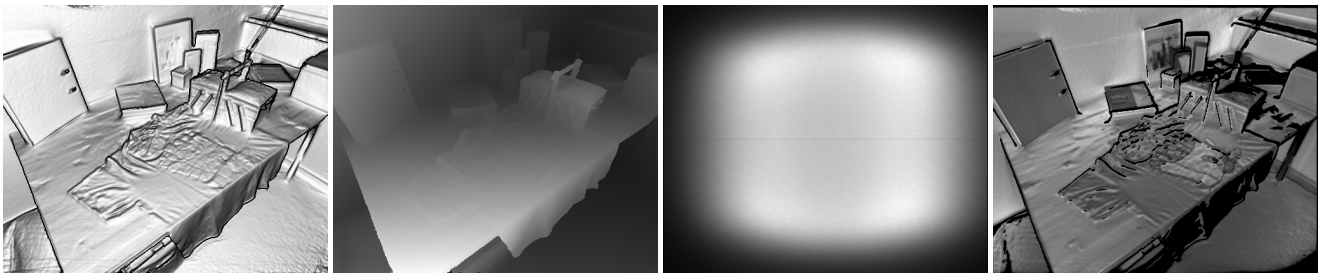
An integral part of our approach is the use of a ToF-based RGB-D camera, because it provides several advantages. First, the infrared intensity image shows only illumination effects due to the integrated light source. Furthermore, the infrared intensity image shows almost no shadows, because the infrared camera and LEDs are mounted close to eachother. These properties alllow us to efficiently remove the illumination from the infrared image by modeling the effects of the single light source. Furthermore, the infrared intensity and depth images are perfectly time-synchronized and registered.

Next we introduce the Kinect One RGB-D camera system. Then we describe the image formation model. Afterwards, we present a noise characterization of the camera system derived from empirical measurements.

3.1. Kinect One Camera

The Microsoft Kinect One consists of two cameras, namely a standard color camera and a time-of-flight (ToF) depth camera. The color camera has a resolution of 1920×1080 and the infrared camera has a resolution of 512×424 pixels.

The Kinect One provides for every frame a color image $\mathcal{C}$, a depth image $\mathcal{Z}$ at a framerate of $30\,\mathrm{Hz}$, and — and this is in contrast to the previous Kinect — a so-called active infrared intensity image $\mathcal{I}$. This active image captures only the illumination induced by the modulated light source of the Kinect One. All unmodulated light, *i.e.*, such as daylight or room light is filtered out by the sensor. The key idea behind our approach is now as follows: As the active



(a) Diffuse term k_d (b) Depth dependent attenuation $\|\boldsymbol{v}\|^{-2}$ (c) Pixel dependent attenuation $\mathcal{K}$ (d) Infrared albedo $\mathcal{A}$

Figure 2: Example of the shading model for the infrared intensity image in Figure 1c: (a) the diffuse component, (b) the depth dependent attenuation term, (c) the calibrated, constant factor, and (d) the infrared albedo image calculated with (6).

image is only illuminated by a known light source, we can reconstruct the albedo (or reflection coefficient) of all surfaces in the scene. Under the assumption that the albedo in the IR spectrum is comparable to the albedo in the visible spectrum (RGB), we can remove shading from the color image and thus create an illumination-free RGB image.

In the following we briefly describe the depth measurement principle and the specific implementation of the Kinect One. The working principle of ToF cameras is that the device sends out amplitude modulated light and measures the reflected light at different position of the light wave, *i.e.*, different phases. From multiple measurements one can derive three quantities: the signal amplitude, the phase shift and the offset. The amplitude is the amount of reflected light originating from the active light source. The amplitude image is also called active image. The phase shift is used to compute the depth of the scene. The offset not only includes reflected light from the active source, but also ambient light. The depth can only be computed up to an ambiguity from the phase shift, because depending on the wavelength, the same phase shift repeats for multiples of a certain distance. This ambiguity can be resolved by using multiple modulation frequencies and selecting the distance on which all phase shifts agree.

Most popular ToF cameras use four measurements to compute amplitude and phase shift [7]. In contrast, the Kinect One only uses three measurements. Additionally, it uses three different signal frequencies. Hence a single amplitude and depth image pair are computed from 9 raw measurement images. An additional 10th measurement is acquired without any active illumination. The infrared camera uses global shutter. Nevertheless, motion artifacts still occur during fast motion, because the active and depth image comprise 9 raw images captured at different time instances. To acquire images, we use the Microsoft SDK, but the free Linux driver also gives access to the intermediate raw images.

3.2. Camera Model

We model both cameras using a standard pinhole camera model, which describes the projection of 3D points to 2D pixel coordinates. Every pixel $\boldsymbol{x}$ is defined by its 2D coordinates $(x, y)^\mathsf{T}$, while a 3D point $\boldsymbol{p}$ consists of three coordinates $(X, Y, Z)^\mathsf{T}$. The projection function π relates the pixel coordinates and the 3D coordinates of a point in the camera coordinate system

$$\boldsymbol{x} = \pi(\boldsymbol{p}) = \left(\frac{Xf_x}{Z} + o_x, \frac{Yf_y}{Z} + o_y\right)^\mathsf{T} \tag{1}$$

where f_x and f_y are the focal length and o_x and o_y the pixel coordinates of the camera center. Given a pixel $\boldsymbol{x}$ and its depth measurement $z = \mathcal{Z}(\boldsymbol{x})$ we can reconstruct the corresponding 3D point by inverting the projection function π, *i.e.*,

$$\boldsymbol{p} = \pi^{-1}(\boldsymbol{x}, z) = \left(\frac{x - o_x}{f_x}z, \frac{y - o_y}{f_y}z, z\right)^\mathsf{T}. \tag{2}$$

As the color, infrared and depth images exhibit radial distortion, we apply a standard distortion model to undistort them. Furthermore, we performed an extrinsic calibration between the color and IR camera so that we can register the color images to the infrared intensity and depth images.

In this section we established the relationship between 3D scene geometry and the depth measurements. Next we describe how the infrared albedo of the scene is related to the observed intensities.

3.3. Infrared Shading Model

Our goal is to specify a shading model, which describes how the albedo $\mathcal{A}$ of a surface point $\boldsymbol{p}$ is related to the observed infrared intensity $\mathcal{I}(\boldsymbol{x})$ in the active image. For this, we assume the diffuse shading model, *i.e.*,

$$\mathcal{I}(\boldsymbol{x}) = k_a k_d \mathcal{A}(\boldsymbol{x}) \tag{3}$$

where k_d is the diffuse term due to Lambert's law, and k_a is a factor describing the light attenuation due to the distance to the light source. Note this model only applies to diffuse (Lambertian) surfaces, *i.e.*, we currently do not model specularities. The diffuse term k_d is defined as

$$k_d = \boldsymbol{n} \cdot \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \tag{4}$$

with $\boldsymbol{n}$ the surface normal and $\boldsymbol{v}$ the vector from the point to the camera center. In our case, the camera center coincides with the origin of the world coordinate system. Therefore, we can set $\boldsymbol{v} = -\boldsymbol{p}$. For the attenuation term k_a, we use

$$k_a = \|\boldsymbol{v}\|^{-2} \mathcal{K}(\boldsymbol{x}) \tag{5}$$

where the first term is the standard light attenuation model for a point light. It describes that the observed intensity of a point decreases with the squared distance to the light source. Note that we approximate the distance to the light source with the distance to the camera center, because on the Kinect One, the modulated light source and the ToF sensor are only a few centimeters apart. From additional experiments, we found that the infrared LEDs are not an ideal point light source (*i.e.*, uniform light in all directions). Therefore, we introduce an additional per pixel factor $\mathcal{K}(\boldsymbol{x})$. The factor accounts for this deviation and other intensity changes introduced by the camera optics such as vignetting. Note that this correction factor is fixed per pixel, because the relative pose of the pixel and the light source never changes.

Figure 2 illustrates the different components of the proposed infrared shading model. Given an infrared intensity image and a corresponding depth map we compute k_d and k_a and solve (3) for $\mathcal{A}$. The formula for each pixel is

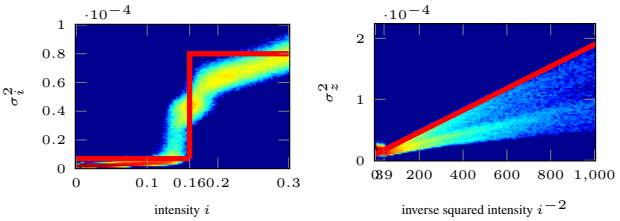
$$\mathcal{A}(\boldsymbol{x}) = \frac{\mathcal{I}(\boldsymbol{x})}{k_d k_a}. \tag{6}$$

To calibrate the pixel-wise factors $\mathcal{K}$, we need to observe a scene with known albedo a. Therefore, we took a series of N images of a white wall, for which we assume a constant albedo value of $a = 1$. We fit a plane to every depth map to obtain practically noise-free estimates of $\boldsymbol{n}$ and $\boldsymbol{v}$. Subsequently, we compute the coefficient $\mathcal{K}(\boldsymbol{x})$ for every pixel $\boldsymbol{x}$ as the weighted average of the N normalized intensity samples, *i.e.*,

$$\mathcal{K}(\boldsymbol{x}) = \frac{1}{\sum_{j=1}^{N} w_j} \sum_{j=1}^{N} w_j \frac{\|\boldsymbol{v}_j\|^2}{k_{d,j}} \mathcal{I}_j(\boldsymbol{x}) \tag{7}$$

where w_j is the inverse of the transformed intensity variance

$$w_j = \frac{k_{d,j}^2}{\|\boldsymbol{v}_j\|^4 \sigma_{i,j}^2}. \tag{8}$$



(a) Infrared intensity noise (b) Depth noise

Figure 3: Empirical noise models and analytical fits for the variance of the infrared intensity (a) and depth values (b). Colors indicate how often a certain variance was observed (dark blue: never, red: often).

We give a definition for the intensity variance σ_i^2 in Section 3.4. The same shading model was used for depth map enhancement in [2].

3.4. Camera Noise Model

Different steps in our pipeline require an estimate of the noise of the measured infrared intensity and depth values. For both quantities, we assume a corruption by zero mean Gaussian noise. To empirically determine the mean and variance for the Kinect One sensor, we recorded sequences of static scenes. The plot in Figure 3a shows the relationship between the mean intensity and its variance σ_i^2. The variance in dark regions $[0, 0.16]$ is almost constant, while in brighter regions $[0.16, 0.3]$, we observed a substantial increase of up to a factor of 10. Note we only observed few pixels above 0.3. Therefore, we decided to model the intensity variance as follows:

$$\sigma_i^2(i) = \begin{cases} 0.7 \cdot 10^{-5}, & \text{if } i < 0.16 \\ 0.8 \cdot 10^{-4}, & \text{otherwise} \end{cases} \tag{9}$$

For the variance of the depth measurements, Frank *et al.* [8] establish the relationship $\sigma(\boldsymbol{x})_z^2 \propto \mathcal{I}(\boldsymbol{x})^{-2}$, i.e, the variance of the depth measurements is inversely proportional to the square of the measured intensity. We experimentally verified this relationship. Figure 3a shows the distribution of the empirical depth variance w.r.t. the inverse squared intensity. We use the following linear model to predict the depth variance σ_z^2 from the inverse squared intensity

$$\sigma_z^2(i) = (0.0018 i^{-2} + 0.1) \cdot 10^{-5}\,\mathrm{m}^2. \tag{10}$$

The red line in Figure 3b represents this model. Our model is a conservative estimate of the true depth variance. We found that the depth variance is higher than predicted for $i^{-2} \in [1, 39]$, which corresponds approximately to $i \in [0.16, 1]$. In this intensity interval we observed an increase in the intensity variance as well. We attribute this to special characteristics of the sensor.

4. Camera Motion Estimation and Fusion

We apply a dense, direct motion estimation algorithm on the illumination free infrared and depth images, which we obtain by inverting the previously detailed image formation model. Once we know the camera motion, we fuse multiple infrared albedo and depth images to reduce the noise. For efficient fusion we represent the scene as a set of keyframes. The following explanations describe our approach for one keyframe, but it can be easily extended to multiple keyframes.

4.1. Camera Motion Estimation

To estimate the camera motion we use a dense, direct image alignment approach, which minimizes the intensity and depth error [10]. Instead of using the intensity image provided by the color camera we use the infrared intensity image. This has the advantage that the intensity and depth image are time synchronized and spatially registered, which improves the performance of the tracking algorithm. The motion estimation algorithm assumes constant illumination, which holds for the intensity image obtained with the color camera, but not for the infrared image. Therefore, we normalize each intensity image by inverting the shading model using (6).

Note that for the normalization of a new infrared image, we have to use the raw depth map because we do not know the relative position to our fused world model yet. However, the noise of the raw depth map disturbs the normals, from which the diffuse shading coefficient k_d is computed. Therefore, we exclude the diffuse coefficient k_d in this normalization step.

After the normalized infrared image has been computed, we estimate the relative transformation $\boldsymbol{T}$ by minimizing the following error over all pixels

$$E(\boldsymbol{T}) = \int_{\Omega} w(\boldsymbol{x}) \boldsymbol{e}(\boldsymbol{x}, \boldsymbol{T})^{\mathsf{T}} \boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1} \boldsymbol{e}(\boldsymbol{x}, \boldsymbol{T}) \mathrm{d}\boldsymbol{x}. \tag{11}$$

The error for one pixel comprises the intensity and depth error

$$\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{T}) = \begin{pmatrix} \frac{\mathcal{I}_2(\pi(\boldsymbol{p}'))}{k_{a,2}} - \frac{\mathcal{I}_1(\boldsymbol{x})}{k_{a,1}} \\ \mathcal{Z}_2(\pi(\boldsymbol{p}')) - [\boldsymbol{p}']_Z \end{pmatrix} \tag{12}$$

where

$$\boldsymbol{p}' = \boldsymbol{T}\, \pi^{-1}(\boldsymbol{x}, \mathcal{Z}_1(\boldsymbol{x})) \tag{13}$$

is the point $\boldsymbol{p}$ transformed into the view of the second camera and $[\,\cdot\,]_Z$ selects the Z coordinate of a point. Note that we denote the image domain by $\Omega \subset \mathbb{R}^2$. As $\pi(\boldsymbol{p}')$ is non-linear in the camera motion $\boldsymbol{T}$, we apply a non-linear Gauss-Newton method to find the minimum. Furthermore, the per error weight w_j and the error covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{e}}$ are derived from a robust, bivariate Student's t-distribution and estimated during the optimization of (11).

4.2. Albedo and Depth Map Fusion

We represent our 3D model as a set of keyframes similar to [10, 13]: Each keyframe comprises an infrared albedo image $\mathcal{A}_{\mathrm{kf}}$, a depth image $\mathcal{Z}_{\mathrm{kf}}$ and a color image $\mathcal{C}_{\mathrm{kf}}$. For the albedo and depth map we additionally store images $\Sigma_{z,\mathrm{kf}}$ and $\Sigma_{a,\mathrm{kf}}$ containing a per-pixel variance estimate. We fuse all raw infrared and depth images that we could successfully register to a keyframe, to obtain a denoised albedo and depth image for each keyframe. Assuming the measurement noise model presented in Section 3.4 the fusion for both images can be performed efficiently using a weighted average.

Given the keyframe depth image $\mathcal{Z}_{\mathrm{kf}}$ and the relative transformation $\boldsymbol{T}$ to the current camera position we can compute for every pixel $\boldsymbol{x}$ in the keyframe its corresponding pixel coordinates $\boldsymbol{x}' = \pi(\boldsymbol{p}')$ in the current camera using $\boldsymbol{p}'$ from (13). With this relationship we can update our current estimates of the infrared albedo $\mathcal{A}_{\mathrm{kf}}$ and the depth $\mathcal{Z}_{\mathrm{kf}}$. The per-pixel update for the keyframe depth image is

$$\hat{\mathcal{Z}}_{\mathrm{kf}}(\boldsymbol{x}) = \frac{\Sigma_{z,\mathrm{kf}}^{-1}(\boldsymbol{x}) \mathcal{Z}_{\mathrm{kf}}(\boldsymbol{x}) + \Sigma_{z,t}^{-1}(\boldsymbol{x}') \mathcal{Z}_t(\boldsymbol{x}')}{\hat{\Sigma}_{z,\mathrm{kf}}^{-1}(\boldsymbol{x})}. \tag{14}$$

The updated variance estimate is

$$\hat{\Sigma}_{z,\mathrm{kf}}(\boldsymbol{x}) = \left(\Sigma_{z,\mathrm{kf}}^{-1}(\boldsymbol{x}) + \Sigma_{z,t}^{-1}(\boldsymbol{x}') \right)^{-1}. \tag{15}$$

We calculate the variance of the current depth image $\Sigma_{z,t}$ according to the noise model (10), *i.e.*, $\Sigma_{z,t}(\boldsymbol{x}) = \sigma_z^2\left(\mathcal{I}_t(\boldsymbol{x})\right)$. We only update the depth estimate, if the following chi-square test is successful

$$\frac{0.5\left(\mathcal{Z}_{\mathrm{kf}}(\boldsymbol{x}) - \mathcal{Z}_t(\boldsymbol{x}')\right)^2}{\Sigma_{z,\mathrm{kf}}^{-1}(\boldsymbol{x}) + \Sigma_{z,t}^{-1}(\boldsymbol{x}')} < 10.83 \tag{16}$$

The value 10.83 corresponds to 99.9% confidence that both measurements belong to the same distribution. This allows us to reject outliers caused by occlusions or sensor failures. In practice we warp $\mathcal{Z}_t$ with $\boldsymbol{T}^{-1}$ to the keyframe instead of using the inverse warping described by (13). The forward warping is efficiently done using a hardware-accelerated rendering pipeline like OpenGL.

To update the infrared albedo image $\mathcal{A}_{\mathrm{kf}}$ we first have to remove the shading effects from the current infrared image to obtain the current albedo image $\mathcal{A}_t$. Using the current depth image for this task would introduce additional noise. Therefore, we use the keyframe depth image. We warp all depth values with a variance $\sigma_z^2 < 0.001^2\,\mathrm{m}^2$ to the viewpoint of the current image. Afterwards, we invert the shading model with these warped depth values using (6) to obtain $\mathcal{A}_t$. Then we can look up the current albedo value and its variance estimate for every pixel in the keyframe albedo

image and fuse them using the following weighted average formula

$$\hat{\mathcal{A}}_{\text{kf}}(\boldsymbol{x}) = \frac{\Sigma_{a,\text{kf}}^{-1}(\boldsymbol{x})\mathcal{A}_{\text{kf}}(\boldsymbol{x}) + \Sigma_{a,t}^{-1}(\boldsymbol{x}')\mathcal{A}_t(\boldsymbol{x}')}{\hat{\Sigma}_{a,\text{kf}}^{-1}(\boldsymbol{x})}. \quad (17)$$

Similar to (15) we update each infrared albedo variance estimate

$$\hat{\Sigma}_{a,\text{kf}}(\boldsymbol{x}) = \left(\Sigma_{a,\text{kf}}^{-1}(\boldsymbol{x}) + \Sigma_{a,t}^{-1}(\boldsymbol{x}')\right)^{-1}. \quad (18)$$

We compute the current infrared albedo variance as:

$$\Sigma_{a,t}(\boldsymbol{x}) = \frac{\sigma_i^2(\mathcal{I}(\boldsymbol{x}))}{(k_d k_a)^2}. \quad (19)$$

We employ the same chi-square test as in (16) to check whether both pixels belong to the same surface point.

Note that in principle, the same strategy should apply to fuse multiple color images into a single keyframe. In practice, however, we found that this degrades the quality of the resulting color image more than it is enhanced. We believe that this is due to the auto white balance feature, rolling shutter effects and imprecise registration between the infrared and color camera. Therefore, in our current implementation, we keep for every keyframe the original color image, but believe that these issues can be resolved when better sensors become available in the near future.

5. Illumination Correction for Color Images

The main idea of our approach is to utilize the infrared albedo image $\mathcal{A}$ to remove the illumination effects from the corresponding color image $\mathcal{C}$. With this additional information it should be easier to resolve the ambiguity inherent in intrinsic decomposition approaches. With the infrared albedo image we know which parts of the image have a constant reflectance and we can use weaker priors. We assume the albedo images are piecewise constant as is commonly done in intrinsic image decomposition [1]. Therefore, the region boundaries in the color and infrared should coincide. The corresponding, constant color albedo and infrared albedo values of a specific material are related by an unknown, constant factor. However, this factor cannot be recovered, because it depends on the material and the illumination.

5.1. Color Shading Model

Our goal is to find the color albedo image $\mathcal{U}$ given a color image $\mathcal{C}$ and a infrared albedo image $\mathcal{A}$. Therefore, we use the standard intrinsic image decomposition formulation:

$$\mathcal{C} = \mathcal{S}\mathcal{U}, \quad (20)$$

i.e. the observed color image $\mathcal{C}$ is composed of a shading image $\mathcal{S} : \Omega \to \mathbb{R}$, which includes all illumination effects, and the color albedo image $\mathcal{U} : \Omega \to \mathbb{R}^3$. Furthermore, we impose the following soft constraint

$$\mathcal{A} = \boldsymbol{G}\mathcal{U} \quad (21)$$

to couple the color and infrared albedo image. $\boldsymbol{G}(\boldsymbol{x})$ is a 1×3 matrix, which combines the three color albedo values to a single intensity value and is constant for the whole domain Ω. We compute $\boldsymbol{G}$ as the average of the infrared albedo image $\mathcal{A}$ divided by the per channel average value of $\mathcal{C}$. This coupling leads to the same average value in the input color image and the color albedo image.

We cast the problem of finding the color albedo image $\mathcal{U}$ and the shading image $\mathcal{S}$ as a variational problem which we solve using a state of the art first-order solver. The energy includes both constraints (20) and (21) and the total variation of $\mathcal{U}$ and $\mathcal{S}$ in order to obtain spatially consistent results.

$$\begin{aligned} E(\mathcal{U},\mathcal{S}) = \lambda \int_\Omega \|\mathcal{C} - \mathcal{S}\mathcal{U}\|_2^2 \mathrm{d}\boldsymbol{x} + \gamma \int_\Omega \|\mathcal{A} - \boldsymbol{G}\mathcal{U}\|^2 \mathrm{d}\boldsymbol{x} \\ + \int_\Omega g_1(\boldsymbol{x})|\nabla\mathcal{U}| + \int_\Omega g_2(\boldsymbol{x})|\nabla\mathcal{S}| \\ \text{s.t.} \quad \mathcal{U}(\boldsymbol{x}) \in [0,1]^3, \mathcal{S}(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \Omega \end{aligned} \quad (22)$$

Note that we utilize a weighted version of the total variation for both $\mathcal{U}$ and $\mathcal{S}$. The weighting functions $g_1(\boldsymbol{x}) = \exp\left(\frac{-|\nabla\mathcal{A}(\boldsymbol{x})|}{\sigma}\right)$ and $g_2(\boldsymbol{x}) = \exp\left(\frac{-|\nabla\mathcal{L}(\boldsymbol{x})|}{\sigma}\right)$ are dependent on the the gradient of the albedo $\mathcal{A}$ and the gradient of the luminance respectively. This ensures that the edges of the obtained albedo image $\mathcal{U}$ and the shading image $\mathcal{S}$ are aligned to the ones of the given images $\mathcal{A}$ and $\mathcal{L}$. Additionally, we constrain the values for the albedo color image $\mathcal{U}$ to lie in the interval $[0, 1]$ and assume that the shading function $\mathcal{U}$ exhibits positive values in each pixel $\boldsymbol{x} \in \Omega$.

5.2. Optimization

Note that functional (22) is a difficult non-convex optimization problem. However, as the energy is convex in $\mathcal{U}$ and $\mathcal{S}$, it can be solved using an alternating optimization scheme where one minimizes the energy $E(\mathcal{U},\mathcal{S})$ alternatively for $\mathcal{U}$ and $\mathcal{S}$ by respectively fixing the other variable. The overall procedure is illustrated in Algorithm 1:

Still, each sub-problem is a huge constrained non-smooth optimization problem which cannot be solved using standard solvers (*e.g.* the interior point method). However, there has been a lot of development in devising first-order solvers which can deal with this class of problems. In particular, we employ the recent primal-dual optimization algorithm [4, 3], which essentially performs a gradient ascent in the dual variable and a gradient descent in the primal variable. For this we need to rewrite energy (22) in its respective

Algorithm 1 Alternating Optimization for Joint Albedo-Shading Estimation

Initialize $\mathcal{U}^0$ and $\mathcal{S}^0$

1: **for** $k = 0, 1, 2, ...$ **do**
2: $\mathcal{U}^{k+1} = \mathbf{argmin}_{\mathcal{U}} E(\mathcal{U}, \mathcal{S}^k)$
3: $\mathcal{S}^{k+1} = \mathbf{argmin}_{\mathcal{S}} E(\mathcal{U}^{k+1}, \mathcal{S})$
4: **end for**

saddle point formulation in order to tackle it via the primal-dual optimization. The primal-dual formulation of energy (22) can be written as follows:

$$\begin{aligned} E(\mathcal{U},\mathcal{S}) = \lambda \int_\Omega \|\mathcal{C} - \mathcal{S}\mathcal{U}\| \,\mathrm{d}\boldsymbol{x} + \gamma \int_\Omega \|\mathcal{A} - \boldsymbol{G}\mathcal{U}\| \,\mathrm{d}\boldsymbol{x} \\ + \sup_{\mathcal{P}} \int_\Omega < \mathcal{P}(\boldsymbol{x}), \nabla\mathcal{U}(\boldsymbol{x}) > \mathrm{d}\boldsymbol{x} \\ + \sup_{\mathcal{Q}} \int_\Omega < \mathcal{Q}(\boldsymbol{x}), \nabla\mathcal{S}(\boldsymbol{x}) > \mathrm{d}\boldsymbol{x} \\ \text{s.t.} \quad \mathcal{U}(\boldsymbol{x}) \in [0,1]^3, \mathcal{S}(\boldsymbol{x}) \geq 0, \\ |\mathcal{P}(\boldsymbol{x})| \leq g_1(\boldsymbol{x}), |\mathcal{Q}(\boldsymbol{x})| \leq g_2(\boldsymbol{x}) \;\; \forall \boldsymbol{x} \in \Omega \end{aligned} \tag{23}$$

Note that the auxiliary dual variables $\mathcal{P} : \Omega \to \mathbb{R}^6$ and $\mathcal{Q} : \Omega \to \mathbb{R}^2$ stem from incorporating the dual formulation of the total variation which is differentiable in both the dual variables and in the primal variables $\mathcal{U}$ and $\mathcal{S}$, in contrast to the original formulation in (22). Using formulation (23) we can tackle each sub-problem in Algorithm 1 by the primal-dual algorithm described in [3]. For details on how to solve the sub-problems we refer the reader to the supplementary materials.

6. Evaluation

We provide a qualitative evaluation for the proposed approach, because there exists no benchmark for intrinsic image decomposition, which also includes infrared image data. For the evaluation of our approach we recorded different scenes with the Kinect One. Figure 4 shows three different indoor scenes. The left column is the input color image. The second column shows the estimated infrared albedo image. The third column displays the recoverd color albedo image $\mathcal{U}$ and the right column is the shading image $\mathcal{S}$. The scene in the first row is illuminated by a ceiling lamp and shows only little shadows. For the desk scenes in the second and third row we added a table lamp, which introduces bright highlights and harsh shadows. In principle our algorithm can remove both effects and recovers plausible color albedo images, e.g, on the red box, table, wall and t-shirt. In case of imperfections in the infrared albedo image, the approach introduces some artifacts. Furthermore, the smoothing effect is too strong for the color albedo image. We hope to address these problems through non-local regularizers. Additionally, regularization on a gamma mapped color image should be beneficial, because small gradients in shadowed areas are amplified and the smoothing term gets less influence. Finally, our simple coupling of the color and infrared albedo images causes sometimes too bright colors in the resulting albedo image. We hope to resolve this issue by better modeling the assumption that both albedo images share piecewise constant regions.

In the supplementary material we show additional scenes and also provide results computed with the algorithm of Chen *et al*. [5]. This algorithm reports state of the art performance for RGB-D intrinsic image decomposition only utilizing color and depth information.

7. Conclusion

The goal of this work is to derive illumination-free color images using commodity RGB-D sensors. As the Kinect One is such as sensor, we first analyzed and modeled the light source, ToF sensor and color camera of the Kinect One. Given these models, we were able to compute the infrared albedo from the active image and the depth image. Subsequently, we demonstrated that the IR albedo can in many cases be transferred to the color domain, so that an illumination-free RGB image can be obtained. As regions with strong highlights or shadows contain only little color information, we apply a state-of-the-art regularization scheme to infer color information from neighboring regions. With this, we hope to contribute to the development of methods for illumination-free 3D reconstructions in the near future.

References

[1] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. *CVPR*, 2013. 1, 2, 6

[2] M. Böhme, M. Haker, T. Martinetz, and E. Barth. Shading constraint improves accuracy of time-of-flight measurements. *Computer vision and image understanding*, 114(12), 2010. 2, 4

[3] A. Chambolle, D. Cremers, and T. Pock. A convex approach for computing minimal partitions. Tech. rep. TR-2008-05, University of Bonn, 2008. 6, 7

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *JMIV*, 40(1), 2011. 6

[5] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *Computer Vision (ICCV), 2013 IEEE Intl. Conference on*. IEEE, 2013. 1, 2, 7

[6] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon. Exploiting shading cues in kinect ir images for geometry refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. 2

[7] S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (tof) cameras: a survey. *Sensors Journal, IEEE*, 11(9), 2011. 2, 3

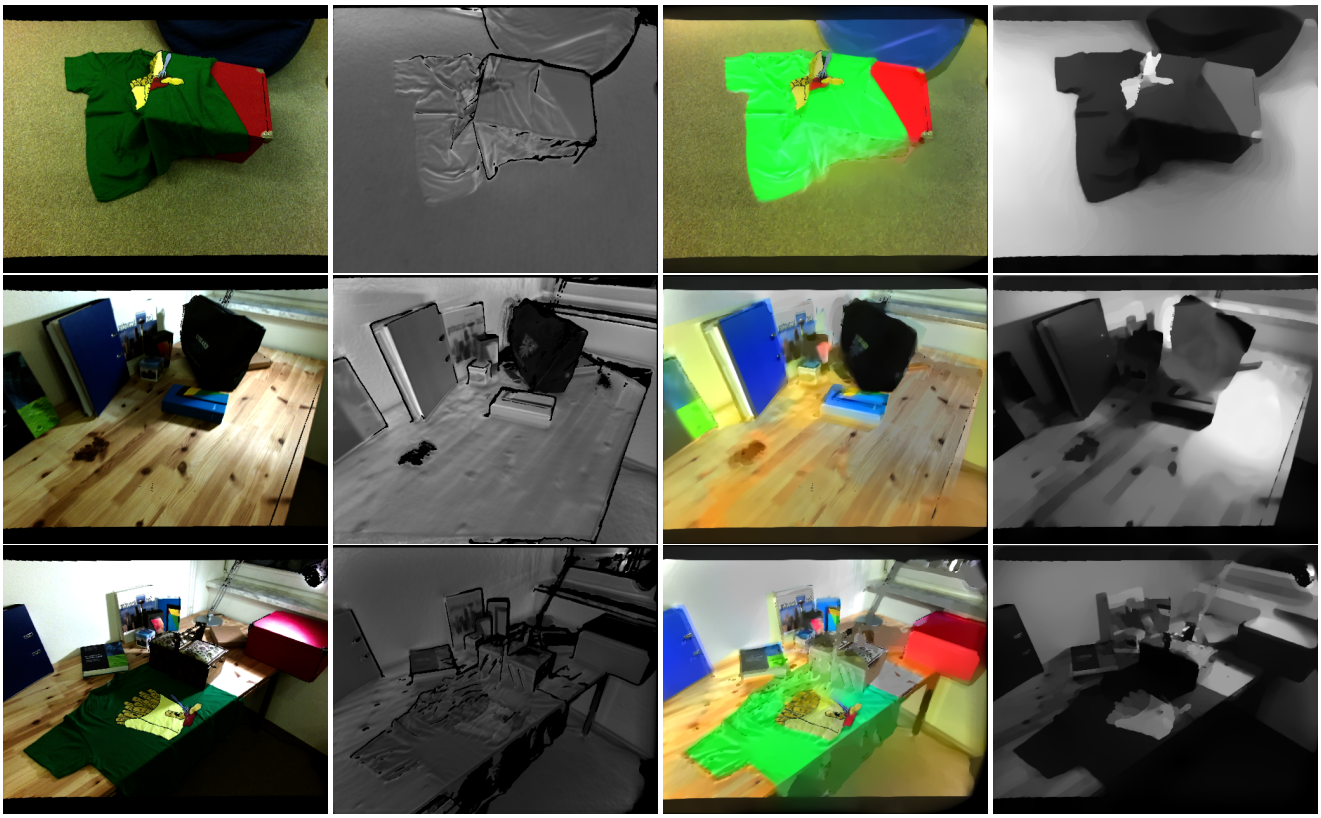


Figure 4: Results of our algorithm on three real world scenes. Left column shows the color image in linear RGB space. The second column is the infrared albedo image $\mathcal{A}$. The third column displays the estimated color albedo image $\mathcal{U}$. On the right is the corresponding shading image $\mathcal{S}$. The proposed approach recovers plausible color albedo images.

[8] M. Frank, M. Plaue, H. Rapp, U. Köthe, B. Jähne, and F. A. Hamprecht. Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras. *Optical Engineering*, 48(1), 2009. 4

[9] J. Jachnik, R. A. Newcombe, and A. J. Davison. Real-time surface light-field capture for augmentation of planar specular surfaces. In *Mixed and Augmented Reality (ISMAR), IEEE Intl. Symposium on*. IEEE, 2012. 2

[10] C. Kerl, J. Sturm, and D. Cremers. Dense visual slam for rgb-d cameras. In *Proc. of the Intl. Conf. on Intelligent Robot Systems (IROS)*, 2013. 5

[11] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29. Wiley Online Library, 2010. 2

[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 *Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, Adelaide, Australia, 2013. 1

[13] M. Meilland and A. Comport. On unifying key-frame and voxel-based dense visual SLAM at large scales. In *Proc. of the Intl. Conf. on Intelligent Robot Systems (IROS)*, Tokyo, Japan, 2013. 5

[14] R. Montes Soldado and C. Ureña Almagro. An overview of brdf models. 2012. 2

[15] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, 2011. 1

[16] N. Salamati, A. Germain, and S. Susstrunk. Removing shadows from images using color and near-infrared. In *Image Processing (ICIP), 18th IEEE Intl. Conference on*. IEEE, 2011. 2

[17] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended KinectFusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012. 1

[18] Y. Xiao, E. Tsougenis, and C.-K. Tang. Shadow removal from single rgb-d images. In *Proc. of the Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[19] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 2