Guided optimization framework for the fusion of time-of-flight with stereo depth

Faezeh Sadat Zakeri,[a] Mårten Sjöström,[b] and Joachim Keinert[a,*]
[a]Fraunhofer Institut für Integrierte Schaltungen, Computational Imaging and Algorithms Group, Moving Picture Technologies, Erlangen, Germany
[b]Mid Sweden University, 3D Realistic Group, Information Systems and Technology, Sundsvall, Sweden

Abstract. The fusion of depth acquired actively with the depth estimated passively proved its significance as an improvement strategy for gaining depth. This combination allows us to benefit from two sources of modalities such that they complement each other. To fuse two sensor data into a more accurate depth map, we must consider the limitations of active sensing such as low lateral resolution while combining it with a passive depth map. We present an approach for the fusion of active time-of-flight depth and passive stereo depth in an accurate way. We propose a multimodal sensor fusion strategy that is based on a weighted energy optimization problem. The weights are generated as a result of combining the edge information from a texture map and active and passive depth maps. The objective evaluation of our fusion algorithm shows an improved accuracy of the generated depth map in comparison with the depth map of every single modality and with the results of other fusion methods. Additionally, a visual comparison of our result shows a better recovery on the edges considering the wrong depth values estimated in passive stereo. Moreover, the left and right consistency check on the result illustrates the ability of our approach to consistently fuse sensors. [DOI: 10.1117/1.JEI.29.5.053016]

Keywords: multimodal sensor fusion; time-of-flight; active and passive sensing; optimization; stereo depth estimation.

Paper 200324 received May 12, 2020; accepted for publication Sep. 22, 2020; published online Oct. 22, 2020.

1 Introduction

Depth acquisition has been investigated for many years as an established area of research due to its many applications, such as view synthesis, autonomous navigation, machine vision, and many more. The stereo depth estimation algorithms are problematic when the scene contains weakly textured areas or occlusions in both indoor and outdoor environments.[1] Active sensing, on the other hand, does not have to deal with these problems. Instead, it suffers from different sources of noise and it performs poorly on non-Lambertian surfaces.[1]

Due to the complementary behavior of passive algorithms and active depth sensing,[2,3] the idea of combining the two modalities got its interest to be one fundamental problem in computer vision.[4] Unfortunately, the combination of active with passive depth is not straightforward. The reason is due to the limitation of manufacturing depth sensors. Most of these depth sensors either use structured illumination or measure time-of-flight (TOF). The first technique can reliably measure depth even in the areas where stereo depth estimation is an ill-posed problem such as occlusions.[4] But its main drawbacks are the lack of performance on the edges and limited depth range.[4] The TOF method can capture depth in various ranges with better performance at boundaries compared to depth sensors using structured illumination.[5] However, the TOF sensors suffer from different sources of noise that make it difficult to get the most benefit from these sensors.[5] Most of these noises are treated by the manufacturers.[6,7] Such noises include quantization and thermal noise and some sources, such as photon shot noise, that can be approximated.[8]

*Address all correspondence to Joachim Keinert, E-mail: joachim.keinert@iis.fraunhofer.de

The main problems of TOF cameras are their low lateral resolution and the "flying pixels" at the borders.[6] These can be the biggest challenge while combining the measured depth from TOF with other sources of depth. To combine two sources of depth, there are various fusion imaging techniques each solving the limitations and sources of errors differently. In Sec. 1.1, we summarize these techniques by categorizing them based on their pros and cons and in the end, we highlight our contributions in this work.

1.1 *Literature Review*

There are several categories of image fusion techniques that are surveyed in detail in Ref. 9. Our focus in this paper is mainly on the category of guided approaches as our work stays within this category. The common assumption in most of the guided image fusion techniques is based on the similarity between the depth and the color image on the local structures.[10,11] There are several methods in this category that additionally assume the homogenous color areas are geometrically similar.[12–15] The well-known Markov random field formulates the fusion as an upsampling problem and defines the data term as the weighted depth map. The weights are computed from the high-resolution image to smooth the estimated depth values on the high-resolution texture map.[13]

The other well-known fusion proposal is the approach of Ref. 15, which uses joint bilateral filtering (JBF)[16] and interpolates depth values on the high-resolution texture map. Due to the interpolation, depth values are oversmoothed mostly on depth discontinuities. Because of the fast performance of the JBF approach reported in Ref. 12, the approach was extended in different variations. For instance, Ref. 12 uses regional statistics on the depth map to perform the JBF. This allows deciding how to combine the result of upsampling and hence performs well in the existence of noisy depth data from a range sensor.[12] Even though the variations of JBF improve the results, the problem of blurring the depth values is the downside of these approaches.[17]

To have a better recovery along the depth boundaries,[18] they formulate the problem of upsampling for the fusion based on a constrained optimization framework. They use the idea of depth from focus by utilizing regularizers computed from nonlocal structures.[19] They re-enforce the similarities by edge-weighting schemes to achieve a better estimation of the fine details. Some approaches[3,20] came up with the idea of guiding the optimization framework by applying image segmentation that results in better edge recovery. Other approaches[21–23] formulate their guidance as an edge-weighting scheme extracted by enforcing spatial similarity between neighboring pixels from both image and depth data without explicit image segmentation. Moreover, their method uses the amplitude values to calculate the reliability of the measured depth in addition to other guided optimization methods. For this reason, it achieves more accurate depth values at the end. Since they stated their problem mainly as a TOF upscaling problem, thus, they consider a sparse value mapping on the TOF samples in their implementation. The sparse value mapping is a process that represents TOF samples after warping on the high-resolution texture map that results in sparse representation of TOF samples due to its low lateral resolution.[21] In the sparse value mapping, first, the TOF depth map is sparsified by keeping values of the regular columns and rows and setting the other depth values as unknown. Keeping values of columns and rows regularly only does not perfectly mimic warping TOF data on the high-resolution texture map. In this way, the lens distortion of the TOF sensor is neglected, and therefore, the irregular representation of TOF data after projection on a high-resolution texture map is not considered. In addition, where the relative resolution of TOF is much lower than the resolution of the high-resolution texture map, their method does not perform well. This might be because the assumed smoothness constraint considers the similarity of a depth value to the depth value of its four neighboring pixels only. The work of Dal Mutto et al.[24] utilizes a locally consistent framework to fuse the two data sources regardless of their relative resolution as a set of local energy minimization problems. The work of Marin et al.[3] improves the one of Ref. 24 by driving the fusion process with the depth map confidences to take into account the different nature and reliability of the data sources. But they report inaccuracies on the recovered edges.[3]

Another improvement of Marin et al.[3] incorporates a convolutional neural network to estimate TOF and disparity confidences using the amplitude information of TOF and semiglobal matching[25] for aggregating matching costs for stereo and finally fuse two sources of data reliably.

Their results represent fusion results with smoother edges in comparison with Ref. 3 but numerically they do not outperform.[3]

Since the inaccuracy of the above methods near the edges is the most common problem, our main aim is to recover precise edges and to not oversmooth the depth values in the fused depth map. Hence, we decided to apply optimization-based approaches to fuse TOF with a stereo depth map. We extend the work of Schwarz et al.[21] by adding stereo constraints into the optimization framework. Additionally, we guide the optimization framework to fuse TOF and stereo depth consistent with the image information. This is a result of the way we generate the edge-weighting values as joint incorporation of edge information from two sources of data and cross masking them with the image information considering the directional manner of the local structures. Thus, the fused map is consistent with the two data sources. Our experiments show stereo consistency between fusion maps on the left and right camera positions in our stereo setup. This permits the usage of our method for applications such as view rendering that mandate consistency between stereo depth maps. Our approach outperforms the state-of-the-art methods both visually and numerically. Our algorithm computes more accurate depth values on the fused map especially on the edge boundaries because we use the two complementary sources of data in a guided optimization framework. Also due to the embedment of confidences for TOF and stereo, our approach leads in an even more accurate fused map specifically on the homogenous areas.

2 Proposed Method

2.1 *Optimization Framework*

The main idea of the weighted optimization framework is introduced in Refs. 21–23. Their proposal incorporates three sets of input sources, a video camera, a TOF sensor, and a preceding time-sequential super-resolution result.[21–23] They use the three sources in a super-resolution task in the time domain to solve the problem of TOF upscaling to the resolution of the video camera regarding a target point of view. We use their framework as the task of TOF super-resolution for TOF plus stereo depth map fusion. We compute a fusion map by minimizing the combination of weighted-error energy terms: spatial error energy Q_S, depth error energy Q_D, and stereo error energy Q_{St}. For projecting the low-resolution TOF depth $D_{\downarrow}$ onto a high-resolution depth map D of the reference camera in the stereo system, we apply a piecewise approach, where piecewise projection means that there is no smoothing happening during or after projection. Consequently, we assume a similarity between spatially neighbored depth values that are expressed as horizontal and vertical errors. We formulate our fusion problem to attain high-resolution fused depth map D^* as

$$D^* = \arg\min_{\hat{\mathrm{D}}}(k_1 Q_{\mathrm{s}} + k_2 Q_D + k_3 Q_{St}), \quad k_1 + k_2 + k_3 = 1, \tag{1}$$

where $\hat{D}$ is the all possible high-resolution estimates and k is the regularizer for each energy term that determines the contribution of each error energy term in the energy system. The value for each k is fixed in our implementation and defined as a constant.

2.2 *Spatial Error Energy*

The spatial error energy term Q_S penalizes discontinuities in the depth values. Hence, it enforces a smooth depth value distribution in the high-resolution depth estimate $\hat{D}$. To construct the piecewise aspect of the depth value distribution, the edge-weighting map W_E is introduced.

$$Q_S = \sum_{x=1}^{X}\sum_{y=1}^{Y}(W_{Eh(x,y)}\{[d_{(x,y)} - d_{(x+1,y)}]^2\} + W_{Ev(x,y)}\{[d_{(x,y)} - d_{(x,y+1)}]^2\}). \tag{2}$$

This construction with the introduction of an edge-weighting map allows holding a similarity assumption between spatially neighboring depth values by penalizing discontinuities. It is important to note that we generate the edge weights in horizontal $[W_{Eh(x,y)}]$ and vertical

$[W_{Ev(x,y)}]$ directions. Consequently, edge-weighting maps consider the direction of edges. This permits the enforcement of similarities between the depth and texture map in a more robust way that yields in more correct edge information.

2.3 *Edge Weights*

The edge weighting allows for sharp depth transitions between objects by relaxing the spatial similarity constraint for neighboring pixels at object boundaries. Weighting elements $W_{E(x,y)}$ are obtained by combining the texture information from the RGB image and the depth information obtained from the TOF and stereo depth estimation. The combination is based on three assumptions. First, because the depth maps describe the scene geometry, the object boundaries should correspond to edges in the corresponding texture from the image. Second, because thorough edge detection on an image will result in many more edges than there are actual objects, the edge information from the low-resolution TOF and stereo depth map can be utilized to select edges that comply with actual depth transitions. Third, due to the higher precision of measured depth from TOF in comparison to the estimated depth from the stereo in most cases, the wrong estimated edges from stereo are invalidated by edge weights of the TOF image. Figure 1 shows the combination of texture and depth information from different sources for generating an edge-weighting map W_E.

As shown in Fig. 1, a horizontal edge filter on the high-resolution $I_{(\text{img})}$ yields the edge map $E_{I(\text{img})}$. The low-resolution depth map $D_{L(\text{TOF})}$ is also edge filtered after it warped to the high-resolution texture map and the result is upsampled to the corresponding texture map of the image to form the edge mask $E_{D_{(\text{TOF})}}$. Furthermore, the high-resolution stereo depth map $D_{(\text{stereo})}$ is edge filtered to produce the edge mask $E_{D(\text{stereo})}$. The edge masking makes the selection of edges in a cohesive manner, which affects having higher quality edges in the fused depth map. References 21–23 explain that the missing or porous edges can lead to depth leakage where erroneous depth values spread into the wrong areas. However, thorough edge detection in the image will result in many more edges than there are actual objects as shown in Fig. 1. This will lead to an unwanted structuration effect in the upsampled depth map.[21–23] A larger threshold for the edge detector reduces the number of unnecessary edges. However, less detected edges will increase the risk of depth leakage. Finding the correct edge threshold for each sequence is difficult. Therefore, it is practical to use a smaller edge detector threshold, i.e., a greater sensitivity, leading to more edges being detected, and validate the resulting edge map with actual depth transitions in TOF and stereo depth. Multiplying each element of $E_{I(\text{img})}$ by its corresponding element in $E_{D(\text{TOF})}$ and $E_{D(\text{stereo})}$ masks out redundant edges in areas with uniform depth and returns the edge-weighting map W_E:

$$W_{E(x,y)} = 1 - E_{I_{(\text{img})}(x,y)}.E_{D_{(\text{TOF})}(x,y)}.E_{D_{(\text{stereo})}(x,y)}. \tag{3}$$

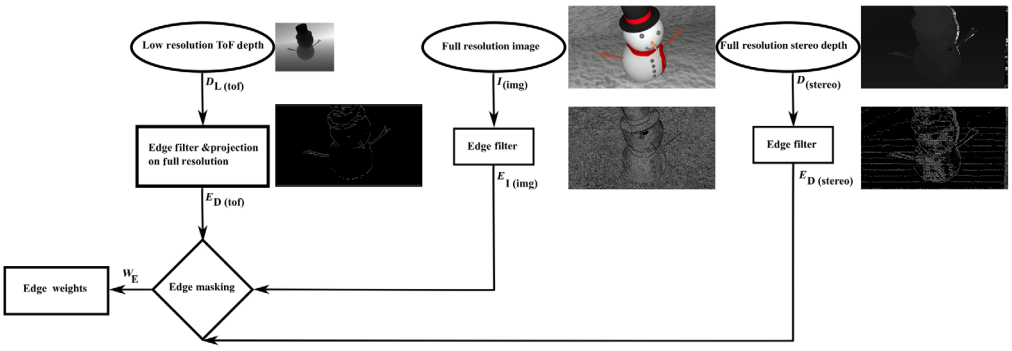

Fig. 1 Scheme showing the generation process for the edge-weighting map W_E. The edge maps correspond to each source and are calculated with different thresholds. The edge maps exhibited here are in a horizontal direction only.

2.4 *TOF Depth Error Energy*

The depth error energy Q_D enforces similarity between the sparse TOF depth and the final high-resolution result. We present our depth error definition in Eq. (4) such that $d_{(x,y)}$ is the unknown depth value on the high-resolution texture map while $D_{(x,y)}$ is the depth value that the TOF sensor sampled. Optimizing our energy system with Q_D as one of its components allows estimating the values for unknown depth values knowing sampled depth TOF values as our data term. However, the piecewise projection of low-resolution TOF on the texture map of the image might be not reliable in terms of mapping points from TOF to the texture map. Therefore, the reliability weights W_D are introduced as

$$Q_D = \sum_{x,y} \mathrm{W}_{D(x,y)}\{[d_{(x,y)} - D_{(x,y)}]^2\}. \tag{4}$$

2.5 *TOF Reliability Weights*

The reliability weights can be used to remove erroneous depth values, suppress low reliability, and emphasize high-reliability depth values in the upsampling process.[21–23] The reliability of the TOF measurements is affected by several issues, e.g., the reflectivity of the acquired surface, the measured distance, multipath issues, or mixed pixels in the proximity of edges, and thus is very different for each different sample.[26–28] In this paper, we have calculated the reliability weights by the method proposed in Ref. 3. They calculate confidences for each TOF pixel by considering both the geometrical and photometrical properties of the scene. P_{AI} considers the relationship between amplitude and intensity of the TOF signal, whereas P_{LV}, accounts for the local depth variance. The two confidence maps P_{AI} and P_{LV} consider independent geometric and photometric properties of the scene. Therefore, the overall TOF confidence map W_D is obtained by multiplying the two confidence maps together. Considering only the amplitude and intensity of the TOF signal is not enough since one of the main issues in the fusion of TOF depth values with other modalities is the finite size of TOF sensor pixels. This leads to a mixed pixel effect[29] that cannot be resolved by confidences estimated based on the photometrical characteristics of the TOF signal only. As a remedy, Ref. 3 introduces another term in the proposed confidence model, P_{LV}. The details of the formulation can be found in Ref. 3.

$$W_D = P_{AI}.P_{LV}. \tag{5}$$

The variable W_D has a value of zero at locations where there is no depth value available from the TOF sensor. This means that the values at those positions are implicitly computed from the smoothness assumption (see more in Ref. 21).

2.6 *Stereo Depth Error Energy*

The stereo error energy Q_{St} encourages stereo consistency on the fusion result and enforces TOF depth to be fused with stereo depth consistently. It also improves the enforcement of the similarity between TOF depth values projected on the high-resolution texture map. Our definition of the Q_{St} error energy term is elaborated in Eq. (6). Like Eq. (4), we aim to estimate unknown depth values on a high-resolution texture map. Adding Q_{St} in our energy equation helps the optimization process in estimating the unknown values. It allows estimating the depth of these pixels by knowing their estimated depth from the stereo as an additional data term to TOF readings.

$$Q_{\mathrm{St}} = \sum_{x,y} W_{\mathrm{St}(x,y)}\{[d_{(x,y)} - D_{\mathrm{St}(x,y)}]^2\}. \tag{6}$$

Due to the lack of accuracy in stereo depth estimation, the danger of introducing nonaccurate depth values in the fused results is inevitable. This means that the addition of a stereo constraint into the total error energy for the fusion can bring fewer benefits. To prevent the mentioned disadvantages from adding the stereo error energy into our energy system, we introduced

W_{St} by estimating stereo confidences into our formulation for Q_{St}. The stereo confidences filter out wrong depth values in the stereo depth map by assigning them low confidences. This allows us to take advantage of the stereo error energy as a supplementary source for depth fusion, especially in areas where TOF performs poorly.

2.7 *Stereo Depth and Confidence Estimation*

We use the ADCensus algorithm (ADC) for stereo disparity estimation as described in Ref. 30. Its cost function relies on both absolute pixel differences and the census transform.[30] To estimate the confidences W_{St}, we used the method proposed in Ref. 31 together with calculating matching costs for each pixel in the depth map. The proposed method in Ref. 31 computes confidence solely based on the cost curve for each pixel in the disparity map. The confidence value for each pixel indicates how likely the assigned disparity is correct. Further details about confidence estimation are explained in Ref. 31.

2.8 *Optimization Framework*

Based on the error energy term, a minimization problem is solved by setting the sum of the partial derivatives to zero.

$$\forall\ (x, y):\ \frac{\partial(k_1 Q_s + k_2 Q_D + k_3 Q_{St})}{\partial d_{(x,\ y)}} = 0. \tag{7}$$

Each partial derivative leads to a linear equation, so we get a linear system $Az = c$ with $N \cdot M$ equations. Each of it has five unknowns, where N and M are the resolutions of the high-resolution texture map. We defined our data term as a sum of known depth values, sample points from the TOF sensor, and estimated depth from the stereo. Eventually, we find a least-squares optimal result via QR factorization.[32]

3 Experimental Setups

In principle, we want to test our algorithm with data that well represent all possible scenarios. Therefore, we chose publicly available synthesized and real datasets. We also want to be able to compare our results with other algorithms. This is simpler if we use the same data as others because then we do not need to implement their methods. Since we did not have access to the implementation of the methods that we are comparing our algorithm with, we are limited to use only a few datasets and no other data. The description of the datasets we have used to test or algorithm and to compare our results with other methods are as described in the following sections.

3.1 *Synthesized Dataset*

The first dataset we used in this work is SYNTH3 that is published by Agresti et al.[27] mainly for machine learning applications. It contains 15 scenes of various objects. They synthesized the images by Blender for stereo-TOF fusion with the corresponding ground truth for each scene. More details of the dataset are elaborated in Ref. 27.

3.2 *Real Dataset*

The second dataset we used contains a collection of real data for TOF and stereo with ground truth that is suitable for the evaluation of our fusion algorithm.[24] The scenes are captured with two Basler video cameras and a Mesa SR4000 TOF range camera located in the middle of the two video cameras. The scenes include challenging object properties. The resolution of the video cameras is 1032×778 while one of the TOF sensors is 176×144. The calibration parameters

are also provided and thus the projection of the TOF depth map onto the texture maps of the stereo cameras is easily possible.

3.3 *Preprocessing*

3.3.1 *Projection of TOF to high-resolution texture map*

To align the TOF depth map with the high-resolution texture map, we must warp the TOF depth to the position of the reference stereo camera. This step is crucial to be done accurately since the lack of precision between the TOF depth and the high-resolution texture map will lead to poor accuracy on the fused results. In addition, it produces inconsistencies between the left and right fused depth maps that can cause rendering artifacts in applications such as view rendering. Using the intrinsic and extrinsic parameters of the cameras, we can warp the TOF depth to the position of the reference stereo camera using a projection as following:[4]

$$z_{\mathrm{I}} m_{\mathrm{I}} = z_D K_D K_I^{-1} [R|\theta] | m_D, \tag{8}$$

where z_D is the low-resolution TOF depth value at the position $(x_{\downarrow}, y_{\downarrow})$, K_D and K_I are the TOF and stereo camera intrinsic calibration matrices, respectively, R is the extrinsic rotation matrix, and θ is the translation vector. We project the TOF depth value at position $m_D = [x_{\downarrow}/z_I, y_{\downarrow}/z_I, 1]^T$ onto its corresponding position $m_{\mathrm{I}} = [x/z_I, y/z_{\mathrm{I}}, 1]^T$ in relation to the high-resolution stereo camera point of view. Using homogeneous coordinates, we implement the depth warping in a matrix multiplication that outputs z_I, which is the new distance of each pixel in the TOF sensor regarding the stereo camera.

3.3.2 *Prefiltering TOF and stereo depth*

Projecting TOF onto the high-resolution texture map results in a very sparse irregular depth distribution. This is due to the low resolution of TOF in comparison with the high-resolution texture map and due to lens distortion. Therefore, we apply a sparse bilateral filter as described in Ref. 33. It replaces each pixel in the depth map by the depth value of that point in the histogram of its neighbor pixels that corresponds with a preset percentile. We repeat the same filtering on the result of stereo depth estimation to get a denser depth map. This removes the outliers without degrading the image quality.

3.3.3 *Converting depth to disparity space*

We chose the disparity space for our fusion algorithm. It is worth noting that working in-depth space instead of disparity space does not have any effect on the result of the fusion algorithm since the spaces are easily convertible. Therefore, we have converted the TOF depth map into the disparity map to fuse it with the disparity map estimated from the stereo. To do so, we use the relationship between depth and disparity as follows:

$$\mathrm{disp} = bf/D, \tag{9}$$

where b is the baseline of the rectified stereo system and f is the focal length of the stereo camera after rectification.

An overview of our experimental setup and the whole of the processing pipeline is shown in Fig. 2.

3.4 *Implementation Details*

Our implementation is based on a heuristic concept that compares the confidences of the TOF and stereo source and choose the source values with higher confidence in the final fused map. In the case of equal confidences or very tiny deviation of the confidences of both modalities, the TOF samples are weighted 0.70, and stereo values are weighted 0.30. This setting is fixed for

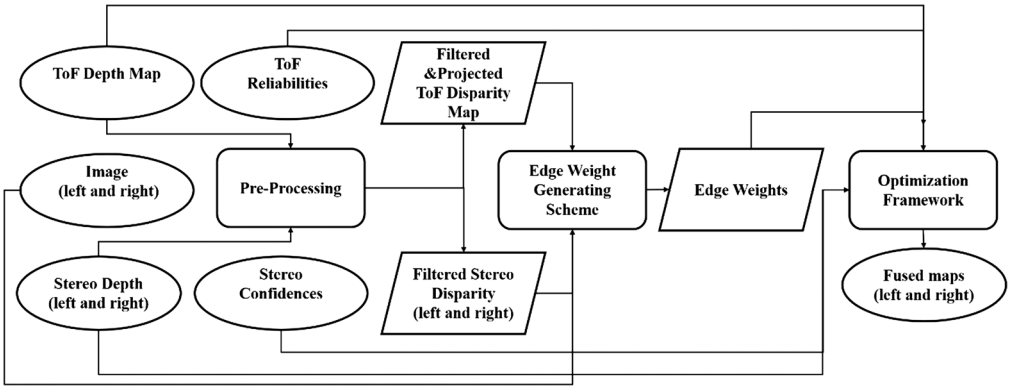

Fig. 2 An overview of the entire processing pipeline.

whole the experiments in this work and is chosen based on our observations that measured TOF signals seem to be more reliable than estimated stereo in the datasets we used.

4 Results and Evaluation

In Fig. 3, we show our fused maps for the synthetic dataset provided in Ref. 27 together with each of the input sources and the ground truth depth maps. The absolute difference (AD) maps, which calculated the differences between the disparity map results and the ground truth, are provided in a red-color-coded format for better visualization in the second column. Mean squared error (MSE) maps are also provided in the third column. We assess the precision of disparity values with average root-mean-square error (RMSE) values over all the scenes in each dataset. MSE maps of AD maps are to more clearly show where the errors occur. We assumed that the combined input data, stereo plus TOF, give the desired output, smaller errors at a certain area of the image, and therefore the RMSE and MSE are employed to give a size of the error related to the energy of the error, which is the basis for the optimization in Eq. (6). In addition, MSE emphasizes large errors more than AD. But it also implies that small errors may be difficult to see. MSE is also consistent with the metric RMSE used for the overall error presented in the table, and MSE is used in the optimization.

We have evaluated our algorithm numerically and compared it against other methods for the synthesized dataset in Table 1 (on the valid pixel mask) and Table 2 (on the whole image). Because the work of Agresti et al.[27] provided us the exact valid pixel masks that they used to compare their method with the method of Ref. 3, we could also include these methods in our objective evaluation as shown in Table 1. Based on the definition in Ref. 27, the valid pixel mask is calculated where there is a valid disparity for a pixel in all the sources (stereo, TOF, and fused map).

As it is shown in Fig. 3, the visual comparison shows improvement in the vicinity of the edges in all the scenes in comparison to the TOF data and stereo data. The TOF disparity map for scene 11 is very poor. The stereo disparity map, on the other hand, is acceptable to some extend even though the values are not very accurate. The poor performances of TOF and stereo for scene 11 can be explained based on our observation, due to the existence of the objects with the non-Lambertian surface.

Since the assumption is made for the synthesizing both TOF and stereo based on directional light, thus, the results are poorly produced for the scenes with non-Lambertian properties. Both of our evaluations on the result for this scene visually and numerically indicate the improvement in the accuracy of the disparity values on the fused map, mainly due to the usage of confidences, W_{St}, and W_D as we introduce them in our proposed method. The edges of the table are an example of this case as we can see its corresponding area in the MSE map is almost showing no error.

We improved RMSE of the state-of-the-art[27] from 2.06 to 1.64 for the synthesized dataset on the valid mask (see Table 1).

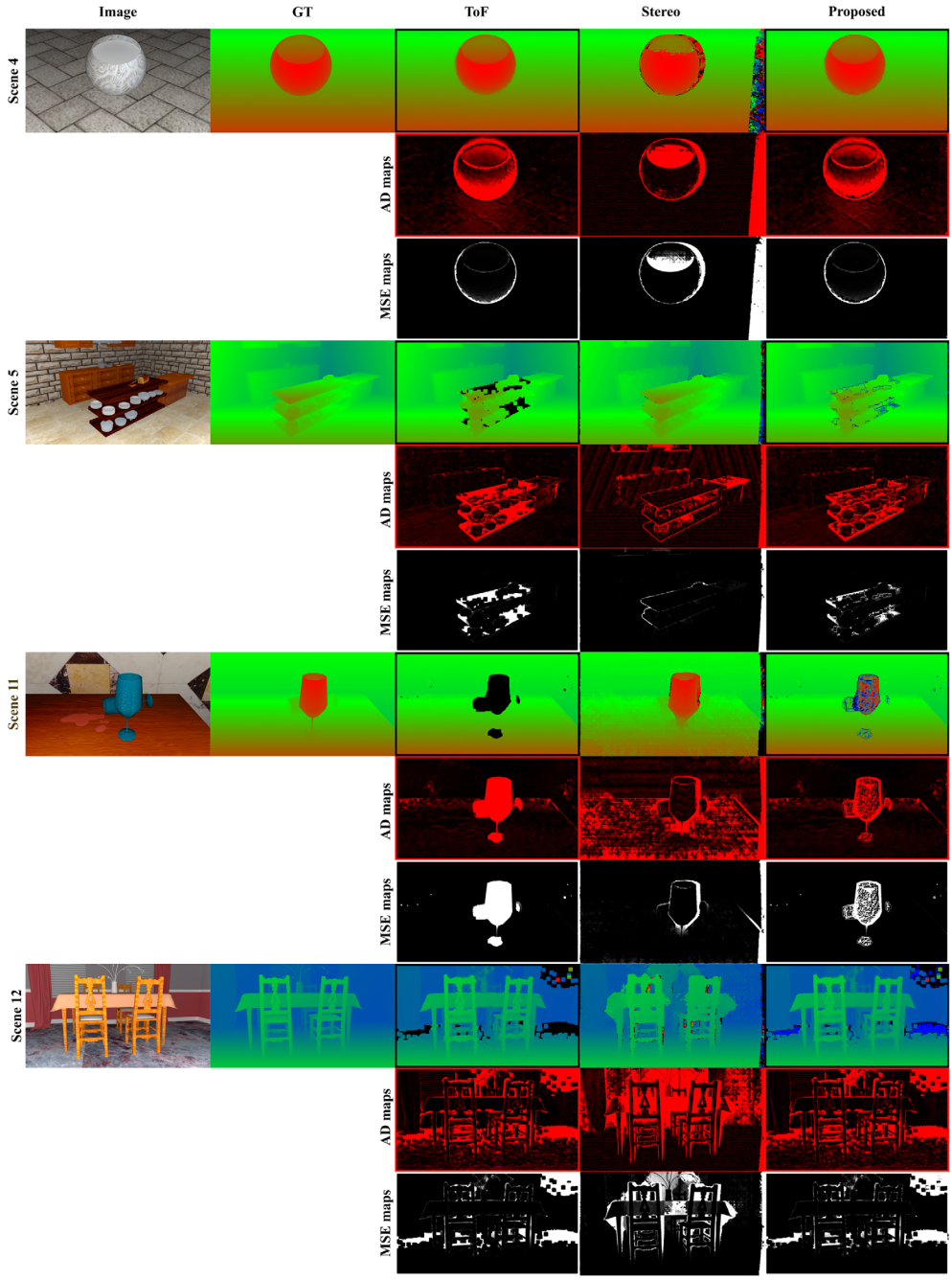

Fig. 3 The input color image is shown in the first column. Also, the ground truth, the input TOF, stereo, and our results are shown in other columns. AD maps calculated from disparity maps and the ground truth are provided in the red color-coded form in the second row for each corresponding column for all the scenes. MSE maps are presented in the third row for each corresponding column for all the scenes. The MSE unit is scaled with 10^4.

Table 1 Numerical evaluation of the result with RMSE on the provided mask from Ref. 27.

Methods	TOF	Stereo	JBF	Ref. 21	Ref. 27	Ref. 3	Proposed
Average RMSE	2.19	3.73	5.58	2.23	2.06	2.07	1.64

Table 2 Numerical evaluation of the result with RMSE on the whole image.

Methods	TOF	Stereo	JBF	Ref. 21	Ref. 27	Proposed
Average RMSE	11.77	14.36	14.34	11.21	19.20	10.49

In addition to evaluating on the valid mask, we also evaluated the input sources and our fused maps on the whole image, see Table 2. Unfortunately, we could not include the method in Ref. 3 in the evaluation of the whole image since their result was only available for the valid mask but not for the whole image. Our objective evaluations on the whole image show in Table 2 that the method of Ref. 21 outperforms the method of Ref. 27. Our proposed method is the best and reduces the RMSE from 11.21 to 10.49.

Figure 4 shows our results (seventh column) for the real image dataset provided in Ref. 24 in comparison with the filtered TOF (TOF BF, third column), the filtered stereo disparity map for the left camera (stereo BF, fourth column), the result of joint bilateral filtering on TOF BF and stereo BF (JBF, fifth column), and the result of the method[21] (sixth column).

As, according to Ref. 27, their method cannot be applied to the real images at the moment, we could not include it in our evaluation. Also, we could not consider the method of Ref. 3 in our evaluation of real images due to technical implementation issues.

The poor resolution and lens distortion of the TOF camera are the reasons why the projected TOF depth map shown in the second column of Fig. 4 is extremely sparse and the samples are irregularly placed. Consequently, applying bilateral filtering cannot recover all the edges. Hence, objects are partially deformed for TOF BF. On the other hand, the maps of stereo BF are smoother. But it is perceptually noticeable that the ADC algorithm fails in texture-less areas, such as the front of the table in all scenes and part of the books in scene C. Considering textured areas, there are fewer areas on the maps of both TOF BF and stereo BF. Hence, it is not a surprise that the map of JBF is not visually looking better. This is caused by the nature of JBF that jointly filters the results. Therefore, the relatively nonaccurate values are jointly included in the fused map instead of being filtered out.

Table 3 shows the RMSE for each scene in addition to the average RMSE for all the scenes.

These RMSE values show that JBF and stereo BF maps do not have good quality. From the AD maps, we see that disparity estimation (ADC disparity estimation algorithm) fails to estimate disparity for the homogeneous areas: in front of the table in all scenes and the body of bears in the scenes A and B. Disparity estimation also performs poorly on the reflective surfaces of the books in scene C. More importantly, it cannot estimate the disparity for the edges as the AD maps illustrate. JBF jointly fuses the TOF BF and stereo BF regardless of the reliability of each source without any edge guide. Therefore, it transfers most of the errors in the stereo map to the final fused map. The other state-of-the-art method considered in our evaluation is one of Ref. 21. Their method objectively performs better than both input data, TOF BF and stereo BF, and JBF. Thus, it reduces the overall disparity error specifically at depth discontinuities. It also performs better visually in comparison with the TOF BF and stereo BF (Fig. 4).

Since the method[21] for TOF disparity upsampling bases on the assumption of similarity between the disparity values of neighboring pixels, the result of Ref. 21 is not very distinguishable from its input data, TOF BF. The reason is that when the resolution of TOF in comparison to the stereo camera is too low, the distribution of the projected TOF pixels is too sparse. Therefore, the assumption of smoothness cannot be applied to two consecutive samples. Finally, the result of our proposed algorithm is shown in the last column of Fig. 4. It is not only visually smoother than each of the input modalities and the other two methods but also numerically more accurate. The proposed method shows a much better recovery on the edges due to our approach of cross-checking the edges from all sources of data: image, TOF, and stereo disparity. MSE maps are indicating this, the edges of small teddy bear on the top of the ball in scene *A*, the right side of the color checkerboard and in scene *C*, and front of the table in both scene *B* and *C*. Also due to the usage of confidences for TOF and stereo, the fusion itself is choosing the more reliable data source for the fused results. This gives better RMSE value than those of the input modalities and other methods. It is visually clear that our approach benefits from both sources as they

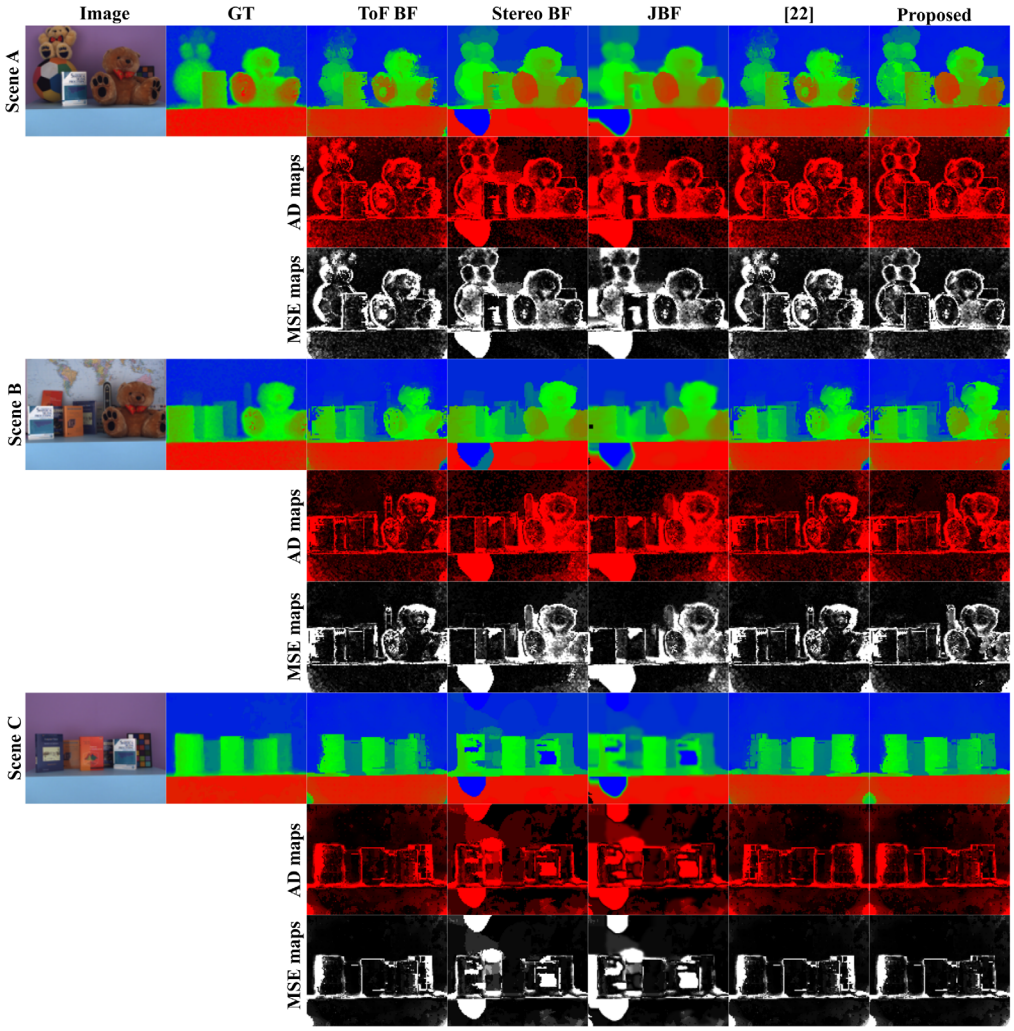

Fig. 4 The input color image is shown in the first column. Also, the ground truth, the input TOF BF, stereo BF, and the results of other methods are demonstrated in other columns. Our proposed fusion map is shown in the final column for all scenes. The AD map is provided in the red color-coded form in the second row for each corresponding column for all scenes. MSE maps are presented in the third row for each corresponding column for all the scenes. The MSE unit is scaled with 10^4.

Table 3 Numerical evaluation of the result with RMSE on the whole image.

Methods	TOF BF	Stereo BF	JBF	Ref. 21	Proposed
Scene A	1.2698	2.4971	2.5449	1.2696	0.9389
Scene B	1.2786	2.4135	2.6467	1.2783	1.0801
Scene C	0.8498	2.6957	0.8498	0.8493	0.8181
Average RMSE	1.1327	2.5354	2.0138	1.1324	0.9457

complement each other, see Fig. 4. The depth map of the one depth modality source in the final fused map improves error-prone areas in the depth map of the other modality.

For instance, the homogeneous area in scene B in front of the table in Fig. 5 shows failure in estimating disparity using the ADC algorithm. In contrast, the TOF sensor did a rather good sampling in this area. Another same example can be observed in scene C in Fig. 5 on the book

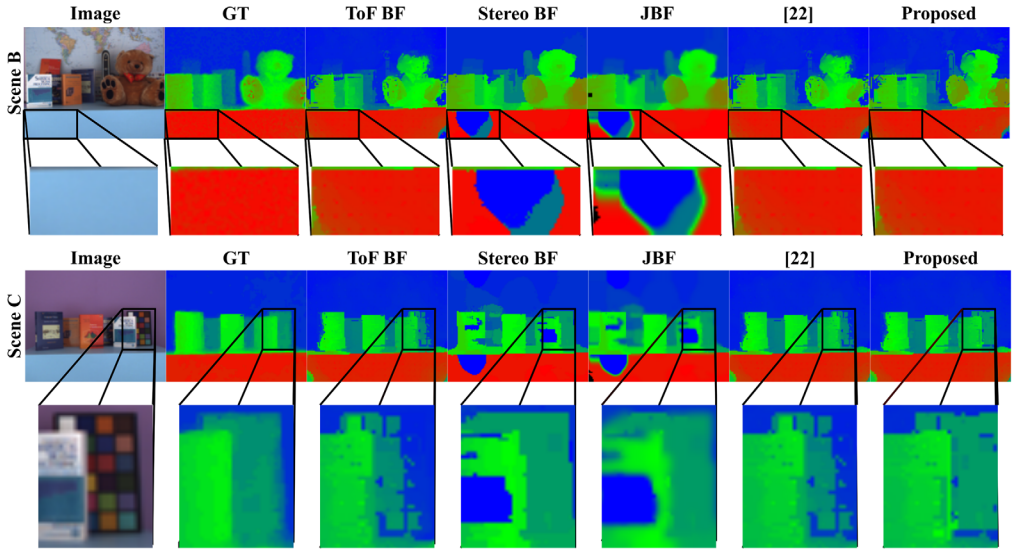

Fig. 5 Results in zoom mode. The first row shows a homogenous part of scene B, in front of the table. The second row shows an occluded area, which includes a color chart as a homogeneous area and the edge of the book in front of it.

that partially occludes the color chart behind it. While the opposite example can be seen on the color chart itself where the TOF samples the disparity poorly, but stereo estimated the disparity rather good. All these examples indicate the aggregation of stereo confidences and TOF reliability weights in the proposed fusion approach. Additionally, these examples explain how this aggregation is managed such that the proposed approach takes advantage of stereo BF over the TOF BF and vice versa based on their accuracy for different areas of the scene.

4.1 *Consistency Check*

There are several applications such as disparity image-based rendering that require consistency between left and right disparity maps. To show that our approach produces consistent left and right fused maps in the stereo setup, we applied a consistency check on the results.

First, we produce fusion maps for both left and right cameras in the stereo system. Next, we check stereo consistency between the two maps and generate consistency maps for both the left and right disparity map (see Fig. 6).

In the consistency check, for each pixel coordinate (x, y), in the left fused map, we add its disparity value to find its correspondence in the right fused map. Then, we compare the disparity value of the left and right correspondences with different thresholds (three) of one to four pixels. If the difference between the disparity value of the left and right correspondences is less or equal than the threshold at pixel coordinate (x, y), we return the disparity value as it is and assign it to

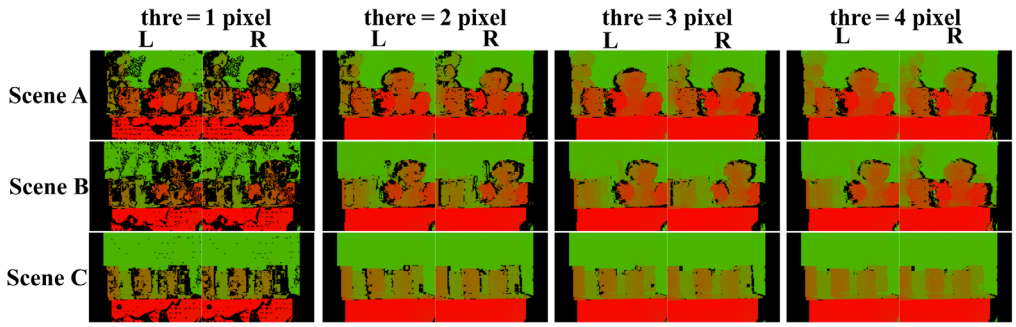

Fig. 6 Color-coded consistency map between the right- and left-fusion maps for three scenes. Each column corresponds to a specific threshold.

the same pixel coordinate in the consistency map. Otherwise, we set a zero value as the disparity value of that pixel coordinate in the consistency map. Figure 6 shows the color-coded consistency maps of different thresholds for all the scenes. The results demonstrate the consistency between left and right fused maps even for a harsh threshold of one pixel.

4.2 *Computational Cost*

The current nonoptimized MATLAB implementation can process each image of the synthetic dataset with a resolution of 540×960 in 572.38 s.

5 Conclusion

This paper introduced an approach for the fusion of TOF and stereo depth maps in a reliable manner. We proposed an approach that is based on a simple optimization framework that utilizes reliability weights and stereo confidences. We introduced stereo constraints into the optimization framework and guided the framework by edge information generated as a result of cross-checking between the RGB image and TOF and stereo disparity. Our approach outperforms state-of-the-art on a synthesized and real dataset both visually and objectively. We examined our result for the left and right consistency and demonstrated the ability of our approach to fusing right and left disparity maps with TOF map in a consistent manner.

The proposed method is straightforward to extend with further depth data from other depth sensing equipment by adding new energy terms in the cost function. Future research consists of identifying new data sources with supplementary accuracy of the depth information in different parts of the image, along with identifying appropriate weighting of those terms such that the respective advantage of the different data is emphasized in the optimization. The proposed method can gain better results in case of incorporating alternative confidence estimation approaches with higher accuracy that could be investigated in the future. The current MATLAB implementation can be optimized to be able to process higher resolutions in shorter computational time.

Acknowledgments

We thank Dr. Marcus Bednara for his guidance at the beginning of this work that greatly helped us with problem understanding. We also would like to thank Dr. Sebatian Schwarz for his assistance that fastens us in reproducing his paper. The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant Agreement No. 676401, European Training Network on Full Parallax Imaging.

References

1. G. Evangelidis, M. Hansard, and R. Horaud, "Fusion of range and stereo data for high-resolution scene-modeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(11), 2178–2192 (2015).
2. S. A. Gudmundsson, H. Aanaes, and R. Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3D estimation," *IJISTA* **5**, 425–433 (2008).
3. G. Marin, P. Zanuttigh, and S. Mattoccia, "Reliable fusion of ToF and stereo depth driven by confidence measures," *Lect. Notes Comput. Sci.* **9911**, 386–401 (2016).
4. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision* **47**, 7–42 (2002).
5. D. Herrera C., J. Kannala, and J. Heikkila, "Joint depth and color camera calibration with distortion correction," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2058–2064 (2012).
6. R. Lange and P. Seitz, "Solid-state time-of-flight range camera," *IEEE J. Quantum Electron.* **37**, 390–397 (2001).
7. B. Buettgen et al., "CCD/CMOS lock-in pixel for range imaging: challenges, limitations and state-of-the-art," Measurement, p. 103 (2005).

8. M. Frank et al., "Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras," *Opt. Eng.* **48**, 013602 (2009).
9. R. Nair et al., "A survey on time-of-flight stereo fusion," *Lect. Notes Comput. Sci.* **8200**, 105–127 (2013).
10. A. Torralba and W. T. Freeman, "Properties and applications of shape recipes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, Madison, Wisconsin, pp. 383–390 (2003).
11. A. Zomet and S. Peleg, "Multi-sensor super-resolution," in *Proc. 6th IEEE Workshop Appl. Comput. Vision*, pp. 27–33 (2002).
12. D. Chan et al., "A noise-aware filter for real-time depth up-sampling," in *Proc. ECCV Workshop Multi-Camera and Multi-Modal Sens. Fusion Algorithms and Appl.* (2008).
13. J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Neural Inf. Process. Syst.*, MIT Press, Cambridge, Massachusetts (2005).
14. J. Dolson et al., "Up-sampling range data in dynamic environments," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, pp. 1141–1148 (2010).
15. Q. Yang et al., "Spatial-depth super-resolution for range images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, pp. 1–8 (2007).
16. J. Kopf et al., "Joint bilateral up-sampling," *ACM Trans. Graphics* **26**, 96 (2007).
17. T. Matsuo, N. Fukushima, and Y. Ishibashi, "Weighted joint bilateral filter with slope depth compensation filter for depth map refinement," in *Proc. Int. Conf. Comput. Vision Theory and Appl.*, Vol. 2, pp. 300–309 (2013).
18. J. Park et al., "High-quality depth map up-sampling and completion for RGB-D cameras," *IEEE Trans. Image Process.* **23**, 5559–5572 (2014).
19. P. Favaro, "Recovering thin structures via nonlocal-means regularization with application to depth from defocus," in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, San Francisco, California, pp. 1133–1140 (2010).
20. G. Pagnutti and P. Zanuttigh, "Joint segmentation of color and depth data based on splitting and merging driven by surface fitting," *Image Vision Comput.* **70**, 21–31 (2018).
21. S. Schwarz, M. Sjöström, and R. Olsson, "A weighted optimization approach to time-of-flight sensor fusion," *IEEE Trans. Image Process.* **23**(1), 214–225 (2014).
22. S. Schwarz, M. Sjöström, and R. Olsson, "Time-of-flight sensor fusion with depth measurement reliability weighting," in *3DTV-Conf. True Vision—Capture, Transmission, and Display 3D Video* (2014).
23. S. Schwarz, M. Sjöström, and R. Olsson, "Temporal consistent depth map upscaling for 3DTV," *Proc. SPIE* **9013**, 901302 (2014).
24. C. Dal Mutto et al., "Locally consistent ToF and stereo data fusion," *Lect. Notes Comput. Sci.* **7583**, 598–607 (2012).
25. H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(2), 328–341 (2008).
26. P. Zanuttigh et al., *Time-of-Flight and Structured Light Depth Cameras: Technology and Applications*, 1st ed., Springer International Publishing (2016).
27. G. Agresti et al., "Deep learning for confidence information in stereo and ToF data fusion," in *IEEE Int. Conf. Comput. Vision Workshops* (2017).
28. B. Huhle et al., "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vision Image Understanding* **114**(12), 1336–1345 (2010).
29. C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo, "Probabilistic ToF and stereo data fusion based on mixed pixels measurement models," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(11), 2260–2272 (2015).
30. X. Mei et al., "On building an accurate stereo matching system on graphics hardware," in *Proc. IEEE Int. Conf. Comput. Vision Workshop*, Barcelona, Spain, pp. 467–474 (2011).
31. R. O. Het Veld et al., "A novel confidence measure for disparity maps by pixel-wise cost function analysis," in *25th IEEE Int. Conf. Image Process.*, Athens, pp. 644–648 (2018).
32. W. H. Steeb, *Problems and Solutions in Introductory and Advanced Matrix Calculus*, World Scientific Publishing Company (2006).
33. R. P. W. Duin, H. Haringa, and R. Zeelen, "Fast percentile filtering," *Pattern Recognit. Lett.* **4**, 269–272 (1986).

Faezeh Sadat Zakeri received her bachelor's degree in computer software engineering from the University of Kashan, Iran. She got her master's degree in image-vision-optics from the University of Saint-Etienne (University de Lyon) and master of computer science degree from the University of Eastern Finland. She was a PhD researcher candidate at Fraunhofer IIS, Germany. Currently, she is continuing her PhD at the chair of Computer Graphics of the University of Tubingen. Her research focuses on 2D/3D computer vision, light-fields imaging, depth reconstruction and fusion, and deep learning.

Mårten Sjöström received his MSc degree in electrical engineering and applied physics from Linköping University, Sweden, in 1992, the Licentiate of Technology degree in signal processing from the KTH Royal Institute of Technology, Stockholm, Sweden, in 1998, and the PhD in modeling of nonlinear systems from the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2001. He was an electrical engineer with the company ABB, Sweden, from 1993 to 1994, and a fellow with the international research organization CERN from 1994 to 1996. In 2001, he joined Mid Sweden University, and he was appointed as an associate professor and a full professor of signal processing in 2008 and 2013, respectively. He has been the head of the computer and system science at Mid Sweden University since 2013. He founded the Realistic 3D Research Group in 2007. His current research interests include multidimensional signal processing and imaging, and system modeling and identification.

Joachim Keinert is heading the group Computational Imaging and Algorithms of the Fraunhofer Institute for Integrated Circuits in Erlangen, Germany. He has diplomas from both the University of Stuttgart, Germany, and Télécom ParisTech (formerly ENST) in Paris, France. In 2009, he received his PhD with honors in the domain of electronic system-level design for image processing applications. His main field of research is on light-field imaging. Beforehand, he also worked on HDR video capture, JPEG 2000, and low-complexity coding.