HDRFusion: HDR SLAM using a low-cost auto-exposure RGB-D sensor

Shuda Li
University of Bristol
Bristol, UK
csxsl@bristol.ac.uk

Ankur Handa
Imperial College London
London, UK
handa.ankur@gmail.com

Yang Zhang, Andrew Calway
University of Bristol
Bristol, UK
{Yang.Zhang,csadc}@bristol.ac.uk

Abstract

Most dense RGB/RGB-D SLAM systems require the brightness of 3-D points observed from different viewpoints to be constant. However, in reality, this assumption is difficult to meet even when the surface is Lambertian and illumination is static. One cause is that most cameras automatically tune exposure to adapt to the wide dynamic range of scene radiance, violating the brightness assumption. We describe a novel system - HDRFusion - which turns this apparent drawback into an advantage by fusing LDR frames into an HDR textured volume using a standard RGB-D sensor with auto-exposure (AE) enabled. The key contribution is the use of a normalised metric for frame alignment which is invariant to changes in exposure time. This enables robust tracking in frame-to-model mode and also compensates the exposure accurately so that HDR texture, free of artefacts, can be generated online. We demonstrate that the tracking robustness and accuracy is greatly improved by the approach and that radiance maps can be generated with far greater dynamic range of scene radiance.

1. Introduction

Dense RGB simultaneous localisation and mapping (SLAM) systems, such as that described in [20], align image frames to estimate sensor pose and build textured 3-D reconstructions. They have been shown to give superior performance in comparison to sparse feature-based approaches, giving high tracking accuracy, robustness to motion blur, and dense 3-D scene reconstructions, making them ideal for applications such as augmented reality. Moreover, when combined with depth-based tracking and mapping, they can maintain tracking even when depth information becomes ambiguous [21, 10, 26, 12, 25].

However, frame alignment relies on a brightness constancy assumption, i.e. that the brightness of 3-D points observed from different viewing positions is constant. Approaches can be categorised into using either a global or a local constancy assumption. The former assumes that any two overlapping frames from an input RGB sequence fulfil the condition [20], while the latter requires only that consecutive frames do[12, 25]. The global assumption enables frame-to-model tracking which is known to accumulate less drift [21], while the local assumption is easier to meet in practice but means that the tracking is done frame-to-frame, with a consequent increase in drift. Violation of either assumption may cause RGB tracking to fail or to become noticeably less reliable than tracking with depth alone.

However, both assumptions are frequently broken in reality when using cameras equipped with automatic exposure (AE). AE is designed to map the high dynamic range of scene radiance into a narrow range suitable for the human eye. When the camera moves from a bright to dark area, the exposure time is increased automatically so that more light can be captured by the camera sensor and vice versa when the camera moves from dark to bright regions. This breaks the global assumption since images in a video sequence are seldom captured with the same exposure time. The local assumption is more likely to be met as exposure usually changes smoothly, but this assumption also breaks when video flickering occurs. This can occur when a camera moves across the boundary between a bright and dark area or moves quickly back and forth between them: in these scenarios, the exposure changes dramatically in a short period of time and results in flickering. Turning AE off can ensure the brightness constancy, but it is often undesirable since it leaves bright areas over exposed and dark areas under exposed, leading to the loss of important visual detail.

AE also poses a problem when texturing a 3-D model of the scene. Overlapping images captured with inconsistent brightness will leave mosaic artefacts when projected back onto the model surface. This is a common problem for many dense mapping systems as illustrated in Fig. 1. The problem has been widely addressed in conventional model texturing, panoramic imaging [3] and video tonal stabilization [1, 6]. This is achieved by compensating the global brightness of input images and blending colours along the boundaries between the images to create consistent texture. But these techniques are computationally expensive and are

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.40

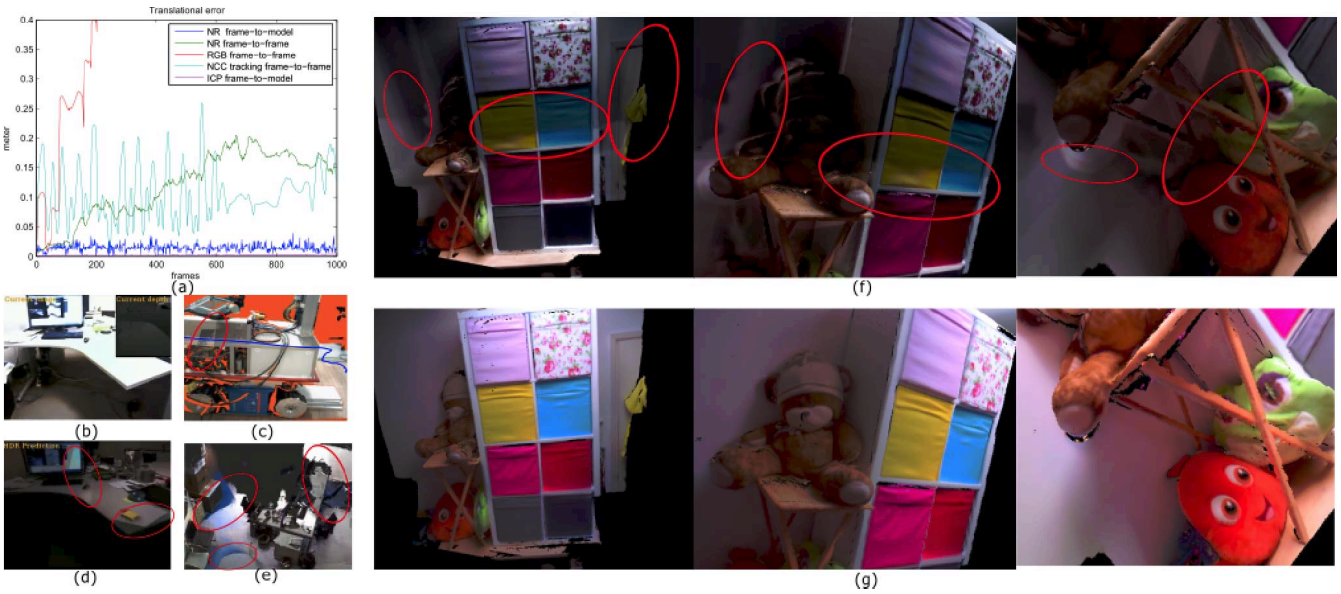

Figure 1: HDRFusion results and comparison: (a) Frame-to-model tracking using normalized exposure/radiance (NR) gives superior tracking accuracy when compared with techniques using raw RGB and NCC on challenging synthetic flickering RGB-D sequences; (b)-(e) Screen captures from videos released with previous approaches illustrate insufficient exposure compensation, leading to artefacts, as highlighted in red. Specifically, (b) and (d) are taken from [19], where (b) is the raw input image and (d) is the predicted scene texture, and (c) and (e) are the predicted scene textures taken from [26] and [11] , respectively; (f) Results from our implementation of [25] showing strong artefacts caused by large camera exposure adjustment when moving from bright areas (top) to dark areas (bottom left); (g) Predicted textures obtained using the proposed HDRFusion, where the HDR textures are visualized using the Mantiuk tone mapping operater [17]. Note the lack of artefacts compared to that in (f).

mainly aimed at delivering visually pleasing results rather than maintaining fidelity to the real world radiance.

In this paper, we introduce a novel technique for dense RGB-D SLAM which allows robust frame-to-model tracking using RGB frames with AE enabled. It is very robust to brightness fluctuation and is capable of capturing a consistent HDR texture on the surface of the 3-D scene reconstruction which is free of artefacts as illustrated in Fig. 1.

The key assumption of the work is that the radiance in a real world scene is constant when the illumination in the scene is static[1]. Thus, if we can base tracking on a measure related to radiance rather than image brightness, then we can avoid the difficulties encountered when AE is operating. To do so, we make use of *normalised exposure* - we compute exposure from image intensity using the inverse camera response function and then normalise the exposure by subtracting the mean and dividing by the standard deviation within a small neighbourhood of each pixel. We show that under the above assumptions the resulting normalised values are independent of exposure time and depend only on scene radiance, hence making them invariant to intensity changes caused by AE. A further advantage is that the normalization operation can be efficiently implemented using down-sampled integral images. The approach is in contrast to jointly tracking and compensating exposure as in previous work [19] which we show is not as robust and reliable as our proposed method.

[1]Scene radiance is the amount light reflected per unit solid angle in a given direction by unit area. Assuming a pinhole camera model, scene radiance is equivalent to sensor irradiance which is the amount of light energy received on the camera sensor per unit area and unit time. Exposure is then the integration of irradiance during exposure time [4].

1.1. Overview

We now give an overview of HDRFusion. The main algorithm is shown in the flow-chart in Fig. 2. The inputs are RGB-D frames from a Xtion Pro sensor. Firstly, we estimate the inverse camera response function and noise level function in radiometric calibration (Section 4). The RGB frames are then converted into exposure maps with estimated confidence maps derived from the noise level function. The camera poses are tracked by aligning incoming frames with predicted frames obtained from the textured 3-D reconstruction, *i.e.*, by registering the live normalized exposure with predicted normalized radiance. This is described in section 5. The predicted normalized radiance is estimated by casting rays into the global volume and the confidence map is used to adaptively weight the error function for tracking, exposure time compensation and radiance fusion. The ray casting module establishes predicted radiance, normalized radiance and depth. The predicted radi-

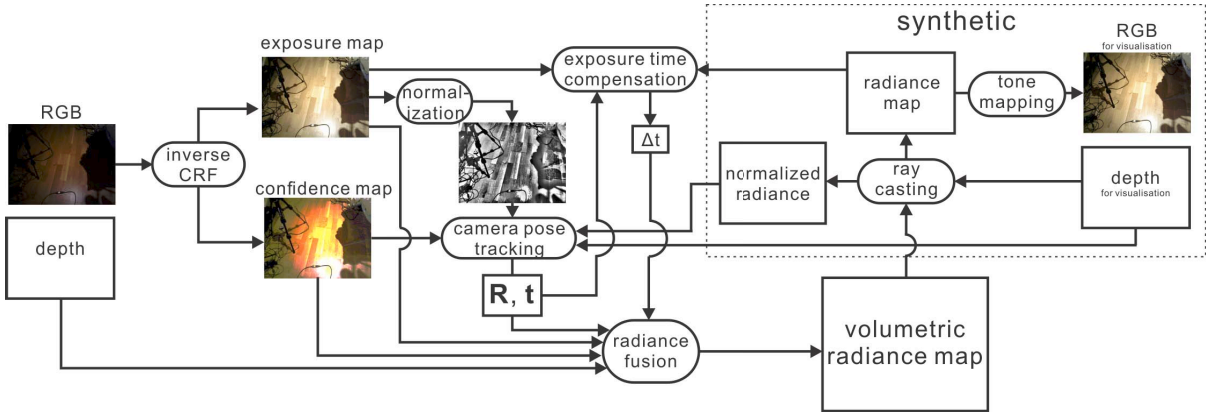

Figure 2: Flow chart of HDRFusion. The boxes represent data structures, eclipses represent data transforming modules and arrows represent data flow. From left, input RGB frames are converted into exposure map through inverse camera response function. The camera pose is tracked in a frame-to-model style. Note that the confidence map is also applied in exposure compensation but for simplicity the data flow is not shown.

ance and depth can then be used for visualization through tone mapping on LDR devices or directly output as HDR data.

2. Related work

There is a huge wealth of literature on dealing with visual odometry or camera motion tracking. However, we will focus on the direct approaches which can track and reconstruct a dense and textured 3-D model in real-time. Motion tracking using depth from an active sensor [21, 10, 23] is independent of lighting but leaves surfaces un-textured. Approaches combining appearance and depth [12, 24, 25, 11] are able to enhance the depth only tracking approaches when dealing with large planar scenes. These approaches are the most relevant to the work described in this paper. However, as pointed out before, they all rely on brightness constancy which limit them to frame-to-frame RGB tracking and as a consequence they are very likely to fail when video flickering happens, for instance. For texturing, [25] describes a simple color blending method but as shown in our experiments, it is inadequate to deal with large exposure changes. Kerl *et al.* [11] describe a key frame based approach by taking the rolling shutter effect into account. This relies on local brightness constancy when tracking live frames with a key frame and is capable of producing sharp super-resolution frames involving no exposure compensation.

Maxime *et al.* [19] pioneered real-time 3-D HDR texture capture and their method is also capable of re-lighting virtual specular objects. The differences between [19] and the work in this paper are two-fold. First, in [19], a gamma function is adopted to approximate the inverse CRF. The gamma function may introduce error when the radiance is high and the resulting radiance is not directly proportional to scene radiance (see Fig. 3). Second, in [19], the exposure is estimated jointly with camera pose, but we find that the shape of the error function when tracking using exposure compensated radiance shows shallow global minima even when the exposure has been compensated for. Consequently, their approach is not as robust as the normalized exposure function used in this work (see Fig. 4). Lastly, mosaic artefacts are clearly visible in their results which indicates inadequate exposure estimation (see Fig. 1(d)).

Normalized cross correlation (NCC) has been widely applied in visual tracking [22] to deal with challenging lighting conditions but its computational cost grows exponentially with the size of patch. Small patches are sensitive to image noise and bring many local minima (Fig. 4). In addition, 3-D HDR texture capturing is not addressed in the paper. HDR video capture using a high-end stereo rig [9, 2] is also relevant to the topic since it involves estimating disparity between binocular views so that LDR frames captured by both frame can be integrated into a single stream of HDR video [2], but the high quality HDR video is the main focus of those methods rather than generating a full 3-D model.

3. Preliminaries

Assuming brightness is constant, camera poses can be estimated by minimizing the intensity difference between corresponding pixels from a reference frame to a live frame. The objective function F can be formulated as:

$$F(\mathbf{R},\mathbf{t}) = \int_{\Omega} \left\| I_r(\mathbf{u}) - I_l(\pi(\mathbf{R}\pi^{-1}(\mathbf{u}, D_r(\mathbf{u})) + \mathbf{t})) \right\|_2^2 d\mathbf{u} \tag{1}$$

where $I : \Omega \to \mathbb{R}_+$ and $D : \Omega \to \mathbb{R}_+$ denote the intensity and depth functions. Note that (1) can be extended to RGB channels. The 2-D image domain is denoted by $\Omega \subset \mathbb{R}^2$ and a pixel coordinate on the image by $\mathbf{u} \in \mathbb{R}^2$. Subscripts

$_r$ and $_l$ denote the reference frame and live frame, respectively. $\mathbf{R} \in \mathbb{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the rigid body motion to transform a 3-D point defined in the reference coordinate system to the live coordinate system. The functions $\pi : \mathbb{R}^3 \rightarrow \Omega$ and $\pi^{-1} : \Omega \times \mathbb{R}_+ \rightarrow \mathbb{R}^3$ are the projection function and its inverse, i.e. $\pi(.)$ projects a 3-D point onto the 2-D image plane and $\pi^{-1}(.)$ transforms a 2-D point into a 3-D point given its depth D.

Equation 1 works as long as brightness constancy holds. Correspondence between $\mathbf{u}'$ from the live frame and $\mathbf{u}$ from the reference must fulfil the equation $\mathbf{u}' = \pi(\mathbf{R}\pi^{-1}(\mathbf{u}, D_r(\mathbf{u})) + \mathbf{t})$. The difference between them is denoted by $e(\mathbf{u}, \mathbf{u}') = I_r(\mathbf{u}) - I_l(\mathbf{u}')$. Equation 1 can be rewritten as $F(\mathbf{R}, \mathbf{t}) = \int_\Omega \|e(\mathbf{u}, \mathbf{u}')\|^2 \, d\mathbf{u}$. The NCC based camera tracking method can be viewed as an extension of (1) by replacing $e(\mathbf{u}, \mathbf{u}')$ with $\sqrt{1 - C^2(\mathbf{u}, \mathbf{u}')}$, where $C(.)$ is the NCC score defined as

$$C(\mathbf{u}, \mathbf{u}') = \frac{1}{|\Omega_N|^2} \int_{\Omega_N} \frac{(N_r(\mathbf{u}, \mathbf{v}) - \mu)(N_l(\mathbf{u}', \mathbf{v}) - \mu')}{\sigma\sigma'} d\mathbf{v} \tag{2}$$

where $N : \Omega \times \Omega_N \rightarrow \mathbb{R}_+$ defines a small image patch, a neighbourhood centred at $\mathbf{u}$. $\Omega_N \subset \mathbb{R}^2$ is the domain of the neighbourhood N and $\mathbf{v} \in \mathbb{R}^2$ is the coordinate w.r.t. N. μ and σ are mean and std. (standard deviation) of image intensity over N_r and μ' and σ' are mean and std. over N_l. NCC-based tracking can then be formulated as $F(\mathbf{R}, \mathbf{t}) = \int_\Omega \left\| 1 - C^2(\mathbf{u}, \mathbf{u}') \right\| d\mathbf{u}$.

4. Camera imaging process

The key observation we rely on in this paper is that the scene radiance is mostly constant and invariant to exposure time of a camera as long as the illumination is static. It is particularly true in indoor environments. Our idea is to replace $e(.)$ with a new error function dependent on scene radiance only. For a pinhole camera model, the scene radiance is equivalent to the camera irradiance since all incident light at a point on camera sensor is from the same 3-D point in the scene along a direction. We therefore denote both as R, from which we can define the exposure as [5] $X = R\Delta t$, where Δt is the exposure time. The mapping from exposure to image intensity I is then defined by [16]

$$I = f(X + n_s(X) + n_c) \tag{3}$$

where $f : \mathbb{R}^+ \rightarrow \mathbb{Z}^+$ is the camera response function (CRF), n_s is a noise component dependent on the radiance, and n_c is a constant noise term. We assume that the image intensity level I ranges from 0 to 255 and the noise statistics are assumed to be $E(n_s) = E(n_c) = 0$, $Var(n_s) = R\Delta t\sigma_s^2$ and $Var(n_c) = \sigma_c^2$ [16, 7].

Our intention is to estimate the CRF in order to obtain the exposure values for given intensities. However, previous works [16, 7] have shown that the CRF can be highly variable, especially when the exposure is very high or very low. This can be seen in the bottom row of Fig. 3, which shows the standard deviation of the variation in intensity in each colour channel for different exposure times. One way of characterising this is to derive a noise level function [16, 7] which relates an input intensity level to the expected noise level. We make use of such a function here to weight the associated pixel data in each of the 3 colour channels, effectively acting as a pixel confidence function (PCF), which we normalise to range from 0 to 1 by dividing by the maximum standard deviation (m) over all 3 channels. Specifically, we define the PCF for a channel as a function of the intensity I and given by [16, 7]

$$p(I) = \frac{1}{m} \left. \frac{\partial f(x)}{\partial x} \right|_{x=f^{-1}(I)} \sqrt{f^{-1}(I)\sigma_s^2 + \sigma_c^2} \tag{4}$$

where $p : \mathbb{Z}_+ \rightarrow (0, 1)$ and $f^{-1}(I)$, the inverse camera response function. Examples of confidence maps obtained from the PCFs for each channel are shown in the right hand column of Fig. 5. Each pixel location in each channel of the confidence map stores a reliability measurement: dark areas show the low probability pixels which usually occur around exposed or under exposed parts of the image. We also experimented with several variants of the PCF defined in (4), namely $p_0(I) = 1$, $p_1(I) = \sqrt{p(I)}$, $p_2(I) = p(I)$, and $p_3(I) = p(I)^2$. Their effects will be discussed in section 5.

The CRF and the noise level function depend on the type of camera sensor used and these can be estimated using a number of different techniques [4, 14, 13]. In our case, we estimate the CRF by putting the RGB-D sensor at fixed position and capturing a sequence of images at different exposure times [4] and the noise level function and PCF are then estimated using the technique described in [7]. Examples of estimated CRFs, their derivatives, PCFs and plots of the standard deviation of the intensity variation for different exposure times (used to compute the optimal values of σ_s and σ_c in (4), see [7]) for the 3 colour channels is shown in Fig. 3.

5. Normalized radiance/exposure

We now show that normalized exposure is equivalent to normalized scene radiance and therefore it is perfect for formulating an error function independent of exposure and relying on scene radiance only. The normalization of the exposure map in a patch of neighbourhood N, centred at pixel

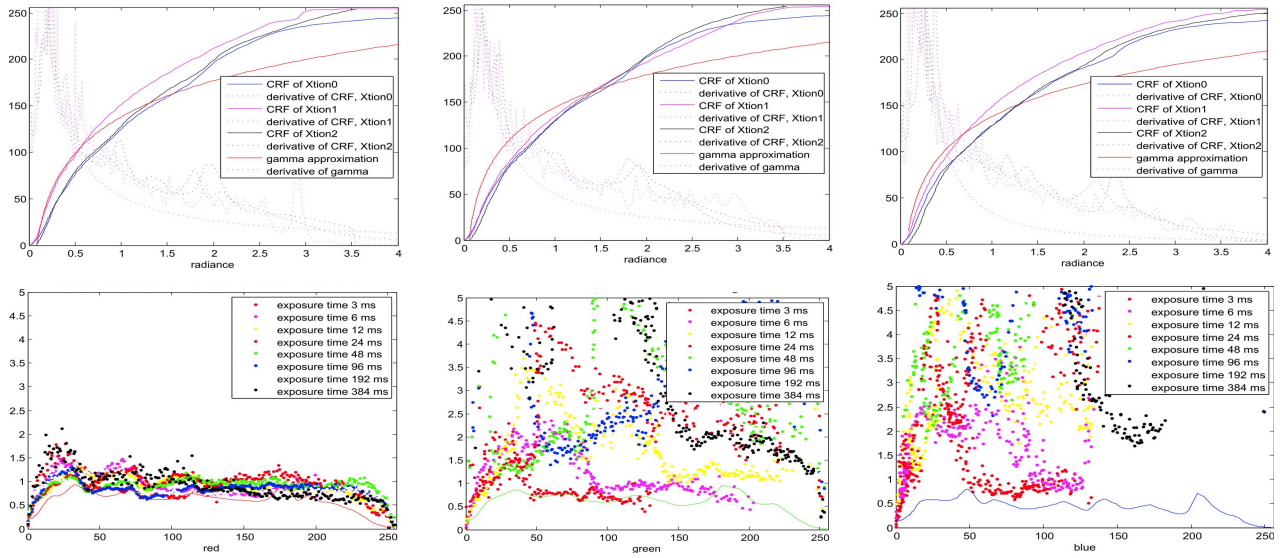

Figure 3: The estimated CRFs and PCFs for each colour channel (R,G and B, from left to right) for 3 Xtion sensors. (top) CRF and its derivative plus the gamma function approximation in red; note the large error in the latter when the radiance is high. (bottom) PCF for each channel (solid line) and the standard deviation of the intensity variation for each intensity level for different exposure times.

location $\mathbf{u}$, is formulated as

$$\bar{X}_N(\mathbf{u}) = \frac{X_N(\mathbf{u}) - E(X_N)}{\sqrt{Var(X_N)}} = \frac{R_N(\mathbf{u})\Delta t - E(R_N \Delta t)}{\sqrt{Var(R_N \Delta t)}} \tag{5}$$

$$= \frac{R(\mathbf{u}) - E(R_N)}{\sqrt{Var(R_N)}} \tag{6}$$

where $X_N : \Omega_N \rightarrow \mathbb{R}_+$ and $R_N : \Omega_N \rightarrow \mathbb{R}_+$ denote the exposure map and the radiance map in N, respectively. From the above equation, it can be seen that $\bar{X}_N(\mathbf{u})$ is independent of exposure Δt. This value is also invariant to viewpoint due to the fact that the radiance distribution in the local region corresponding to N is roughly constant irrespective of viewing position, as long as the surface is Lambertian and the illumination is constant. Fig. 5 shows the mean, standard deviation, normalized exposure and pixel confidence map of two consecutive frames captured at different exposure times. Note that the normalized exposure maps extracted from frames captured at different exposures are strikingly similar while the mean and standard deviation maps are smooth and blurry which indicates good resistance to viewpoint changes. Therefore, we define a new error function as follows

$$e'(\mathbf{u}, \mathbf{u}') = (\bar{X}_r(\mathbf{u}) - \bar{X}_l(\mathbf{u}'))p(I_l(\mathbf{u}')) \tag{7}$$

where $p(I_l(\mathbf{u}'))$ serves as a dynamic weight to balance the noise introduced during image formation such that less reliable pixels will be assigned with a smaller weight. $p(.)$ can be chosen from the family of PCFs we defined above, i.e. $p(.) \in \{p_0(.), p_1(.), p_2(.), p_3(.)\}$.

The error functions obtained when aligning two flicker affected frames along the horizontal axis using NCC, raw intensity, radiance with exposure compensated and normalized exposure are shown in Fig. 4, where the true alignment is at 0. Note that the error function using normalized exposure and weighed by the square root PCF $p_1(.)$ is the most ideal for optimization.

The camera poses can then be obtained by optimizing the error functions using the forward compositional approach described in [25].

5.1. Exposure time compensation

When the camera pose is estimated, the exposure time can then be compensated using the following equation:

$$\Delta t = \frac{1}{|\Omega|} \int_{\Omega} p_l(\mathbf{u}) \frac{X_r(\mathbf{u})}{X_l(\mathbf{u}')} d\mathbf{u} \tag{8}$$

where $p_l(.)$ is the confidence of a pixel in the live frame. After Δt is estimated, then the exposure map for the live frame can then be scaled by Δt in order that it is compatible in terms exposure with the reference frame, enabling fusion onto the scene reconstruction as described next.

6. Radiance Fusion

The compensated exposure map $X_l \Delta t$ is aligned with X_r, which is the synthetic radiance map proportional to

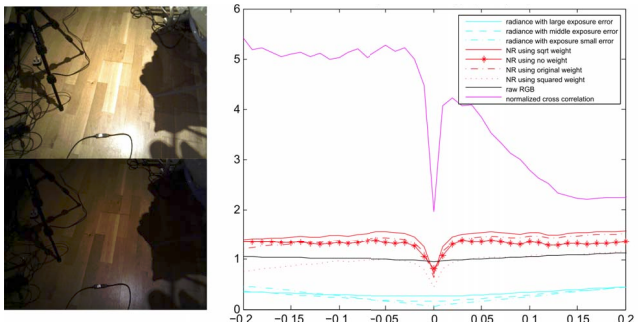

Figure 4: Comparison of error functions obtained when aligning two flicker affected frames in the horizontal direction, where 0 offset is the true alignment. The functions obtained using various forms of normalized exposure (red) are the most ideal for optimisation, with a relatively deep single global minimum. The NCC based error function [22] is also strongly convex but has many local minima. Tracking using exposure time compensated [19] is slightly better than tracking raw RGB but its global minimum is shallow even when the exposure compensation has small error.

scene radiance up to a scalar. $X_l \Delta t$ is then fused into the global volume using a fast parallel approach similar to [25]. The volumetric data structure stores not only the truncated signed distance function (TSDF) and its weights as KinectFusion does [21], but also stores the radiance, normalized radiance and confidence values for each of the 3 channels. The normalized exposure is also fused into the global volume so that synthetic normalized radiance map can be efficiently extracted using ray casting. Note that the radiance weight is different from TSDF weights. The fusion of radiance with depth for each voxel is given by the following equations:

$$F = \frac{w_F * F + w'_F * F'}{w_F + w'_F} \tag{9}$$

$$R = \frac{w_R * R + w'_X * R'}{w_R + w'_X} \tag{10}$$

$$\bar{R} = \frac{w_R * \bar{R} + w'_X * \bar{X}'}{w_R + w'_X} \tag{11}$$

$$w_F = w_F + w'_F \qquad w_R = w_R + w'_X \tag{12}$$

where F and R are the TSDF values and the radiance in the global volume, respectively, and F' and $R' = X_l \Delta t$ are those from the live frame. Similarly, w_F and w_R are the global weights and w'_F and w'_X are the weights from live frame. The weights are defined as follows. $w_F = |\mathbf{n}^T \mathbf{v}|$ is the absolute cosine distance between the surface normal $\mathbf{n}$ and the viewing direction $\mathbf{v}$ at the live pixel location, where $\mathbf{n}, \mathbf{v}$ are unit vectors. It down weighs the TSDF values captured at high angle between the normal and viewing direction. We found that using view dependent weights enables the map can capture more details. $w_R = \frac{p_r + p_g + p_b}{3}$, where

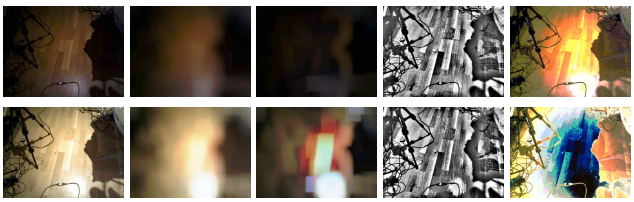

Figure 5: Exposure normalization. From left to right, figures correspond to raw RGB, mean, standard deviation, normalized radiance, and confidence map. The 1^{st} and 2^{nd} row correspond to 2 consecutive frames when flickering happens. Although the image brightness changes significantly, the normalized radiance map is similar. The mean, standard deviation maps and normalized radiance are tone mapped from the HDR domain.

p_r, p_g and p_b are the confidence values of 3 colour channels, respectively. In experiments, we find that storing individual confidence of 3 colour channel is the global volume is unnecessary and may introduce color distortion as well. In addition, to ensure the quality of radiance, only the pixels whose maximum confidence are above a threshold τ_0 and the angle between surface normal and viewing direction is above threshold τ_1 are allowed to be fused into the volume.

7. Experiments

In all experiments, we used 3 Xtion Pro sensors whose exposure times could be manually set and the estimated CRFs are shown in Fig. 3. The intrinsic parameters for the cameras were obtained from the sensor driver library OpenNI. We implemented HDRFusion based on the framework described in [15] and tested it on two commodity workstations, PC0 equiped with NVIDIA GTX 680 and PC1 NVIDIA GTX Titan Black GPU. Both PC are hosted by an i7 quad-core CPU. The volume resolution are set as 256^3 and 480^3 for PC0 and PC1 respectively with volume size ranges from 2^3 to 3^3m according to the size of the scene. The size of the normalisation patches was set to 40 pixels. The normalisation is implemented using integral images for efficiency. Frame resolution was set as QVGA for PC0 and VGA for PC1. Both of them operate at about 10Hz.

7.1. Tracking with AE enabled

We first used the synthetic dataset ICL to evaluate our approach [7] which provides high quality CG HDR frames and ground truth camera poses. First, photo realistic LDR RGB frames are simulated using real CRF and noise level function of a randomly chosen Xtion sensor. We generated two sequences of video to simulate video flickering and smooth AE behaviour. The flickering sequence is sim-

ulated by randomly choosing exposure time from the set $3, 6, 12, ..., 96$ (ms). The second sequence is generated using the equation $\Delta t = C/L$, where $C = 4.8 \times 10^5$ and L is the average HDR intensity of the 10 by 10 patch in the center of the original HDR frames. The exposure changes in the second sequence are smooth. Kinect like depth noise is also added using the approach from [8]. Typical flickering pairs are illustrated in Fig. 4. The tracking approach using normalized intensity, NCC object function based on [22] and approach similar to the tracking of [19] are used as the baseline approaches. For fairness, the ICP-based frame-to-model tracking are disabled for all above methods. The tracking accuracy in terms of rotational and translational error are plotted in Fig. 6.

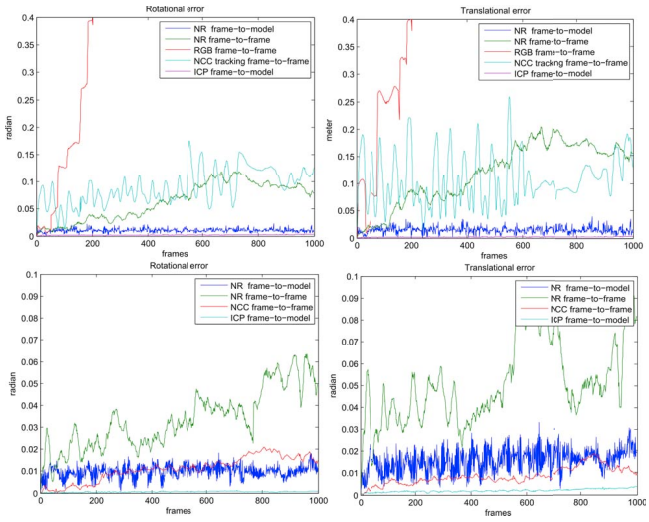

Figure 6: Tracking synthetic sequence. The top two figures are the rotational and translational error using synthetic flickering sequence. The bottom two figures are using synthetic smooth AE sequence. In the flickering sequence, we can see that raw RGB based tracking quickly get lost, while the NCC and the proposed frame-to-frame tracking (NR) and frame-to-model tracking using normalized radiance remains working well. The tracking NR in frame-to-model mode gives the best performance in the flickering sequence. Due to the rich geometric variance, the ICP-based frame-to-model tracking give the best results in this sequence. In smooth sequences, the NCC and ICP performs better but the proposed tracking remain working reasonably accurate. The frame-to-model tracking is within 3cm in the 1000 frames testing sequence. In the experiment, we observe that the NCC based method provides some resilience to AE but it is extremely sensitive to color consistent area where the std of the patch is 0 as shown in the top row. Instead, our approach allows patch size of 40 pixel so that the variance are unlikely to be zero.

We also performed a qualitative comparison using real data between the proposed tracking and tracking using the approach described in [25], which is representative of dense direct SLAM [20, 26, 12]. Two sequences of RGB-D video with flickering were captured. In these sequences, the sensor is overlooking a planar scene, a floor and a white board, respectively. The experiment is designed to test a system when the geometry is at minimum. Note that to the best knowledge of the authors, currently there is no SLAM system capable of tackling the situation since there is minimum amount of geometric features for ICP and the AE effects will impact on dense direct RGB tracking. We found that the method described in [25] exhibited significant drift and resulted in a blurry textured map. In contrast, HDR-Fusion dealt with it effectively and the reconstructed floor and white board are shown in Fig. 7. The reconstructed 3-D models are accurate with sharp HDR textures.

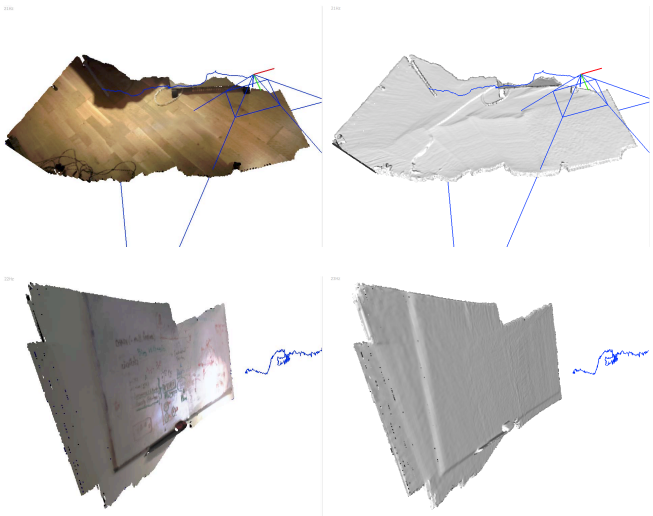

Figure 7: Tracking under flickering using real data. The blue curves are the camera trajectories. The frustums in the top figure show the camera pose. Qualitative comparison between our method and RGB-D SLAMs is available in the accompanying video showing that RGB-D SLAM drifts badly and results in a distorted map with blurry texture.

7.2. HDR Radiance map

To demonstrate the effectiveness of the HDR texture capture, we tested our system on three scenes, namely 'Bear', 'Desk' and 'Sofa', frames from which are shown in Fig. 8. The bear scene is illuminated by indirect sun light, the desk scene by fluorescent light and the sofa scene by both fluorescent lighting and a Dedolight-400D metal halide lamp. In Fig. 8, HDR scene textures are compared with the ground truth. The ground truth is captured using a Canon 5D MarkII SLR camera. Three exposure LDR images with a 2-fstop interval of the scene were captured and then merged to

form an HDR image. Both are rendered using tone maping operator(TMO) [17]. In contrast, standard RGB-D fusion fails to capture the details of scene textures such as under the desk and sofa, and leaves artefacts around overlapping frames. Note that the camera poses for RGB-D fusion are provided by the proposed "HDR-D" tracking, because simple RGB-D tracking fails during the mapping stage due to AE.

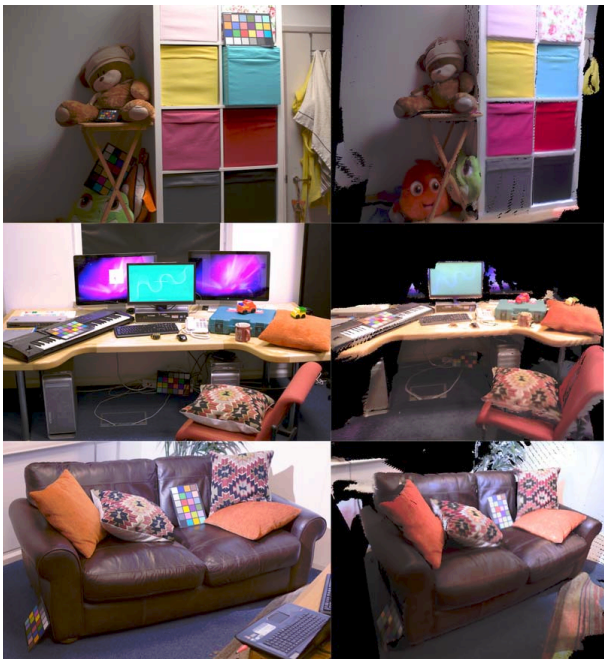

Figure 8: The left are the ground truth HDR radiance and HDR radiance generated using HDRFusion are rendered using [17] where the colour saturation is set as 1. We can see that estimated HDR texture closely matches the HDR radiance captured using the high-end SLR camera.

8. Conclusions

In this paper, we propose a novel HDRFusion system capable of capturing high quality HDR scene texture using a low cost RGB-D sensor. Tracking normalized exposure allows the decoupling of the tracking from exposure compensation which improves the accuracy of both. Tracking normalized exposure was also shown to be robust to camera AE adjustment. The tracking is runing in frame-to-model mode which accumulates less drift. In future work, calibrating the CRF function online will be investigated as in some sensors the exposure time can not be changed. Another limitation of the system lies in its large memory footprint. Storing both the normalized radiance and radiance seems unnecessary. Reducing the size of memory cost by combining the both will also be investigated.

Figure 9: Desk. The LDR frames generated using [25] are shown in the first row and HDR frames produced by HDRFusion are shown in the second row. The HDR radiance is rendered using [17], where the colour saturation is set as 1.5.

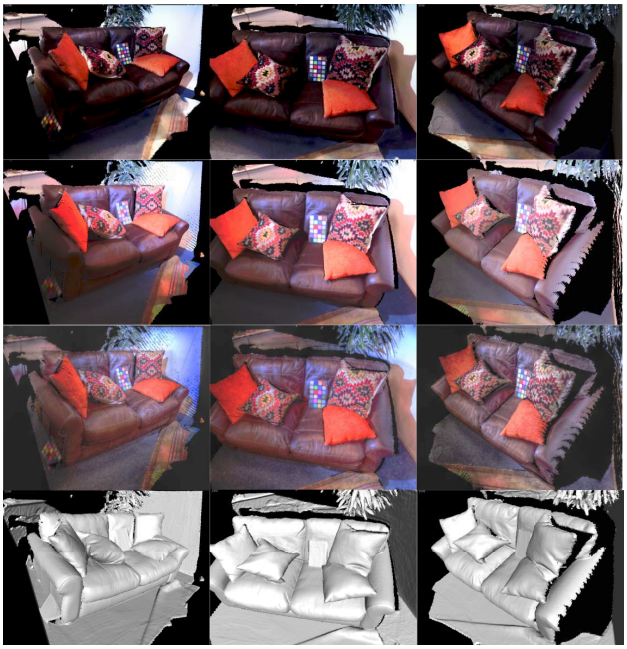

Figure 10: Sofa. The LDR frames generated using [25] are shown in the first row and HDR frames produced by HDRFusion are shown in the second row and third row. The second row is generated using [17] where the colour saturation is set as 1. Comparing with raw RGB fusion [25], the dynamic range of the radiance texture is much higher. The details in dark area are well preserved. The third row is generated using [18], where the colour saturation is set as 1.25. [18] visualizes the rich details captured by HDRFusion. The bottom row shows the recovered surface geometry.

References

[1] T. Aydin, N. Stefanoski, and S. Croci. Temporally coherent local tone mapping of HDR video. *ACM Trans. on Graphics (ToG)*, 33(6), 2014. 4321

[2] M. Batz, T. Richter, J. U. Garbas, A. Papst, J. Seiler, and A. Kaup. High dynamic range video reconstruction from a stereo camera setup. *Signal Processing: Image Communication*, 29(2):191–202, 2014. 4323

[3] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *Intl. Journal of Computer Vision (IJCV)*, 74(1):59–73, 2007. 4321

[4] P. E. Debevec and J. Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *ACM SIGGRAPH (SIGGRAPH)*, number August, pages 1–10, 1997. 4322, 4324

[5] P. E. Debevec and J. Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *ACM SIGGRAPH (SIGGRAPH)*, number August, 1997. 4324

[6] Z. Farbman and D. Lischinski. Tonal stabilization of video. *ACM Trans. on Graphics (ToG)*, 30(4):1, 2011. 4321

[7] A. Handa, R. A. Newcombe, A. Angeli, and A. J. Davison. Real-Time Camera Tracking : When is High Frame-Rate Best? In *European Conf. on Computer Vision (ECCV)*, 2012. 4324, 4326

[8] A. Handa, T. Whelan, J. Mcdonald, and A. J. Davison. A Benchmark for RGB-D Visual Odometry , 3D Reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014. 4327

[9] Y. S. Heo, K. M. Lee, and S. U. Lee. Robust Stereo matching using adaptive normalized cross-correlation. *IEEE Trans. Pattern Anal. Machine Intell. (PAMI)*, 33(4):807–822, 2011. 4323

[10] 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 *3DV*, pages 1–8. Ieee, jun 2013. 4321, 4323

[11] C. Kerl, J. Stuckler, and D. Cremers. Dense Continuous-Time Tracking and Mapping with Rolling Shutter RGB-D Cameras. In *Intl. Conf. on Computer Vision (ICCV)*, 2015. 4322, 4323

[12] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 3748–3754, 2013. 4321, 4323, 4327

[13] S. J. Kim, J. M. Frahm, and M. Pollefeys. Joint feature tracking and radiometric calibration from auto-exposure video. In *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2007. 4324

[14] S. J. Kim and M. Pollefeys. Radiometric Self-Alignment of Image Sequences. In *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 645–651, 2004. 4324

[15] S. Li and A. Calway. RGBD Relocalisation Using Pairwise Geometry and Concise Key Point Sets. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015. 4326

[16] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman. Automatic estimation and removal of noise from a single image. *IEEE Trans. Pattern Anal. Machine Intell. (PAMI)*, 30(2):299–314, 2008. 4324

[17] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Trans. on Graphics (ToG)*, 27(3):1, 2008. 4322, 4328

[18] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception*, pages 87 – 94, 2006. 4328

[19] M. Meilland, C. Barat, and A. Comport. 3D High Dynamic Range dense visual SLAM and its application to real-time object re-lighting. In *IEEE/ACM Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 143–152, 2013. 4322, 4323, 4326, 4327

[20] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM : Dense Tracking and Mapping in Real-Time. In *Intl. Conf. on Computer Vision (ICCV)*, 2011. 4321, 4327

[21] R. A. Newcombe, D. Molyneaux, D. Kim, A. J. Davison, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion : Real-Time Dense Surface Mapping and Tracking. In *IEEE/ACM Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 4321, 4323, 4326

[22] G. G. Scandaroli, M. Meilland, and R. Richa. Improving NCC-based direct visual tracking. In *European Conf. on Computer Vision (ECCV)*, pages 442–455, 2012. 4323, 4326, 4327

[23] J. Serafin and G. Grisetti. NICP : Dense Normal Based Point Cloud Registration. In *Intl. Conf. on Intelligent Robot Systems (IROS)*, page 8, 2015. 4323

[24] T. Whelan, H. Johannsson, M. Kaess, J. J. Leonard, and J. McDonald. Robust real-time visual odometry for dense RGB-D mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 5724–5731, 2013. 4323

[25] 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. *Intl. Journal on Robotics Research (IJRR)*, 34(4-5):598–626, 2015. 4321, 4322, 4323, 4325, 4326, 4327, 4328

[26] T. Whelan, S. Leutenegger, R. F. Salas-moreno, B. Glocker, and A. J. Davison. ElasticFusion : Dense SLAM Without A Pose Graph. *Robotics: Science and Systems (RSS)*, 2015. 4321, 4322, 4327