Enhanced Depth Estimation using a Combination of Structured Light Sensing and Stereo Reconstruction

Andreas Wittmann[1], Anas Al-Nuaimi[1], Eckehard Steinbach[1] and Georg Schroth[2]
[1]*Chair of Media Technology, Technische Universitt Mnchen, Arcisstrae 21, Munich, Germany*
[2]*NavVis GmbH, Blutenburgstrae 18, Munich, Germany*

Keywords: Stereo Vision, Google Project Tango, Structured Light, 3D Scanning.

Abstract: We present a novel approach for depth sensing that combines structured light scanning and stereo reconstruction. High-resolution disparity maps are derived in an iterative upsampling process that jointly optimizes measurements from *graph cuts*-based *stereo* reconstruction and *structured light* sensing using an accelerated α-*expansion* algorithm. Different from previously proposed fusion approaches, the disparity estimation is initialized using the low-resolution structured light prior. This results in a dense disparity map that can be computed very efficiently and which serves as an improved prior for subsequent iterations at higher resolutions. The advantages of the proposed fusion approach over the sole use of stereo are threefold. First, for pixels that exhibit prior knowledge from structured lighting, a reduction of the disparity search range to the uncertainty interval of the prior allows for a significant reduction of ambiguities. Second, the resulting limited search range greatly reduces the runtime of the algorithm. Third, the structured light prior enables a dynamic tuning of the smoothness constraint to allow for a better depth estimation for inclined surfaces.

1 INTRODUCTION

In recent years research and applications in 3D vision have been experiencing strong growth driven by the increasing availability of low-cost hardware that can sense the environment in 3D. Critical to these advancements was the release of the Microsoft Kinect in 2010 which is the first consumer device that can reliably sense its immediate vicinity in 3D and in real time using the *structured light* (SL) sensing technique.

Other recent trends, including the proliferation of smart portable devices, which are equipped with powerful processors and a plethora of sensors, have called upon the deployment of 3D sensors to enable new applications including Augmented Reality. Indeed, Google launched the first smartphone with an active 3D sensor, similar to that of the Kinect, named *Project Tango* device (Piszczor and Yang, 2014). Other companies, such as Occipital (Occipital, 2015), offer smartphone attachable 3D sensors which use the same sensing principle.

The deployment of such 3D sensors in portable devices is significant but their impact may be limited by the characteristic limitations inherent to the used sensing technique. Most notably, SL-based 3D sensing has a relatively limited range of a few meters. Critically, 3D sensing fails on glossy surfaces, edges, fine structures and elements that are illuminated by bright light owing to the infrared-based depth perception. The latter also severely degrades the usability outdoors.

Unlike depth from SL, 3D reconstruction from stereo can provide 3D data for larger distances, fine structures as well as outdoor scenes. Contrary to SL techniques, it benefits from bright illumination and sharp object boundaries. However, local stereo matching requires well-textured surfaces, something SL-based 3D scanning does not depend on. Also, global stereo reconstruction algorithms are often slow and depend on the structure of the underlying scene, which forbids its sole use for applications that require a consistent performance.

The complimentary nature of depth from stereo and depth from SL, as further demonstrated by a real example in Section 3, motivated us to develop a joint depth estimation algorithm that uses image and 3D data acquired on a smartphone prototype to generate enhanced depth images. It builds upon the strengths of both sensing techniques while compensating for their shortcomings. For this purpose, we utilized the Google Project Tango device, henceforth also termed "smartphone".

Wittmann, A., Al-Nuaimi, A., Steinbach, E. and Schroth, G.
Enhanced Depth Estimation using a Combination of Structured Light Sensing and Stereo Reconstruction.
DOI: 10.5220/0005724605100521
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 3: VISAPP*, pages 512-523
ISBN: 978-989-758-175-5

To achieve our goal we combine a graph cuts-based state-of-the-art stereo algorithm (Kolmogorov et al., 2014) with the depth maps captured by the SL sensor of the smartphone. To that end we redesigned a global energy function typically used in stereo imaging to incorporate the SL sensor data while considering its error characteristics as explained in Section 4.1. Beyond solving for depth values using a joint energy function, the algorithm utilizes the smartphones's active range measurements to limit the disparity label space for the stereo algorithm, resulting in a substantially faster convergence (see Section 5). In Section 4.2 we explain how this technique, which exploits depth priors from the smartphone, is applied in an iterative process involving sequential upsampling and stereo reconstruction to produce higher resolution depth maps. The results in Section 5 demonstrate the gains achievable with our proposed fusion scheme. Prior to the detailed explanations of the contributions, we give background information in 3D sensing and survey related work on stereo-range data fusion in Section 2. Also, the Project Tango device is briefly presented in the same section.

2 RELATED WORK & BACKGROUND

3D sensors can be broadly categorized into passive and active sensors. The former rely on the ambient lighting of the environment while the latter project (visible or invisible) patterns or beams of light to sense the environemt. Passive and active 3D sensors are introduced in Section 2.1 and Section 2.2, respectively. In Section 2.4 the Project Tango device, which we simultaneously use for active and passive 3D sensing, is briefly introduced. In Section 2.3 we survey related work on enhanced depth estimation with active and passive 3D sensors.

2.1 Passive 3D Sensing

Standard passive 3D sensing involves capturing the scene from multiple perspectives and using visual correspondences to infer the 3D shape. There exists a variety of approaches including *multi-view* stereo (MVS) and *structure-from-motion* (SfM), that take two or more images to reconstruct a 3D scene. A special case is *two-view* stereo that only considers a pair of images to derive the depth in a scene.

The aforementioned approaches either utilize several cameras, or as in case of SfM, a single moving camera. Seitz et al. (Seitz et al., 2006) compare different multi-view approaches which are typically slow and are generally applied in offline processing. However, real-time capable SfM algorithms such as Dense Tracking and Mapping (DTAM) (Newcombe et al., 2011) and Large-Scale Direct Monocular SLAM (LSD-SLAM) (Engel et al., 2014), which both employ GPU computing to significantly speed up processing, exist. Nevertheless, SfM-based multi-view methods suffer from drift and an inherent scale ambiguity that does not allow us to restore the true scale of a scene. Multi-view stereo reconstruction typically allows for a true scale representation of a scene by utilizing a setup that involves multiple cameras with rigid transformations among the individual camera frames. Obviously, reconstruction via MVS requires extensive efforts and is therefore not suited for consumer use. The accuracy for multi-view approaches is not clearly related to the number of input images. Furthermore, SfM reconstruction approaches are only suitable for static scenes.

For two-view stereo matching approaches, Scharstein and Szeliski (Scharstein and Szeliski, 2002) present a large number of techniques including a performance assessment. Their datasets are considered standard for performance evaluations in stereo vision and their online database lists the latest algorithms in this domain (Scharstein and Szeliski, 2015).

Advanced stereo reconstruction (henceforth always referring to two-view stereo except otherwise noted) algorithms consider a global energy formulation and find the set of disparities that minimizes the energy function. The minimization is usually performed with inference algorithms such as belief-propagation (Felzenszwalb and Huttenlocher, 2006) or graph cuts (Boykov et al., 2001). Tappen and Freeman (Tappen and Freeman, 2003) compare the two approaches. They conclude that belief-propagation is in general faster than graph cuts but the results are less smooth. Although Tappen and Freeman (Tappen and Freeman, 2003) found graph cuts and belief propagation to perform similar on their dataset, we observed a significantly better performance of graph cuts for our real world datasets, which is why we decided to utilize the state-of-the-art graph cuts stereo algorithm by Kolmogorov and Zabih (Kolmogorov et al., 2014) in our work.

2.2 Active 3D Sensing

Two major types of active 3D sensors exist: Time-of-Flight (ToF) and Structured Light (SL).

ToF sensors emit infrared-light (IR) and capture its reflection. The distance assigned to a pixel is inferred from the time delay between emission and reception of the IR-signal. ToF sensors run in real time and provide good results even on textureless surfaces. Yet, the sensors suffer from limited resolution as well as various error sources such as noise, multi path, ”flying pixels” and are susceptible to background illumination (Foix et al., 2011). Due to size and power limitations it is difficult to deploy ToF on smartphones.

SL sensors work by projecting a light pattern onto the scene and then capturing it with a camera. The distortion is used to infer the 3D geometry as it is a function of the 3D shape (Scharstein and Szeliski, 2003a). The light pattern acts as texture and hence texture-less scenes can also be sensed. The Microsoft Kinect performs SL sensing on a dedicated chip achieving real-time 3D imaging. Since the projected pattern is made up of IR light, it does not work in the presence of sunlight. The projected pattern is relatively weak due to power limitations thus limiting the sensing range to < 10m. Since SL sensing essentially performs stereo vision, the sensing accuracy is a function of the IR camera resolution and the depth of the scene. In our paper we show how to properly account for the decreasing accuracy with increasing depth in the fusion algorithm (Section 4.1.1).

2.3 Enhanced Depth Estimation through Fusion

In (Wei-Chen Chiu and Fritz, 2011), a promising approach that utilizes cross modal stereo reconstruction, known as IR-image RGB registration, is proposed to find correspondences between the IR and RGB images of the Kinect. By combining the RGB channels with appropriate weightings, the image response of the IR-sensor is resembled, which allows for depth estimation for reflective and transparent objects via stereo reconstruction. Fusing the stereo reconstruction results with the structured light measurements extends the abilities of the Kinect without the need for additional hardware. The stereo reconstruction approach proposed in our work does not require an optimization as it is proposed by Wei-Chen et al. By utilizing the same camera for both stereo images we avoid a degradation of stereo resulting from the use of two different cameras.

(Li et al., 2011), (Scharstein and Szeliski, 2003b) and (Choi et al., 2012) achieve a highly accurate fusion of structured light scans and stereo reconstruction by recording a projected pattern with a set of stereo RGB cameras. The structured light sensor used in our work provides reliable depth measurements out of the box. Moreover, it records RGB and depth images from SL with the same sensor chip and therefore achieves a highly precise alignment as well.

(Gandhi et al., 2012) generate a high-resolution depth map by using ToF measurements as a low-resolution prior that they project into the high-resolution stereo image pair as an initial set of correspondences. Utilizing a Bayesian model allows propagating the depth prior to generate high-resolution depth images.

In (Somanath et al., 2013), high-resolution stereo images are fused with the depth measurements from the Microsoft Kinect. The authors use a graph cuts-based stereo approach for the fusion. Therefore, the influence of the individual sensors is considered with a confidence map, which is determined by the stereo images as well as the Kinect measurements. For the fusion, Somanath et al. project the SL measurements into the high-resolution stereo images, which results in a reduced confidence of the Kinect data. Instead, our setup allows capturing the RGB as well as the depth images from SL with a single camera and avoids the resulting alignment errors in the fusion. Therefore, our confidence consideration is not affected by alignment and projection issues and is solely based on the error characteristics of the smartphones’s SL sensor.

The aforementioned stereo-range superresolution approaches (Li et al., 2011), (Gandhi et al., 2012) and (Somanath et al., 2013) perform a fusion of passive stereo vision and active depth measurements by projecting a low resolution prior into the stereo images to perform a fusion at high-resolution. In contrast, we propose an iterative fusion approach that is initialized at the low-resolution of the SL depth images. This approach results in a tremendous acceleration of the correspondence computation, since both, the number of pixels that have to be assigned a disparity and the considered label space, are much smaller at low-resolution. Iteratively launching the algorithm with the disparities found in the stereo-SL depth fusion allows us to retrieve a superresolution depth image in much shorter time than the previously mentioned approaches.

2.4 Google Project Tango

Figure 1 depicts the Google Project Tango device that we use in our experiments to perform a fusion of SL and stereo depth maps. The Project Tango device uses essentially the same sensing technique as the Kinect. In fact, it is equipped with a Primesense chip for hardware-based disparity computation (Goldberg et al., 2014), just as is the Kinect. Contrary to the Kinect, however, the Project Tango device uses the same camera (Identified as “4MP” in Figure 1)

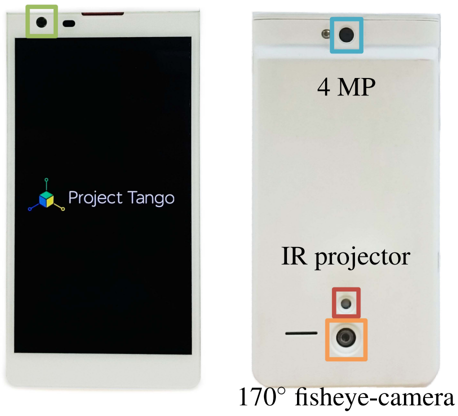


Figure 1: Front and back view. The image sensors are highlighted. According to (Goldberg et al., 2014), the 4MP camera is used to capture RGB images as well as the IR pattern, projected by the IR projector, for depth estimation.

to capture RGB images as well as the projected IR pattern for depth estimation. This is important since both, the RGB and the corresponding SL depth images are readily aligned.

In our experimental setup, we achieve a horizontal shift of b_{stereo} between the stereo image pairs, as shown in Figure 2, by moving the Project Tango device on a slider. Strictly speaking, this is SfM (see Section 2.1). However, due to the controlled movement, we are simulating a stereo camera pair.

The stereo depth image is computed w.r.t. the initial position and is fused with the SL depth image captured at the same position.

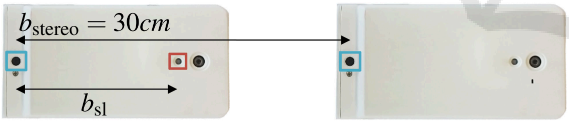


Figure 2: Recording depth images with the Project Tango device: The smartphone is moved by b_{stereo} to capture a stereo image pair. Simultaneously, the internal SL chip computes a depth image using the projected IR pattern which has an effective baseline of b_{sl}, which we determined in a calibration process.

3 PROBLEM FORMULATION

Here we illustrate the problem of depth estimation using solely SL sensing or only stereo reconstruction more formally and demonstrate their complementary properties to motivate our proposed solution for depth map fusion using stereo and SL vision. For that, we use the setup introduced in Section 2.4.

Figure 3 reveals the strengths and weaknesses of the Project Tango device's SL depth sensing abilities. The figure shows an RGB image along with a heatmap that encodes depth values up to a distance of 10 meters. It can be seen that the sensor performs well on nearby and smooth elements, but fails on illuminated or glossy surfaces. We note that although the shown depth map has been generated with the Project Tango device, it is exemplary for common SL 3D sensors.

Similar to Figure 3, Figure 4 shows the depth map for the same scene, however this time computed using standard stereo vision. In this case, the depth estimation provides good results for elements that are affected by the projector image, as well as the glossy poster in the scene. As expected, it fails on textureless elements such as the wall or the nearby chair.

The complementary properties of the two depth maps that correspond to Figures 3 and 4 are shown in Figure 5. From a naive fusion, where the depth estimate for a pixel is adopted from SL whenever available and filled with a value from the depth map from stereo otherwise, it can be seen that a depth map with a substantially reduced amount of "holes" is produced. The figure clearly shows that a substantial amount of depth data is contributed by either depth map, hinting only the combination of both can lead to significant improvement of the depth maps. Hence, the goal of this paper is to develop a fusion approach for joint stereo & structured light depth map estimation. The complementary character of both techniques inspired us to think of a more sophisticated fusion than a simple naive fusion, since not only lots of information gained from stereo vision is wasted, but also geometric priors, typically incorporated in stereo reconstruction through regularization, are enabled.

4 PROPOSED FUSION SCHEME

In this section we propose a fusion scheme that exploits the complementary properties of stereo reconstruction and structured light sensing beyond the simple fusion approach presented in Section 3. To that end, a global energy formulation that models the data cost as well as the cost for the spatial configuration of the disparities, as is typically the case in state-of-the-art stereo vision, is considered. The data term of the energy function is extended to incorporate prior knowledge derived from the SL data as explained in Section 4.1.1. This prior knowledge is also incorporated in the smoothness term, as explained in Section 4.1.2, to compensate for the shortcomings of typically employed smoothness constraints in state-of-the-art stereo reconstruction. The resulting fusion scheme is applied in an iterative process involving sequen-

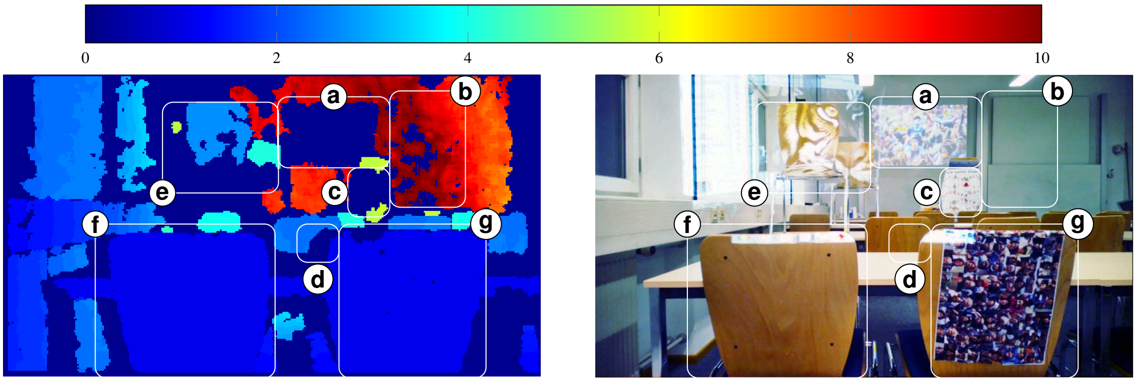


Figure 3: Sample scene captured by Tango's SL sensor. Depth values are color coded. 0 depth (dark blue) represents *missing* depth data. Key regions are highlighted (ⓐ-ⓓ), where ⓐ does not include any depth information due to the projector image that is confusing the IR-pattern; ⓑ is punctured with multiple holes indicating the maximum range of the SL projector has been reached (exceeded). Also the depth values are strongly varying despite the planar shape which is due to the high depth uncertainty at large distances; ⓒ highlights a glossy poster which makes the IR-pattern unusable for the IR-camera; ⓓ exhibits missing depth values next to the edge of the chair due to IR projection pattern occlusion; ⓔ displays the partly reconstruction of the backs of two chairs illuminated by the projector image; ⓕ,ⓖ show the well estimated depth values for the backs of the two nearby chairs.

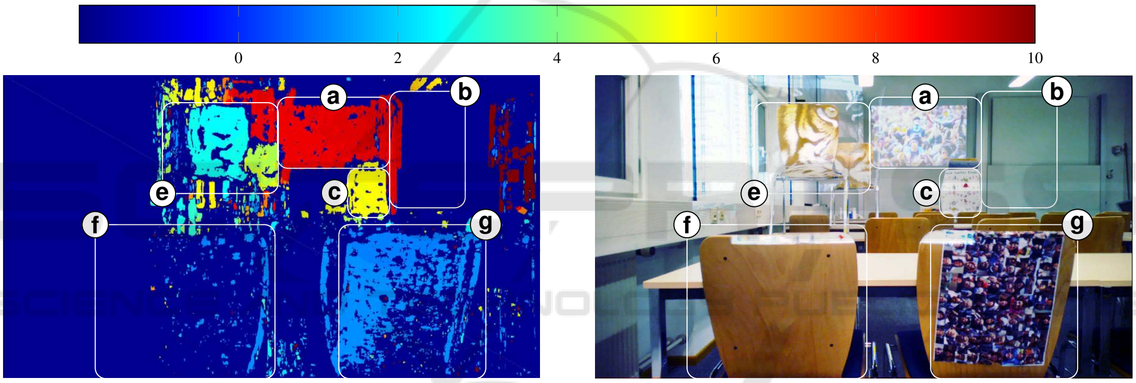


Figure 4: Depth map from stereo using block-matching. Key regions are highlighted by boxes (ⓐ-ⓑ), where ⓐ shows a richly textured area that allows a good reconstruction; ⓑ is poorly reconstructed as a result of missing texture in this area; ⓒ depicts a glossy poster that is well reconstructed; ⓔ provides rich texture for a good reconstruction ; ⓕ only allows a fragmented reconstruction resulting from missing texture of the scene; ⓖ performs better than ⓕ despite having the same shape owing to the poster.

tial upsampling and joint reconstruction to produce higher resolution depth maps as explained in Section 4.2.

Before diving into the details we want to note that we interchangeably use the terms disparity and depth since they are related directly to one another. As will be apparent, however, all mathematical formulas and the actual implementation are based on disparities and disparity maps.

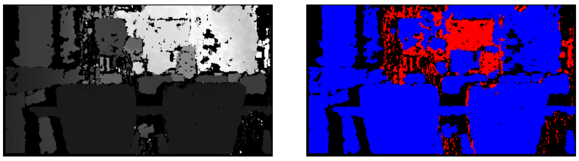
4.1 Joint Optimization using Stereo and Strucured Light

We propose a fusion of structured light and stereo depth maps based on the built-in sensors of the smartphone shown in Figure 2. The energy formulation used to find the optimal set of disparities is inspired by (Kolmogorov et al., 2014). The disparity $f_p \in \{L_1, L_2, \ldots, L_n\}$, for every pixel $p \in P$, is found such that the resulting configuration of the disparities f minimizes the energy

$$E(f) = E_{\text{data}}(f) + E_{\text{smoothness}}(f) + E_{\text{A}}(f), \qquad (1)$$

where $E_{\text{data}}(f)$ represents the data cost that also incorporates the structured light measurements. $E_{\text{smoothness}}(f)$ considers the spatial configuration of the disparities and $E_A(f)$ aggregates the occlusion and the uniqueness term which are explained in (Kolmogorov et al., 2014).

We note here that whenever using the word "orig-



ⓐ Fused depth map

ⓑ Color coding of depth source

Figure 5: Naive fusion of depth maps from stereo and structured light. ⓐ shows the depth map obtainable with a naive fusion scheme adopting the SL value whenever available for a pixel and using the stereo value otherwise. ⓑ Red implies a depth value from stereo reconstruction was used as opposed to blue which encodes values from SL. Black means no depth value sensed using either technique. Both sensing techniques are seen to contribute significantly towards the final outcome.

inal" we refer to Kolmogorov and Zabih's algorithm in its original form and as implemented by them.

4.1.1 Data Cost

We propose to extend the data term introduced in Equation (1) which typically accounts for the stereo matching cost C_{BT} with an additional term to account for the structured light measurements as follows:

$$E_{\mathrm{D}}(f) = \sum_p C_{\mathrm{BT}}(p) + C_{\mathrm{SL}}(p), \tag{2}$$

where C_{SL} describes the cost contributed by the structured light sensor. We propose to model C_{SL} according to

$$C_{\mathrm{SL}}(p) = w_1\, c_{\mathrm{SL}} \left(1 - e^{\frac{(f_p^{\mathrm{SL}} - f_p)^2}{\sigma_{\mathrm{SL}}^2}}\right) = w_1\, c_{\mathrm{SL}}\, C'_{\mathrm{SL}}(p). \tag{3}$$

The weighting factor w_1 controls the influence of the cost term and c_{SL} is the maximal penalty that can be assigned. $C'_{\mathrm{SL}}(p)$ represents an inverse Gaussian function and is motivated by the assumption of a normally distributed structured light disparity measurement error. This cost term ensures that whenever a prior measurement f_p^{SL} is available and the considered disparity f_p deviates from it, a penatly is contributed. More specifically, the cost added by $C_{\mathrm{SL}}(p)$ depends on the *assigned* disparity f_p as well as the disparity measured by the SL sensor f_p^{SL}, weighted using the Gaussian function. We define the weighting factor w_1 according to

$$w_1 = \begin{cases} \rho & f_p \in [f_p^{\mathrm{SL}} - 3\sigma_{\mathrm{SL}}, f_p^{\mathrm{SL}} + 3\sigma_{\mathrm{SL}}] \\ 0 & \text{else,} \end{cases} \tag{4}$$

where ρ can be chosen from the interval $(0,1]$ according to the weight to be assigned for the prior. We are inspired for this particular design by the work of (Khoshelham and Elberink, 2012) in which they investigated the Kinect's sensing accuracy. Assuming a normally distributed disparity measurement of constant variance, they derived the depth measurement error. They concluded and experimentally verified that the depth measurement uncertainty has a standard deviation that quadratically increases with the sensed distance. Since the Project Tango device's SL sensor essentially works the same way, we use the same uncertainty model assuming the error in the disparity domain is distributed in a Gaussian fashion with a constant variance σ_{SL}.

For the standard data term $C_{\mathrm{BT}}(p)$ in stereo matching, we use the Birchfeld-Tomasi pixel dissimilarity measure (Birchfield and Tomasi, 1998) as follows:

$$C_{\mathrm{BT}}(p) = w_2 \min\left(T(I_l(p), I_r(q))^2,\ c_{ST}\right), \tag{5}$$

$T(\bullet,\bullet)$ computes the data cost according to the commonly used Birchfeld-Tomasi dissimilarity measure for the pixel at position p in the left image I_{l} and the pixel at position q in the right image I_{r}. c_{ST} trims the cost and the weighting factor is set to be $w_2 = 1 - w_1$. We obtained the best results for considering the structured light related term with a larger weighting than the Birchfeld-Tomasi term, resulting from the higher reliability of structured lighting. We empirically determined the best fusion results for $w_1 \in [0.6; 0.8]$, depending on the underlying scene.

This particular energy formulation allows for a significant acceleration of the computation of the disparities. Figure 6 illustrates how the speed-up is achieved. In the figure, the grid represents the pixel positions for which the disparities have to be estimated, which have to be evaluated for the whole label space $[L_{\min}; L_{\max}]$, if no prior information from the structured light depth map is available. Whenever prior information is available from the SL data, indicated by the red dots, the label search can be limited to the uncertainty interval of the structured light measurements. We explicitly limit the label space for f_p to a range of $f_p^{\mathrm{SL}} \pm 3\sigma_{\mathrm{SL}}$, since there is a 99% chance that the true disparity $\hat{f}_p$ is within that interval assuming normally distributed SL disparity measurements. This reduces the complexity in case of large search intervals and high-resolution stereo images significantly as shown in the results (see Section 5).

4.1.2 Smoothness Cost

The smoothness term is based on the assumption that neighboring pixels with similar intensity values should be assigned the same disparity. If this assumption is violated, a large penalty is added to the cost function in Equation (1). In case of a strong contrast

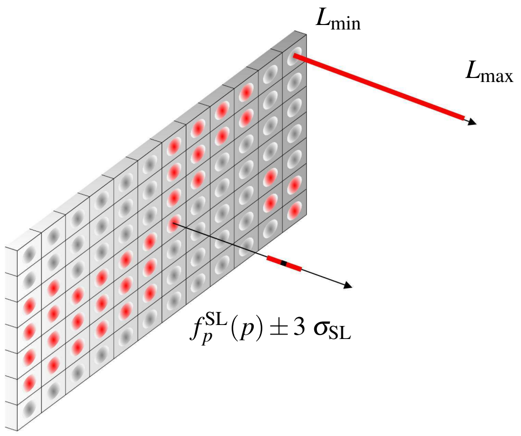


Figure 6: Disparity label search space with and without prior knowledge on the disparities.

among neighboring pixels, the penalty for changing neighboring disparities is smaller and equal disparities among neighboring pixels do not contribute any cost. In (Kolmogorov et al., 2014), the authors set the threshold for similar pixels according to

$$\max\left(|I_{\mathrm{l}}(p_1) - I_{\mathrm{l}}(p_2)|_\infty, |I_{\mathrm{r}}(q_1) - I_{\mathrm{r}}(q_2)|_\infty\right) < 8, \quad (6)$$

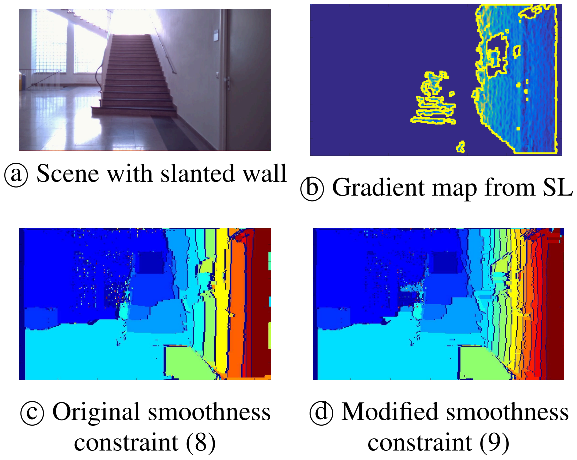
where p_1, p_2 are neighboring pixels in the left and q_1, q_2 in the right image. The smoothness term in (1) can therefore be expanded to

$$E_{\mathrm{S}}(f) = \sum_{a_1 \sim a_2} V_{a_1,a_2} \cdot 1\left(f(a_1) \neq f(a_2)\right), \quad (7)$$

where $a_1 \sim a_2$ indicates that pixels p_1 and p_2 are adjacent and share the same disparity $f_{p_1} = f_{p_2}$. In that case, both assignments should be either *active* or *inactive*, hence $1\left(f(a_1) \neq f(a_2)\right)$, otherwise the smoothness term contributes a penalty V_{a_1,a_2}. The terms active and inactive refer to whether a disparity is necessarily assigned to a pixel or a pixel is otherwise labeled as occluded and accordingly not active. In (Kolmogorov et al., 2014) the smoothness cost is defined as follows

$$V_{p_1,p_2} = \begin{cases} 3\lambda & \text{if} \quad \max(|I_{\mathrm{l}}(p_1) - I_{\mathrm{l}}(p_2)|_\infty, \\ & \qquad |I_{\mathrm{r}}(q_1) - I_{\mathrm{r}}(q_2)|_\infty) < 8 \\ \lambda & \text{if} \quad \max(|I_{\mathrm{l}}(p_1) - I_{\mathrm{l}}(p_2)|_\infty, \\ & \qquad |I_{\mathrm{r}}(q_1) - I_{\mathrm{r}}(q_2)|_\infty) \geq 8 \end{cases}, \quad (8)$$

where λ is a constant that models the influence of the smoothness term. However, in case of *slanted* surfaces (here we refer to surfaces that are not parallel to the imaging plane and hence appear slanted) that exhibit a homogeneous coloring, such as the wall on the right hand-side of Figure 7 ⓐ, the assumption that neighboring pixels with similar intensity should be assigned equal disparities is no longer valid and results in a clustering of the assigned disparity values as shown in 7 ⓒ. To overcome this deficiency, we again exploit the prior information available from the SL sensor. We introduce an additional term that is based on the gradient map of the structured light sensor data (which is shown for the same example in Figure 7 ⓑ) such that the smoothness penalty becomes aware of the existence of slanted surfaces and is adapted accordingly. More specifically, the smoothness penalty is extended by


ⓐ Scene with slanted wall

ⓑ Gradient map from SL

ⓒ Original smoothness constraint (8)

ⓓ Modified smoothness constraint (9)

Figure 7: Benefits of the modified smoothness constraint compared to the original one used in (Kolmogorov et al., 2014). The modified constraint exploits prior depth knowledge from structured lighting to compute a disparity gradient map which is used to downweight the smoothness constraint for slanted surfaces disallowing the enforcement of a uniform disparity on surfaces of homogeneous pixel intensity however with a varying depth. Notice, the floor is still largely assigned a single disparity value as no SL measurements are available on this image region and accordingly a gradient map for this region cannot be computed.

$$V_{p_1,p_2} = 0.1\,\lambda \quad \text{if} \begin{cases} c_{\nabla,l} \leq \nabla D_{\mathrm{SL}}(p_1) \leq c_{\nabla,u} \\ \qquad \& \\ c_{\nabla,l} \leq \nabla D_{\mathrm{SL}}(p_2) \leq c_{\nabla,u}, \end{cases} \quad (9)$$

where ∇D_{SL} is the gradient map of the SL sensor measurement. The constant $c_{\nabla,l}$ (respectively $c_{\nabla,u}$) serves as a lower (upper) threshold for the gradient values. The thresholds are introduced to assign a low smoothness penalty for slanted surfaces for which the gradient values are supposed to vary within an interval $[c_{\nabla,l}; c_{\nabla,u}]$. The gradient constraint in Equation (9) is dominant and overrules Equation (8). The additional constraint was found to significantly improve the performance of the algorithm for slanted surfaces as shown in Figure 7 ⓓ.

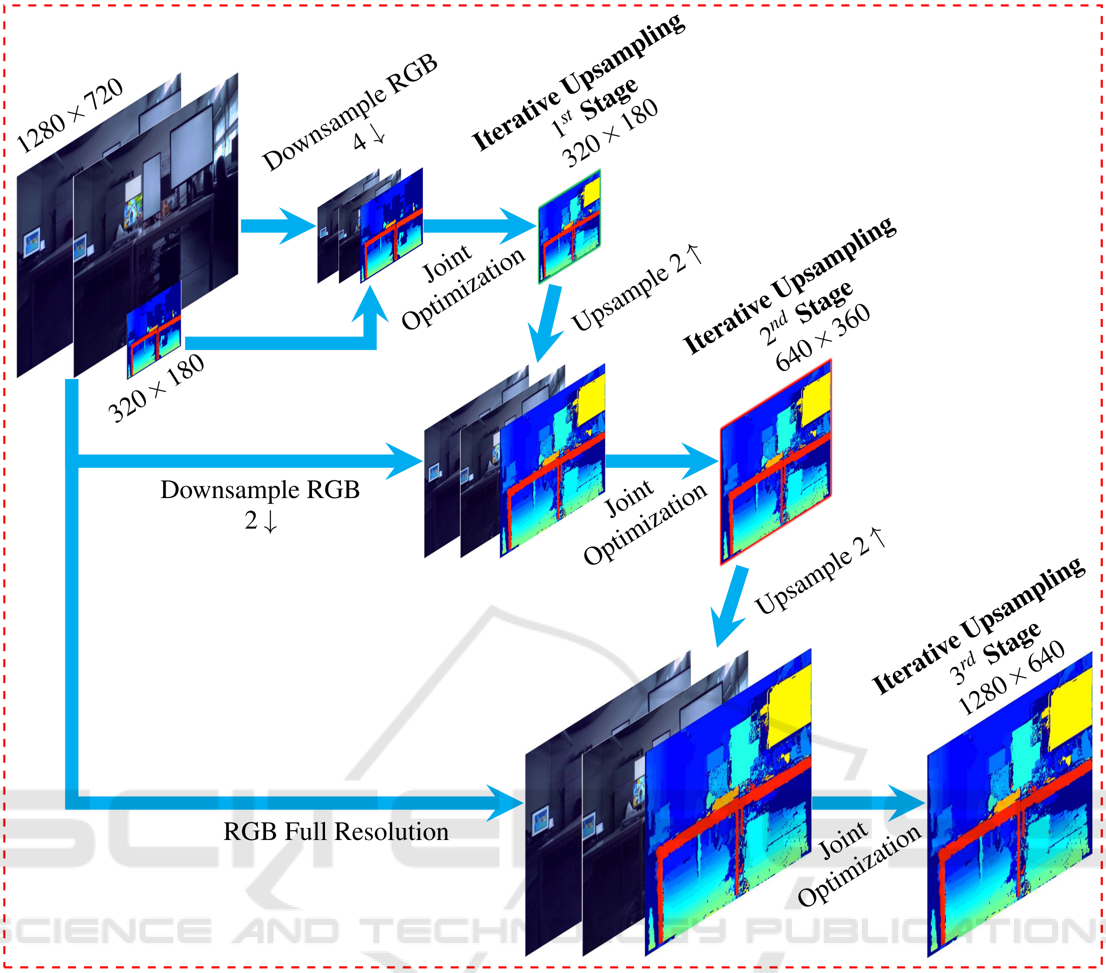
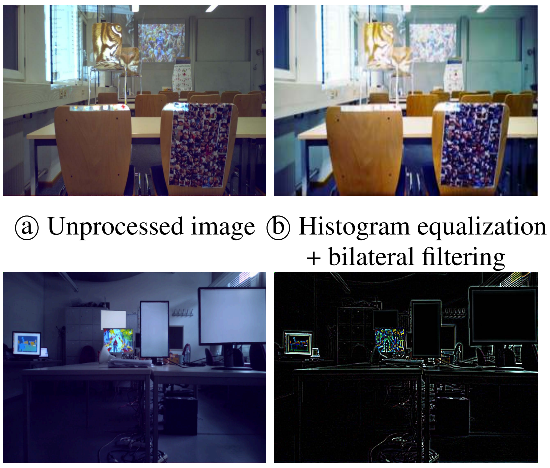


Figure 8: Iterative fusion: Starting with down-sampled stereo images, a first iteration of fusion is very efficiently achieved resulting in a dense depth map, which, after upsampling, serves as an enhanced prior for a second iteration of fusion at a higher resolution. The process can be continued until achieving the desired resolution. In all stages, a joint optimization of the prior and the RGB-images is performed by utilizing a global energy formulation.

4.2 Depth Map Superresolution

Since the structured light measurements and the stereo images recorded with the Google Project Tango device do not have the same resolution, we designed an iterative algorithm that increases the resolution of the disparity map from the low-resolution of the structured light sensor to the high-resolution of the stereo images. As opposed to other approaches that project a low resolution depth prior into the high resolution stereo images to initiate a fusion, we propose an iterative approach that initializes the fusion at low-resolution to obtain a dense disparity map, which is then upsampled and used to reinitialize the algorithm in a second iteration. Again, a nearest neighbor upsampling is applied to the disparity map found in the second stage and the algorithm is initialized a third time and converges at the resolution of the stereo images. The underlying rationale behind starting the fusion at the lower resolution of the SL depth maps is related to Equations (3) and (4). In essence, an upsampling of the SL depth image by a factor s to adapt its resolution to the stereo image pair also implies extending the disparity search range by the same factor. Accordingly, we do the converse by first downsampling the stereo image pair. Hence, we obtain a great reduction of the disparity search intervals for pixels with and without a disparity prior, leading to a very fast computation of the first fused depth map at the same resolution of the SL depth map. The results in Figure 12 show that this first fused depth map is computed at a fraction of that with the full stereo image pair resolution. This first fused depth map in turn makes up a reliable prior for the second fusion stage



ⓐ Unprocessed image ⓑ Histogram equalization + bilateral filtering

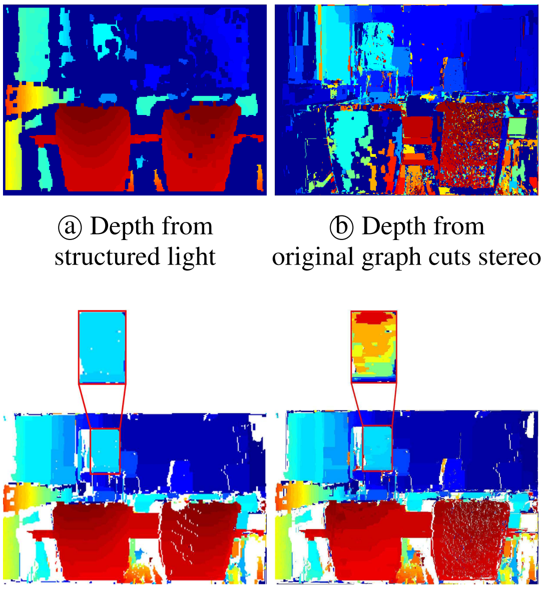
ⓒ Unprocessed image ⓓ Bilateral filtering with background subtraction

Figure 9: Effect of histogram equalization + bilateral filtering (ⓐ → ⓑ) and bilateral filtering with background subtraction (ⓒ → ⓓ) on sample images.

at the doubled resolution and in turn significantly reduces the disparity search space for the pixels. For the first iteration, the disparity search range for a pixel p that exhibits prior knowledge is set to $f_p^{\text{SL}} \pm 3\sigma$ as explained in Section 4.1.1. However, if the disparity map resulting from the first iteration of the algorithm can be assumed to exhibit the true disparities, upsampling it by a factor of 2^n would result in an uncertainty of $\pm 2^{n-1}$ pixels. Therefore, unlike in the first iteration, the initialization of the proposed fusion algorithm with the disparity map obtained from the previous stage upsampled by a factor of 2, only requires a search interval of $f_p^{\text{SL}} \pm 1$ pixels for a pixel p that exhibits prior knowledge in the 2^{nd} and 3^{rd} stage. An initialization of the algorithm with a dense prior therefore ensures a tremendous acceleration of the disparity computation for the 2^{nd} and 3^{rd} iteration of the iterative fusion. Figure 8 shows the 3 stages of the proposed algorithm.

5 RESULTS

This section presents results that we obtained using the fusion approach explained in Section 4. Before running the algorithm, the stereo images are preprocessed to adjust their brightness and remove noise. Depending on the image scene, we found a pipeline of histogram equalization (Liling et al., 2012) and bilateral filtering (Ansar et al., 2004) (see Figure 9 ⓐ, ⓑ), or bilateral filtering with background subtraction (Ansar et al., 2004) (see Figure 9 ⓒ, ⓓ) to improve the performance of the algorithm.



ⓐ Depth from structured light ⓑ Depth from original graph cuts stereo

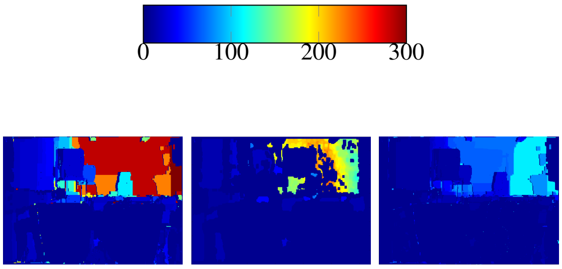
ⓒ Our approach 1^{st} stage ⓓ Our approach 3^{rd} stage

Figure 10: From the depth map obtained using SL sensing ⓐ many regions have no depth values due to bright illumination or their surface properties. ⓑ shows a depth map computed using state-of-the-art stereo reconstruction via graph cuts. Of particular note is that several objects in the scene are assigned a wrong disparity (e.g. the backs of a chair, table top) which can be related to ambiguities in the cost computation. Also, the slanted wall on the left-hand side of the disparity map contains a large number of missing assignments and exhibits a strong clustering of disparities. ⓒ and ⓓ show the result achieved with the fusion algorithm proposed in this work. Both disparity maps contain a significantly higher number of disparity assignments than ⓐ and ⓑ. At a first glance, both images look alike, but ⓓ has a higher spatial resolution (pixels) and depth resolution than ⓒ. The latter can be observed from the magnified parts in ⓒ and ⓓ, where only two disparities are assigned for the enlarged part of ⓒ, while for the enlarged part in ⓓ, a range of disparity assignments results in a finer discretization of the depth space (a seperate colormap is used for this enlargement).

5.1 Datasets

To the best of our knowledge there exists no benchmark dataset for joint stereo-SL depth estimation. Accordingly, we show the results obtained on three sample scenes captured with the Project Tango device. Through detailed analysis we highlight how individual blocks contribute towards providing a final estimation that is better than that obtainable with either stereo or SL reconstruction alone.

Lecture Hall: We generated depth images of the scene shown in Figure 9 ⓐ, with SL sensing and



ⓐ Uncertainty 1st stage ⓑ Uncertainty structured light ⓒ Uncertainty 3rd stage

Figure 11: Uncertainty/resolution in the depth computation for the first stage of the fusion algorithm ⓐ, the structured light sensor ⓑ and the third stage of the iterative fusion ⓒ in mm. For ⓑ, the uncertainty/resolution is computed using the focal length and the baseline b_{SL} assuming a standard deviation in disparity measurement σ_{SL} of $\frac{1}{8}$-pixel disparities, a value adopted from a well-known fact about the Kinect's Primesense chip. The uncertainty/resolution plot shows that the 3rd-stage of the iterative fusion approach has a higher depth resolution than ⓐ and ⓑ.

stereo reconstruction as well as using our proposed fusion algorithm. The resulting depth maps are depicted in Figure 10. The regions of the depth maps covering surfaces illuminated by the strong light from the projector reveal how beneficial the proposed fusion of SL sensing and stereo reconstruction is. The noise in the stereo depth map (Figure 10 ⓑ) observed on the back of the close chair no longer appears in the fused depth map, as the adapted data term (Section 4.1.1) reduces the ambiguity in disparity assignments. Furthermore, our approach reduces the clustering of disparities, also observable in ⓑ, by utilizing a gradient map based on the SL measurements as explained in Section 4.1.2.

Figure 11 shows the depth resolution of our approach vs. the depth resolution of the structured light sensor of the Project Tango device. It can be seen that the final stage of the suggested algorithm achieves a better depth resolution than the SL measurements.

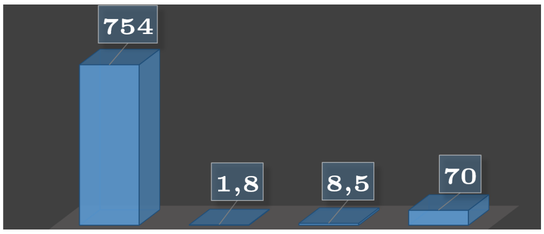
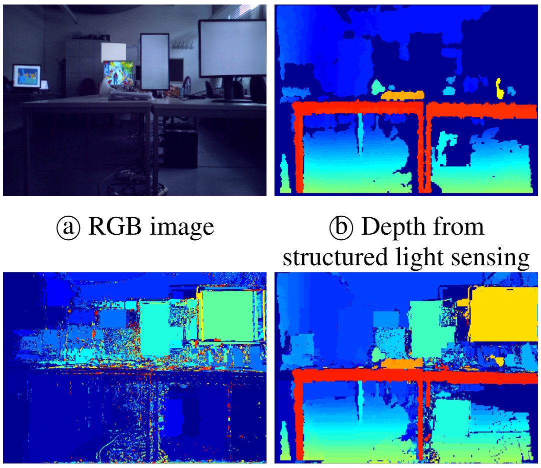


Figure 12: Computation time required by the original graph cuts stereo reconstruction algorithm vs. computation time of the iterative fusion approach for the dataset ”lecture hall” in seconds. From left to right: original graph cuts stereo reconstruction, iterative fusion approach 1st stage, - 2nd stage and - 3rd stage.



ⓐ RGB image ⓑ Depth from structured light sensing

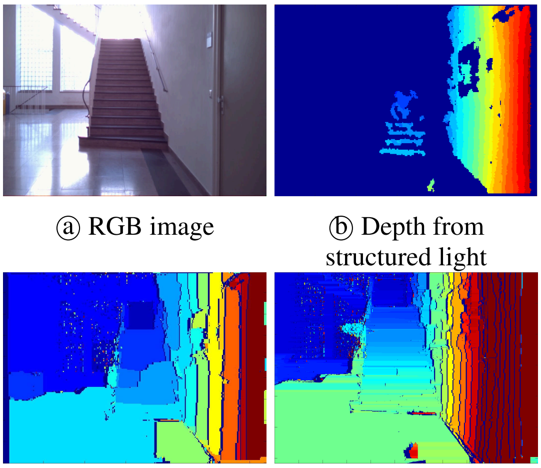
ⓒ Depth from original graph cuts stereo ⓓ Depth using proposed approach

Figure 13: The structured light sensor ⓑ could not capture the depth measurements for the monitors shown in ⓐ, since the infrared pattern cannot be retrieved in that case. The original graph cuts stereo reconstruction algorithm ⓒ shows many wrong disparity assignments, resulting from occlusion and ambiguities. The corresponding disparity map also contains many fragments and is rather noisy. The iterative fusion approach ⓓ instead, provides a dense disparity map that outperforms the individual measurements by far and even reconstructs the slanted wall and floor, which is challenging for the original graph cuts stereo reconstruction algorithm.

Beyond the enhanced depth image quality, the proposed iterative fusion algorithm allows us to tremendously reduce the computational complexity of the disparity search, as illustrated in Figure 12. It shows the processing times of the original graph cuts stereo reconstruction algorithm vs. the three stages of the iterative upsampling approach. It can be seen that the computation time of the depth image is reduced by almost 90%.

Student Lab: Figure 13 depicts the evaluation of the dataset ”student lab”, which shows a common office environment. In this case, reconstruction using structured light scanning fails for the monitors, which disturbs the IR-pattern. However, the global stereo reconstruction approach successfully reconstructs most of the monitors. The figure clearly shows how the proposed iterative fusion approach compensates for the shortcomings of both modalities and also allows a strong reduction of the required computation time from 1464s to 83.7s. In other words, the calculation time is reduced by 94.3%.

Stairs: Figure 14 shows the evaluation of the dataset ”stairs”. The scene exhibits a large number of



ⓐ RGB image

ⓑ Depth from structured light

ⓒ Depth from original graph cuts stereo

ⓓ Depth using proposed approach

Figure 14: The structured light sensor ⓑ is significantly disturbed by the sunlight that can be observed through the large glass windows. The original graph cuts stereo reconstruction algorithm ⓒ fails to reconstruct the floor as well as to resolve the large number of depth levels associated with the steps. The slanted wall is represented by a few disparity clusters and does not allow to recognize the geometry of the scene. The iterative fusion approach ⓓ instead, performs a good reconstruction of the slanted wall exploiting the SL measurements and also resolves the fine depth levels of the steps in the scene to a large extent. Stereo reconstruction also benefits the fusion by partly reconstructing the wall in the background.

discrete depth levels at the steps of the stairs, as well as a continuously increasing depth along the slanted wall on the right hand side of ⓐ. The iterative fusion approach allows resolving the fine depth levels of the stairs and also outperforms the standard graph cuts stereo reconstruction of the slanted wall in the scene. The computation time of the graph cuts stereo algorithm was measured with 560s, while all three stages of the iterative fusion approach only require 48.7s, which corresponds to a reduction 91.3%.

6 CONCLUSION

In this paper, we presented a novel approach to generate high-resolution depth maps from a fusion of a low-resolution structured light prior with state-of-the-art stereo reconstruction in an iterative process. Unlike other approaches that perform a fusion of active and passive methods in 3D imaging, we do not upsample a low-resolution prior to match the high-resolution of the stereo images, but instead perform a fusion at low-resolution and use the resulting disparity map as a prior to iteratively reinitialize the algorithm until it converges. This strategy has two major advantages. Limiting the disparity search to the uncertainty interval of the prior greatly reduces ambiguities and also allows for a significant reduction of the runtime of the algorithm. Initializing the fusion at low-resolution amplifies the effect, as not only the disparity search range for low-resolution images is smaller, but also the number of pixels that have to be assigned a disparity. High-resolution disparity maps are then inferred in an iterative upsampling process that ensures a consistent computational complexity, also for sparse priors. For the fusion approach discussed in this paper, we found the following points to have potential for further improvement in future work.

- Without a prior, surfaces with continuously changing disparities are challenging for the algorithm. This problem could be reduced with an additional constraint based on the stereo images.
- A GPU-implementation of the algorithm could help to significantly accelerate the computation.
- Initializing the algorithm at even lower resolution can further decrease the runtime.

REFERENCES

Ansar, A. I., Huertas, A., Matthies, L. H., and Goldberg, S. (2004). Enhancement of stereo at range discontinuities. In *Defense and Security*, pages 24–35. International Society for Optics and Photonics.

Birchfield, S. and Tomasi, C. (1998). Depth discontinuities by pixel-to-pixel stereo. In *Computer Vision, 1998. Sixth International Conference on*, pages 1073–1080.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239.

Choi, S., Ham, B., Oh, C., gon Choo, H., Kim, J., and Sohn, K. (2012). Hybrid approach for accurate depth acquisition with structured light and stereo camera. In *Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on*, pages 1–4.

Engel, J., Schöps, T., and Cremers, D. (2014). Lsd-slam: Large-scale direct monocular slam. In *Computer Vision–ECCV 2014*, pages 834–849. Springer.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. *International journal of computer vision*, 70(1):41–54.

Foix, S., Alenya, G., and Torras, C. (2011). Lock-in time-of-flight (tof) cameras: A survey. 11(9):1917–1926.

Gandhi, V., Cech, J., and Horaud, R. (2012). High-resolution depth maps based on tof-stereo fusion. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4742–4749.

Goldberg, A. O., S., L., M., D., T., t., J., S., J., D., and T., G. (2014). Google tango teardown. www.ifixit.com/Teardown/Project+Tango+Teardown/23835.

Khoshelham, K. and Elberink, S. O. (2012). Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454.

Kolmogorov, V., Monasse, P., and Tan, P. (2014). Kolmogorov and Zabihs Graph Cuts Stereo Matching Algorithm. *Image Processing On Line*, 4:220–251.

Li, Q., Biswas, M., Pickering, M. R., and Frater, M. R. (2011). Accurate depth estimation using structured light and passive stereo disparity estimation. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 969–972.

Liling, Z., Yuhui, Z., Quansen, S., and Deshen, X. (2012). Suppression for luminance difference of stereo image-pair based on improved histogram equalization. *Proceedings of the computer science and technology*, 6:2.

Newcombe, R. A., Lovegrove, S. J., and Davison, A. J. (2011). Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE.

Occipital (2015). Occipital depth sensors. http://occipital.com/.

Piszczor, M. and Yang, C. (2014). Project tango. https://sites.google.com/a/google.com/project-tango-sdk/home.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42.

Scharstein, D. and Szeliski, R. (2003a). High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–195.

Scharstein, D. and Szeliski, R. (2003b). High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1.

Scharstein, D. and Szeliski, R. (2015). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. http://www.vision.middlebury.edu/stereo/.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528. IEEE.

Somanath, G., Cohen, S., Price, B., and Kambhamettu, C. (2013). Stereo+kinect for high resolution stereo correspondences. In *3D Vision - 3DV 2013, 2013 International Conference on*, pages 9–16.

Tappen, M. F. and Freeman, W. T. (2003). Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 900–906. IEEE.

Wei-Chen Chiu, U. B. and Fritz, M. (2011). Improving the kinect by cross-modal stereo. In *Proceedings of the British Machine Vision Conference*, pages 116.1–116.10.