A Survey on Time-of-Flight Stereo Fusion

Rahul Nair[1,2], Kai Ruhl[3], Frank Lenzen[1,2], Stephan Meister[1,2], Henrik Schäfer[1,2], Christoph S. Garbe[1,2], Martin Eisemann[3], Marcus Magnor[3] and Daniel Kondermann[1,2]

[1]Heidelberg Collaboratory for Image Processing (HCI), Heidelberg University, Germany
[2]Intel Visual Computing Institute, Saarland University, Germany
[3]Technical University Braunschweig, Germany,

September 11, 2013

Abstract

Due to the demand for depth maps of higher quality than possible with a single depth imaging technique today, there has been an increasing interest in the combination of different depth sensors to produce a "super-camera" that is more than the sum of the individual parts. In this survey paper, we give an overview over methods for the fusion of Time-of-Flight (ToF) and passive stereo data as well as applications of the resulting high quality depth maps. Additionally, we provide a tutorial-based introduction to the principles behind ToF stereo fusion and the evaluation criteria used to benchmark these methods.

1 Introduction

Will there ever be one depth sensor to rule them all? While this will hopefully be true one day, all current depth sensing modalities fall short of obtaining this title. Passive stereo works well on textured scenes and has a high lateral resolution due to readily available mega pixel cameras. Conversely, there are issues at occlusion boundaries and when the textures are ambiguous or when no texture is present at all. Also, due to the number of pixels that have to be compared, especially when global optimization techniques are used, stereo matching algorithms are often computationally demanding. Time-of-Flight(ToF) imaging on the other hand delivers images at high frame rates independent of surface texture, but at the cost of a lower resolution and systematic errors. For a more detailed description of Time-of-Flight cameras please refer to [41]. Finally, there is active stereo (e.g. Kinect), which triangulates correspondences between a structured active illumination and a camera. While the effects at occlusion boundaries (shadowing, edge fattening) remain, unstructured surfaces are no longer a problem. This comes at a cost though, as the lateral resolution is now limited by the resolution of the projection system. To summarize, the major drawback of all of these methods is that they usually only work in a limited domain and lack the robustness often required in various application domains. As these modalities often differ in the areas where they excel or fail, it appears natural to combine them to create a "super-sensor".

Depending on the camera systems used different methods ensue. With a single additional camera typically edge information from the high resolution intensity image is used to guide the upsampling of the depth image[50, 35]. In [10], Castañeda et al. present a system using two

Time-of-Flight cameras. In this survey paper we will focus on techniques to fuse of Time-of-Flight and passive stereo data.

The remainder of this paper is organized as follows. In Section 2 we shall further clarify, what we expect from such a fusion system and what use there actually is in having high resolution depth maps. Next, in Section 3 the basic fusion pipeline including common preprocessing steps will be introduced in a tutorial like fashion. As benchmarking such systems is as important as the innovation of new fusion systems we will dedicate Section 4 to common evaluation strategies. Finally, in Section 5 we will summarize the specifics of current fusion systems.

2 Requirements Engineering and Application Domains

2.1 Requirements

We have identified four basic requirements an application can have on a fusion system.

Speed up while retaining Quality Current stereo algorithms are often quite time consuming. This is due to the vast search space that has to be analyzed. Given real time ToF imaging it may now be possible to reduce the search space and therefore make real-time implementations of the stereo methods possible.

Robustness/Self Awareness Fusion methods should be able to be at least as good as the (locally) better of the two modalities and degrade gracefully in presence of small calibration/synchronization errors.

Increase in Quality Other than identifying regions of erroneous values the system should also be able to use this information to produce depth maps that are better than either method alone.

Backward Compatibility In many application areas it is easier to just add an additional camera to the working system than to completely alter the existing system.

It is clear that it is difficult to accommodate for all requirements simultaneously. A speed - quality trade off has always to be made depending on the application. For Human Computer Interaction and Robotic/Navigation applications a fast system that is able to detect and eliminate erroneous values [40, 20] may suffice. More sophisticated multimedia application on the other hand require high quality depth maps partly with speed constraints imposed on the system (e.g. in 3DTV, Augmented Reality). As other application domains for depth data have been discussed in [47], we will focus on the application of high quality depth maps in multimedia systems.

2.2 High Quality Depth maps for Multimedia Application

In movie and film productions many post-production steps are commonly conducted including color corrections,(green-screen) matting, integration of computer generated imagery, compositing and many more. High-resolution depth or disparity maps can help to ease many of these steps. As edges in the depth maps depict object boundaries in general, they can be used to guide local color corrections, in the spirit of cross-bilateral filters [18, 51]. Integration of virtual objects is possible with correctly handled occlusion [12]. For stereoscopic movie productions the aforementioned tasks become even more important due to additional challenges including color matching between the stereoscopic views, vertical alignment, disparity compensation, 3D compositing and image interpolation. To faithfully deal with these tasks, correspondences between the left and right image in the form of disparity or depth maps build the foundations of all these algorithms.

Figure 1: From left to right: Stereo-ToF rig on set, example image, high resolution disparity map

Depth maps as a form of 2.5D scene representation ease the integration of computer generated imagery or video footage with depth information [43]. Precise depth maps also allow for image interpolation [70] which in turn can be used for disparity compensation [17]. To prevent a flickering appearance in stereoscopic video footage, appropriate local color corrections are necessary that consistently correct for color mismatches in both views [19]. The problem is even more difficult for specularities, here the solution is usually to replace the specular parts in the image by information from the other view [62] or to synthesize a consistent specularity for both views [59].

Not only post-production but also display and transmission of stereoscopic content requires high-resolution and high quality depth. In depth-image-based rendering the video stream consists of the typical RGB images plus an additional depth channel [45]. From this information the stereoscopic views are recreated by warping of the RGB image based on the depth and desired ocular distance.

As most modern production settings already employ several cameras (cf. Figure 1), the idea of using an additional ToF camera lends itself to assist the depth map generation. It should be noted though that algorithms intended to work in such a setting require a higher amount of robustness as compared to the lab setting. At any given time many different people are (to a certain extent) independently monitoring several different aspects of the scene such as lighting, camera movement, focus of camera or the stereo baseline, often changing parameters frequently to accommodate for the requirements of the director. Also the time plan is quite strict, such that any additional in-between calibration steps need to be avoided. So if the ToF-Stereo setup is to be attached to the principal camera it will be more difficult to obtain high precision measurements and alignment than what is common for a lab setting. Therefore, robustness of the algorithms, especially towards slightly misaligned cameras, is extremely important. On the other hand, often post-production crews acquire their own footage of the scene separately beforehand, so that they can start working on set reconstructions etc. before receiving the main plates. In Section 5.3 two methods will be discussed in detail that cater to these different settings.

3 Setting up Fusion Systems

In the following we describe the general aspects of ToF stereo fusion systems. These include the general pipeline (Section 3.1), possible camera setups (Section 3.2), calibration (Section 3.3) and data preprocessing (Section 3.4).

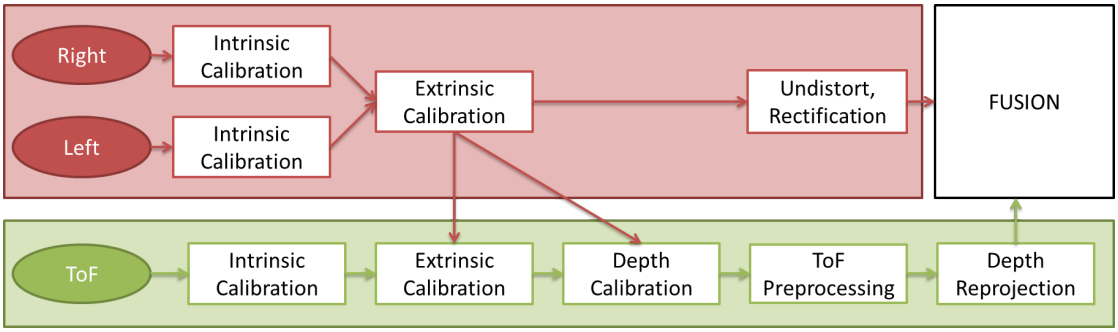


Figure 2: Basic Fusion Pipeline

3.1 Pipeline

Most fusion systems differ mainly in how the data is merged, once it has been brought into the same reference frame. Figure 2 illustrates the basic pipeline employed. After choosing a specific camera setup the standard camera intrinsics have to be estimated for all three cameras, i.e. focal length, principal point as well as radial and tangential distortion coefficients. Next the spatial relationship(Roto-Translation) between the three cameras have to be found by means of pairwise or joint stereo calibration methods.

For the ToF camera, additionally a depth calibration has to be undertaken due to the systematic errors described in [41, Section 2]. This can either be done after the intrinsic calibration with methods proposed in [41, Section 3] in or by jointly using the additional stereo information. After applying preprocessing steps to clean up the ToF data (i.e. to reduce effects by noise pixels), the images must be brought into the same coordinate frame by means of rectification and reprojection. Finally, the fusion step involves a combination of the following:

- The ToF depth and the output of a Stereo algorithm are computed individually and then fused.
- The ToF data is used as an initial guess and/or to reduce the search space for subsequent stereo refinements
- The depth reconstruction algorithm uses both stereo and ToF costs as data terms.

Additionally various regularizers have been applied to obtain depth maps of sufficient smoothness despite noise. In the following we will describe the steps commonly employed by the methods presented in Section 5 before the data is fused in detail.

3.1.1 Choice of Depth Cues

In essence a ToF stereo fusion system corresponds to a trifocal camera system, with the third camera sensor having a lower spatial resolution, but a high temporal sampling. Therefore, there exist many different sources of dense or sparse depth information that may be exploited in a fusion system. In the following we would like to discuss these depth cues in detail. Though some of the cues are rarely used or not used at all, we believe that future algorithms may additionally use these modalities.

ToF depth from demodulation This is the standard output of the Time-of-Flight sensor that is used in all fusion systems. The advantages of these modalities as opposed to stereo is that

the depth estimation *a*) works on textureless surfaces, *b*) has a arguably simpler behavior at depth discontinuities (unlike edge fattening in stereo) *c*) is real-time capable out of the box. Major downsides are the limited lateral resolution and the various strong error sources such as noise, multi path, flying pixels, wiggling and to a certain extent susceptibility to background illumination.

A detailed description of these errors and methods to compensate for them are given in [41].

Photo consistency/Stereo Depth from stereo is a well studied field of research [54]. In stereo depth estimation, dense correspondences between left and right view are found and the depth is inferred via triangulation. Unlike ToF cameras, the lateral resolution of this modality can be very high. Depth from stereo will fail in areas with little texture or in presence of highly repetitive patterns due to ambiguous matching. While this imposes a problem for fast local-evidence based stereo methods, global methods use regularization techniques to utilize prior knowledge about "normal" scenes such as temporal coherence. It should also be noted that with the large (Full HD, 4K) images commonly used for multimedia applications such global techniques often reach their limit in terms of computational cost without any search space reduction.

Cross-modal Stereo While all current systems only use the photo consistency constraints between the stereo heads, additional information is available via the intensity image of the ToF camera. Unfortunately, due to the difference in resolution and wavelength sensitivity traditional photo consistency can not be used here. A promising line of research is cross-modal stereo [13], also known as IR/Thermal image-RGB registration [34, 4, 60], which tries to find correlations between the near or far IR with RGB/Intensity image to either infer depth as in the former case or find a warp as in the later case.

Structure from Motion If the fusion system is moving and the scene mostly static, it is possible to use structure from motion (SFM) techniques [63, 48] to additionally infer depth at some locations. Here, it has to be ensured that the synchronization is sufficiently accurate or that the fusion system is capable of handling slight misalignments robustly.

Monoscopic Cues Lighting, Shading and Silhouettes are mostly used in monoscopic depth estimation [33, 1]. Shape from Shading with unconstrained lighting is yet a difficult problem. But for the ToF camera, as the primary light source is around the camera, this could still be feasible and should be investigated. Indeed, Stürmer et al. [57] observed that the amplitude image in observed an inverse square falloff with distance. Finally, silhouettes constrain the direction of normals of the depth map to be perpendicular to the pixel ray.

Current fusion systems typically only use the ToF as a black box depth imager and the photo consistency constraint between the stereo heads. A notable exception is the method by Kim et al. [37] that uses additional silhouette constraints and the technique by Zhu et al. [67] that uses an optical flow based temporal smoothing (though no SFM information is used here).

3.2 Camera Setup

The camera setup employed should suit the requirements of the application and additionally aim to reduce the effects of visible errors due to alignment issues. Figure 3.2 illustrates various common camera configurations, though naturally many more are possible.

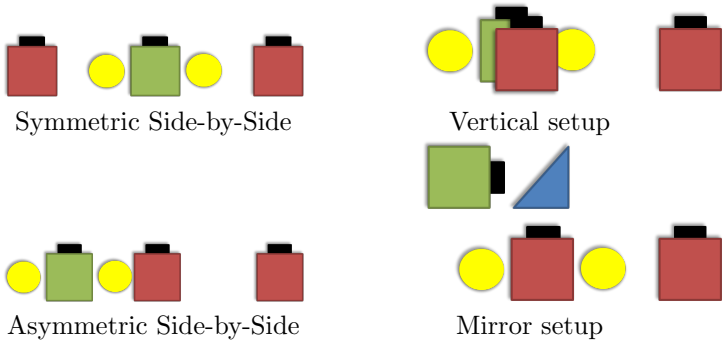


Figure 3: Different possible camera setups. Green: ToF Camera, Red: RGB Camera, Yellow: ToF Lights

Symmetric Side-by-Side This is the approach most commonly employed by most fusion systems with the stereo heads symmetrically placed left and right of the ToF camera. Though this approach seems to be the best on first instance, it depends on whether the paralax between the ToF camera and the stereo heads is actually being used to additionally infer depth. If the ToF data is going to be reprojected on to one of the stereo frames then more information is lost than in the assymmetric setup.

Asymmetric Side-by-Side This approach tries to compensate such parallax effects between the ToF and Stereo heads by placing the ToF camera closer to the primary camera. Depending on whether the ToF imager or one of the stereo heads is the primary camera, also different setups may make sense. In the latter case the ToF camera may additionally be placed on the other side of the primary camera to ensure available depth information, otherwise not obtainable due to occlusion between the stereo heads.

Vertical Setup The vertical setup mostly corresponds to the asymmetric side-by-side setup. It is employed for the same reasons as above in situations, where placing the camera next to the primary camera is not feasible, such as in stereo-production rigs (cf. Section 1) which are often huge in size and would induce a bigger parallax in a side-by-side setup.

Mirror Rig Ideally - if only ToF and passive stereo are used, there shouldn't be any parallax between the primary and ToF camera. This is achievable using a beam-splitting mirror/prism (commonly known as hot or cold mirrors) and sharing the same optical axis. The center for sensor systems (ZESS) in Siegen has produced such a prototype system[23]. The Arri group, manufacturers of production grade film cameras, have recently introduced another RGB-Z camera that works on the same principal. It should be noted though that such a setup still requires manual alignment of the mirror and cameras to actually achieve zero-parallax and in practice this may be difficult to achieve.

3.3 Calibration

Intrinsics and Extrinsics Once the hardware is set up the system needs to be calibrated intrinsically for focus, principal point and distortion coefficients and the ToF camera additionally for depth. The extrinsic calibration is concerned with finding the Roto-Translation between the three cameras. Both procedures are straight forward for the stereo heads and can be done using standard libraries [8, 6]. In our experience, the same methods often work

on the intensity image (with parameter tweaking) for ToF imagers with higher resolutions. More details on calibration can be found in [41].

Joint Calibration The extrinsic calibration between the ToF camera and the stereo heads using the standard method will be less precice due to the lower resolution of the ToF camera. Some fusion techniques account for this by simply adjusting uncertainties that they utilize during fusion [27, 46]. If the ToF depth calibration has already been obtained, DalMutto et al. [14] suggest using the depth information and planar calibration targets to obtain a precise extrinsic calibration. Schiller et al.[55] jointly do the intrinsic ToF calibration with the extrinsic calibration, as the depth estimates delivered by the ToF and stereo will not only be consistent but the ToF calibration can also be achieved more precisely. Similar methods were also proposed in [68, 38] and can be summarized as follows:

1. Obtain pictures of planar calibration targets via a (calibrated) stereo setup and the ToF image.
2. Fit a Plane into this target via the triangulated target points in the stereo setup. To obtain dense stereo "ground truth points".
3. If the extrinsic calibration has not been estimated yet, use Horn's method [32] to find the transform between the stereo plane and ToF plane.
4. Finally, store the residuals between ToF depth and the plane for the ToF depth correction. This can be done in form of a 4D look-up table or by fitting a polynomial spline per pixel.

Finally, Guan et al. [25] use spherical targets that are detected in the RGB and ToF imagers.

3.4 Preprocessing

Stereo Rectification Stereo rectification [29] reduces the search space to a line search along one image dimension by finding two homographies such that the epipolar lines between the two stereo heads become parallel.

Depth Reprojection ToF delivers radial depth which has to be converted into z-depth before comparing with the stereo depth. Given the focal length f and centralized pixel coordinates p_x, p_y (i.e. principal point in $(0, 0)$) and radial depth d the coordinates (X, Y, Z) can be computed via:

$$(X, Y, Z)^T = (p_x, p_y, f)^T \cdot \frac{d}{(f^2 + p_x^2 + p_y^2)}. \tag{1}$$

These points can then be rotated and translated into the reference coordinate frame. If a dense ToF depth map is required the values for reference frame pixels without a corresponding ToF pixel have to be interpolated. Finally, the z-depth z can be converted into disparities $disp$ using the baseline b:

$$disp = \frac{b \cdot f}{z}. \tag{2}$$

Depth Preprocessing The depth data may be additionally filtered before reprojection to avoid false occlusions due to noise. This ranges from simple median filtering to remove flying pixels to more complex denoising techniques as presented in [42].

4 Evaluation of Fusion Methods

In this section we will discuss various evaluation datasets and performance metrics to benchmark fusion algorithms. Additionally, based on the requirements discussed in Section 2 we will propose some new experiments and performance measures that we believe will help in a better understanding of the fusion system.

4.1 Datasets

4.1.1 Available Stereo-ToF Datasets[1]

Currently, very few ground truth datasets for ToF stereo fusion are actually available. Nair et al. [46] used the HCI Box[2] for quantitative evaluations. The target consists of a box with various geometric primivitives that was hand measured to 1mm accuracy and aligned to PMD[Tec] CamCube 3 data. It contains little texture and shows strong multi-path effects on the box sides. The Padua[3] datasets introduced by Dal Mutto et al.[14, 16] contain simple synthetic scenes as well as measured tabletop scenes containing a varied amount of textured objects. The reference data was obtained using space time stereo[66] and aligned with ToF data from a MESA SR4000.

4.1.2 Semi-Synthetic GT

Since ToF stereo fusion ground truth datasets are not as readily available as datasets for assessing ToF or stereo alone, authors often resort to use existing datasets, by simulating the missing modality.

Synthesizing the ToF Image Often the Middlebury ground truth dataset [54] is used [16, 64] and the ToF view is synthesized from the ground truth data. Though an interesting way to compare the results, the naive implementation currently used is to just downsample the GT depth and add some noise to the obtained depth map.

This approach does not account for a) the different camera positions and b) the complex noise behavior of ToF cameras. We therefore believe that it can be improved in two important aspects.

- **Alignment** The effects due to the ToF and the reference camera not sharing the same optical axis are completely ignored in this simple approach. This is fine as a baseline evaluation to isolate alignment effects from the fusion part or if a mirror rig is used. Otherwise, the depth map should be first synthesized in the ToF view before warping the data back, possibly adding alignment noise.
- **Simulation** Also some care should be taken into properly simulating the ToF sensor. Evaluation using the GT without any noise can only be used as a proof of concept. We suggest to use one of the simulators described in [47, Section 3.2].

Synthesizing Stereo Similarly, if a ToF GT dataset is available where the reference data has been obtained including RGB/Intensity information such as the datasets in [52] and the HCI- Laser scanning dataset (cf. Section 4.2 in [47]), this can be used to synthesize additional views. If only a GT depth map is available occlusion effects need to be handled consistently.

[1] Up to date list: http://hci.iwr.uni-heidelberg.de/Benchmarks/document/tofstereo
[2] http://hci.iwr.uni-heidelberg.de/Benchmarks/document/hcibox/
[3] http://freia.dei.unipd.it/nuovo/research/ToF.html

4.2 Performance Measures

As ToF stereo fusion aims at finding high quality 3D reconstructions, the same evaluation criteria that are discussed in [47, Section 5] can be employed. Here, we will give an overview of the performance criteria reported in the ToF- Stereo fusion literature and propose some performance criteria specifically for ToF stereo fusion we deem useful.

4.2.1 Used Measures

Accuracy and precision Conventional depth measuring approaches such as laser scanning always state precision (variance of measurement) and accuracy (systematic bias between GT and measurement). Assuming independently and identically Gaussian distributed errors in each pixel, then mean and standard deviation of the signed error would correspond to these measures. As many real life distributions often have heavy tails, skewing or more than one mode, robust statistics such as median and interquartile range should be used. Finally, as there often is a strong correlation between error and external factors such as viewing angle or texturedness, such scalar error metrics may not give the complete insight into the behavior. Therefore, wherever possible we suggest to additionally supply either the complete (1D) error distribution, or even the error images [46].

Mean squared error, median absolute error In fusion literature [69, 16] often the mean squared error is reported instead of accuracy and precision. For real valued functions this corresponds to the sum of variance and squared bias. Again, due to the inherent quadratic weighting of large errors a better metric would be the median absolute error instead. The same arguments against the scalarization of the error as above apply here as well.

Application specific evaluation For many applications geometry reconstruction is not the final goal but just a intermediate step. Song et al. [56] evaluate the edge quality by comparing the obtained depth edges with pre-labeled silhouette boundaries in a plant phenotyping application (cf. Section 5.2 in [47]). This is not only interesting for plant phenotyping, where the leaf silhouettes have to be extracted reliably, but also in multimedia applications where the location and shape of silhouettes are of vital importance. Zhu et al.[68, 69] analyze the deviation between a box model fitted into the depth data and the GT box by analyzing the angular deviation of the three observed box sides from 90 degrees. Finally, for view synthesis, the quality criteria is the credibility of the synthesized view. This evaluation could be achieved having an additional camera capturing the scene and comparing a synthesized view with the real view.

Eyeballing The evaluation of ToF stereo fusion methods is still largely qualitative in nature due to the lack of sufficient ground truth datasets. For certain applications (e.g. visual effects) the users can often judge best, how useful the algorithm results are to them. This process, also called eyeballing, requires many different scenes to be visually inspected by one or more independent expert users. While all proposed methods show qualitative results, a proper user study has yet to be undertaken.

4.2.2 Proposed Measures/Experiments

Graceful degradation - Alignment As spatial and temporal alignment (i.e. extrinsic calibration and synchronous triggering) is a big issue one possible quality criterion is the robustness towards misalignments. We propose the following experiments to assess this.

First a calibrated dataset using the standard setup is captured. Fusion results for a spatially misaligned setup are then generated by artificially varying the calibration between ToF and the stereo setup. Temporal alignment can either be evaluated by capturing the stereo data in a higher framerate than the ToF images or by interpolating in between frames.

Speed vs. Quality One claim that all Fusion papers make is that using ToF data speeds up computation considerably compared to a baseline and many authors also state the running times of their algorithms. Additionally, an assessment of execution time (number of iterations, change of search range, etc.) vs. quality improvement could be made. While it is clear that the speed of algorithm execution depends heavily on the implementation platform, hardware and implementation details, we think that a speed over quality assessment of the algorithms is still necessary. Quality can mean any quality criteria from endpoint error to precision of edges.

Effect of Fusion The final claim that many fusion algorithms make is that the depth maps obtained is better than either depth map alone. The question that remains is how much better is the algorithm? And how does the scene composition affect this performance. A fusion algorithm should at least be as good as the better of the two modalities irrelevant of scene composition.

5 Overview of Fusion Methods

Following [54] we will group the fusion methods based on the optimization strategy that is employed. *Local methods* [40, 26, 2, 28, 15, 64, 46, 3] tend to be faster and parallelizable but cannot cope with locally erroneous data. They are often based on a line search that is guided by the ToF data. *Global methods*, [20, 27, 68, 67, 69, 37, 46, 53, 22, 16, 56] add the ToF information as an additional data term in a global energy functional is then jointly optimized. While the depth maps obtained are smoother due to the usage of prior information/regularizers, this is at the cost of additional computational resources. In this overview, we will further group the global techniques depending the framework that was chosen for optimization. While [27, 68, 67, 69, 56] employ different *graphical models* for inference, [53, 46] formulate the problem in a *variational* framework. The last sub-group of the global methods[20, 37, 22, 16] contains those which use *other non-local* optimization strategies.

After a discussion of commonalities in each group, we will proceed to describe each method in detail. The description will start from the point we left in Section 3 – that is after all data have been brought to the same reference frame and after all preprocessing has occurred – except for some special kinds of preprocessing not already mentioned in Section 3. The notation used in the following is summarized in Table 5. Please note that some algorithms work in the disparity (d) space while others operate in the depth (z) space. This doesn't impose any additional contraint as one representation can be converted into the other using the extrinsic calibration and Eq. 2.

5.1 Local Methods

The methods presented here have in common that the basic optimization employed only takes a local sets of pixel values are taken into account. Note, that the aggregation over support windows, whenever applied, make implicit assumptions (e.g. piecewise planar, fronto parallel patches) on surface regularity.

i, j	pixel location
$i \in \Omega$	pixel i in image domain Ω
$j \in N_i$	pixels j in neighborhood of i
$\{x_i\} = \mathbf{x}$	Value x at pixel i collectively referred to as $\mathbf{x}$
ste, ToF	stereo, ToF
L, R	Left/Right image
$\hat{\mathbf{x}}, \hat{x}_i$	Final estimate for $\mathbf{x}, x_i$
$\mathbf{d^{ToF}}, \mathbf{z^{ToF}}$	Disparity, Z-depth from ToF (as converted using Sec. 3.4)
$\mathbf{d^{ste}}, \mathbf{z^{ste}}$	Disparity, Z-depth from stereo
$\mathbf{A^{ToF}}, \mathbf{I^{ToF}}$	Amplitude, Intensity
$\mathbf{I^L}, \mathbf{I^R}$	Intensity in Left/Right image
$E(x), E_i(x)$	Objective energy to be minimized (at location i)
$R(x)$	Regularizer
c	Confidence / Weights
$\mathbf{1}_{ToF}(z), \mathbf{1}_{ste}(z)$	Range indicator functions for ToF/stereo
$\chi_{ToF}(x), \chi_{ste}(x)$	Spatial indicator functions for valid/trusted ToF/stereo
$\gamma_1, \gamma_2 \ldots$	User Parameters

Table 1: General notation used

Kuhnert et al. 2006 [40] Kuhnert and Stommel proposed the first ToF stereo fusion algorithm in 2006. Unlike many methods that incorporate the ToF data into the stereo matching, this methods first independently computes depth maps and uncertainties for ToF and for stereo and then fuses both data sources (see also [16]). The stereo algorithm (Winner Takes All [54]) is only applied to confident regions, using a thresholded Sobel operator response to obtain a binary confidence map. Then, for each data source per pixel ranges are estimated for ToF by adding and subtracting 2 sigma of the previously measured noise. Using indicator functions $\mathbf{1}_{ToF}$ and $\mathbf{1}_{ste}$ for the depth ranges, the fused depth is then given as.

$$\hat{z}_i = \int_0^\infty z_i \mathbf{1}_{ToF}(z_i) \mathbf{1}_{ste}(z_i) dz_i. \tag{3}$$

This amounts to choosing the mid point of the depth range where the two ranges from ToF and stereo overlap. Otherwise the depth is set to 0 (invalid).

Beder et al. 2007 [3] Beder et al. derive a closed form solution to estimating patch orientation based on ToF and Stereo data. The patchlet is initialized with the ToF depth data. Next, by deriving analytical formulas for the gradient direction the patch orientation is optimized using stereo and ToF data. Beder et al. also give a thorough analysis on planar wall scenes as ground truth.

Gudmundsson et al. 2008 [26] applies a hierarchical stereo matching algorithm directly on the remapped ToF depth data. The reprojected depth is input into the 4th coursest level of a hierarchical stereo matching algorithm by van Meerbergen et al. [61] (see also [53]).

Hahne et al. 2009 [28] First, a binary confidence map is obtained by thresholding the amplitude image. The depth data in unconfident areas are discarded and the holes filled via linear scan line interpolation. Next, only the unconfident regions are then further refined via correlation based block matching. The support window shape that is used guided by a watershed segmentation of the color image. The segmentation is seeded using an eroded

version of the confident and unconfident regions.

$$\hat{d}_i = \begin{cases} \underset{d_i}{\text{argmin}} E_{\text{ste}}(d_i) & \text{if } A_i^{ToF} < \gamma \\ d_i^{ToF} & \text{otherwise} \end{cases} . \tag{4}$$

DalMutto et al. 2010 [15] The technique is build around a confidence-based matching in a probabilistic framework. It computes pixel wise probabilities of ToF and probability of stereo in a cost volume. The ToF probability is assumed to be a Gaussian centered around the ToF depth. The stereo probability is given by the truncated absolute difference. The energy resembles the one used in Eq. 8 without the regularizing terms.

Bartczak et al. 2009 [2] Bartczak et al. propose an iterative line search based fusion scheme. After each iteration of the algorithm the obtained depth map is fed back into the matching score in order to enforce local minima.

The local matching cost after the nth iteration is given by

$$E^n(d_i) = \frac{\sum\limits_{j \in N_i} w_j(d) \left(E_{PC}(d_i) + \sum\limits_{k<n} E_D(d_i|\mathbf{d}^k) \right) / (N+1)}{\sum\limits_{j \in N_i} w_{i,j}(d_i)} \tag{5}$$

with $\mathbf{d}^0 = \mathbf{d}^{ToF}, \mathbf{d}^k = \underset{\mathbf{d}}{\text{argmin}} \left(\mathbf{E}^k(\mathbf{d}) \right)$. In Eq. 5 E_{PC} is the photo consistency cost based on truncated L_1 cost weighted and offset by a confidence in the cost given by the min max range of the cost function. The pixel-wise depth contribution E_D is a truncated L_2 cost. Finally, the weights used for aggregation are given as the product of normal distributions centered around center pixel *a)* color *b)* location and *c)* photo consistency.

Yang et al. 2010 [64] The approach by Yang et al. is based on plane-sweeping stereo [21]. As a preprocessing the technique employs a fast RGB-assisted bilateral filter. The energy being minimized is

$$E(z_i) = cE_{ToF}(z_i|z_i^{ToF}) + (1-c)E_{ste}(z). \tag{6}$$

with E_{ToF} being modeled as a truncated quadratic loss between the depth and the ToF depth. E_{Stereo} corresponds to the plane sweeping cost based on the sum of square (SSD) distance per pixel costs. The confidence c used for matching is given as

$$c = \frac{(1-c_{ste})c_{ToF}}{((1-c_{ste})c_{ToF}) + (1-c_{ToF})c_{ste})}. \tag{7}$$

Here, c_{ste} is the stereo confidence given as the likelihood of the current matching assuming a Gaussian distribution of matching costs in the aggregation window centered around the center pixel cost and c_{ToF} is the ToF confidence, a Gaussian with the amplitude image used as standard deviation.

5.2 Graphical Models

Graphical models have frequently been used in the past to solve the stereo matching problem[39, 7, 58]. Here the problem of correspondence estimation is treated as a labeling problem, where each discrete label corresponds to a disparity value. The energy is interpreted as the negative logarithm of a joint probability distribution defined on a graph, where each node corresponds

to an observed (data term) or latent (depth) random variables the probabilities are defined on cliques of these graphs. Though continuous extensions of graphical models do exist[36] the methods presented here still operate on a discrete domain and differ in how the graph is defined as well as the optimization method used for inference.

Zhu et al. 2008, 2010, 2011 [68, 67, 69] In a series of publications starting with[68] Zhu formulates the problem in a Maximum a priori-MRF framework. In [69] the adaptive weight terms are added and finally in [67] a temporal smoothing term is added. The graph structure represented by the energy functional is given by a temporally layered graph. Each layer represents a normally 4 connected pixel neighborhood graph. The connections between layers are given by an optical flow estimate

$$\begin{aligned} E(d) &= c_{stereo}E_{stereo} \\ &+ c_{ToF}E_{ToF}(d|d_{ToF}) \\ &+ R_{smooth} \\ &+ R_{temp} \end{aligned} \tag{8}$$

with $E_{ToF}(d|d_{ToF})$ being a function of the truncated L_1 distance between the estimated disparity and ToF disparity and $E_{stereo}(d)$ based on the Birchfield and Tomasi matching cost [5]. The spatial smoothness term R_{smooth} is a truncated quadratic penalization. Finally, the temporal regularization R_{temp} is given by the complete cost without the temporal term for the previous and next frame. The weights are set according to the confidence in each point. Stereo confidence is the peak to peak ratio of the cost function. ToF reliability is given by a normal distribution (cf. [15]). Optimization is done using Loopy Belief Propagation.

Hahne et al. 2008 [27] The approach by Hahne et al. is based on Graph Cuts and regularizes the first order Total Variation (TV) of the reconstructed surface. The graph is defined on the cost volume with the optimal cut between foreground and background nodes being the desired surface. Each voxel is associated with an consistency edge in z direction. Additionally, smoothing edges connect the nodes in x and y direction. The nodes themselves reside in between voxels. The energy considered is

$$E(z) = \sum_i (E_{fused}(z) + c_{fused,x}\partial_x z + c_{fused,y}\partial_y z) \tag{9}$$

where E_{fused} is a linear combination of photo consistency and truncated quadratic cost for the ToF and the smoothing weights c_{fused} determined by a linear combination of the color difference in the primary stereo image and the difference of median depth of the ToF output. Note that the variable z corresponds to the edges in z direction that are chosen in the cut.

Song et al. 2011 [56] Song et al. use a the standard graph cut stereo approach [7]. Unlike the previous approach the graph is defined over the image grid using multiple labels. Inference is done using alpha expansion. The Time of Flight data is used to reduce the label space in each graph node.

5.3 Variational Fusion

Different to fusion approaches based graphical models as considered in the previous section, *variational* fusion approaches consider both a continuous image domain $\Omega \subset \mathbb{R}^2$ and continuous variables (functions), indicated by a dependency on the image coordinates $x \in \mathbb{R}^2$.

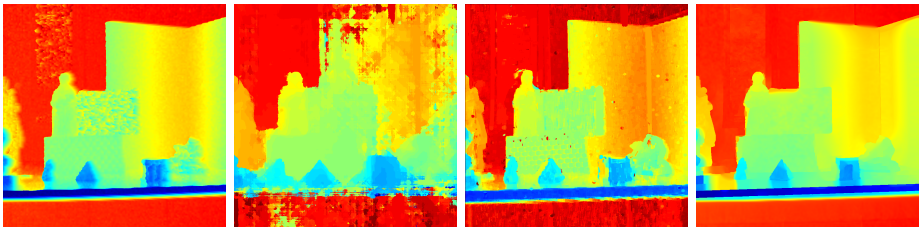


Figure 4: Nair et al. – comparison of ToF only, stereo with semi global matching (SGM) [31], local and variational fusion. We observe that the local approach gives a rough estimate of the disparity, which still compares favorably with SGM. The variational fusion approach provides the most regular result.

We start with a brief overview of a general variational framework, before discussing in detail two recently proposed ToF stereo fusions approaches [46, 53], which are based on this framework. Since one of these approaches assumes unsynchronized data, the restriction to solely horizontal correspondences between the (rectified) stereo images I^L and I can not be applied. We therefore describe this correspondence in terms of an optical flow field $u = (u_x, u_y)^\top : \Omega \to \Omega$, also referred to as displacement field.
We recall the general form of an variational approach given as

$$E(u) := E_{data}(u) + \lambda R(u), \tag{10}$$

to be minimized w.r.t. u, where $E_{data}(u)$ is the data term, $R(u)$ is a regularization term and $\lambda > 0$ is a regularization parameter. A standard data term for optical flow based on the linearized brightness constancy assumption [9, 65] is

$$E_{data}(u) := \|\rho(u)\|_{L^1} := \int_\Omega \rho(u(x))\, dx, \tag{11}$$

where

$$\rho(u(x)) := \left|\mathbb{I}^{\mathbb{L}}(x + u_0(x)) + \langle \nabla \mathbb{I}^{\mathbb{L}}(x + u_0(x)), u(x) - u_0(x)\rangle - \mathbb{I}^{\mathbb{R}}(x)\right| \tag{12}$$

with some approximation u_0 of u. The above framwork is typically used in combination with a coarse-to-fine multi-scale approach (image pyramid), see e.g. [65], where u_0 is updated on each scale. The two fusion approaches considered below differ to this standard form in the way how additional information on the image correspondence from a different modality is introduced into this framework and how the initial approximation u_0 is obtained.

Nair et al. 2012 [46] Nair et al. consider a synchronized camera setup which allows rectification of the stereo images. As a consequence the displacement field can be assumed to be horizontal ($u_x = d$ with disparity d, $u_y = 0$).

The proposed approach consist of two stages, which both make use of confidence measures to determine regions where the ToF or the stereo data might be corrupted. These confidence measures cover problems with low signal intensity and flying pixels for the ToF data, and regions with weak textures and occlusions in the stereo data. A detailed review of the exact definition of these measures is out of the scope of this section; instead we refer the reader to [46].

The first stage of the proposed approach consists in a local fusion using block matching combined with these fidelity measures. Later on, in the second stage, the result of the local

(a) (b) (c)

Figure 5: Ruhl et al. – input data example: approximate depth data from multiple unsynchronized Kinects to be used as uncertain prior to image correspondence estimation. (a) HD camera image, (b) VGA depth map (invalid depth data marked in red), (c) depth points projected into world space.

method is used as initialization. Alternatively, the local fusion approach can serve as a stand-alone method with low numerical costs.

To improve the result of the local fusion approach in a second stage, a modification of the variational framework in Eq. 10 is considered, where the standard data term is replaced by

$$E_{data}(u) := \int_{\Omega} \chi_{ToF}(x)\rho_{ToF}(u(x)) + \chi_{ste}(x)\rho_{ste}(u(x))\, dx \tag{13}$$

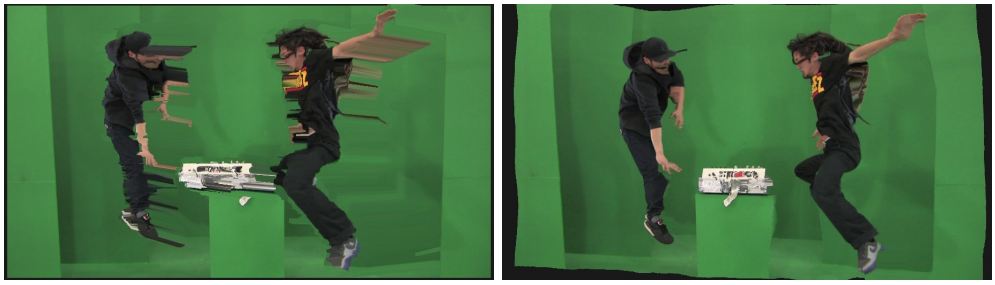
with two local terms $\rho_{ToF}(u)$ and $\rho_{ste}(u)$ penalizing the deviation from the ToF and stereo data, respectively. (We refer the reader to [46] for the exact definition of these two terms.) The aforementioned confidence measures are used to determine locally which of the two data modalities is preferable to the other by defining the indicator functions χ_{ToF} and χ_{ste} in Eq. 13. Thus, the individual data terms are active only in the corresponding image regions. As regularization term an adaptive approach based on first- and second-order total variation is used.

We refer to Fig. 4 for a comparison of the results from both stages.

Ruhl et al. 2012 [53] The authors consider a fusion system, which focuses on settings with unsynchronized cameras. Such a setting complicates reconstruction as typical algorithms require input data captured at the same time instance. In particular, here, the images correspondences do not only depend on the camera geometry, but also on a change of the scene between the individual image recordings. As a consequence, the correspondences in general can not be assumed to be horizontal. The approach makes use of a given depth proxy to guide an image correspondence algorithm that establishes the necessary connections between the input RGB images. The proposed method is not restricted to ToF, as the depth data can be obtained with any available method, but it can be used directly in a ToF stereo fusion setting.

The two stage approach can be briefly summarized as follows:

First stage: Different alternatives are considered to obtain a prior for the stereo correspondence. One alternative is to use depth sensors such as ToF or Kinect (cf. Fig. 5). The second one is to use very coarse, manually modeled geometric proxies, which are e.g.



(a) (b)

Figure 6: Ruhl et al. – stage 1 vs. 2: Approximate prior vs. estimation guided by approximate prior. (a) source image warped directly by approximate prior (b) source image warped after dense image correspondence estimation guided by the approximate prior. The large-displacement, occlusion and low-texture matching properties have been preserved while detail errors are much less present.

a common byproduct of visual media productions. In both cases, the core idea is, after assuming fully calibrated camera systems, to project the 3D world coordinates from the scene into image planes of the stereo cameras to obtain (possibly sparse) correspondences that ideally map to a disparity field $u(x)$, but may also deviate from epipolar geometry to some extent.

Second stage: The variational framework introduced in Eq. 10 is used with the data term defined as in Eq. 11 and using total variation regularization. The correspondence prior $\tilde{u}(x)$ from the first stage enters the approach in the interpolation phase, when the initialization u_0 for the next finer step of the image pyramid is set up. Values of $u_0(x)$ from the coarser level are replaced by the values of prior $\tilde{u}(x)$, if the employed confidence measure allows it.

We refer to Fig. 6 for an example comparing direct application of a depth-based prior $\tilde{u}$ against the results of a dense image correspondence estimation merely guided by $\tilde{u}$ using a confidence measure.

5.4 Other Methods

Kim et al. 2009 [37] Kim et al. propose a volumetric approach. The initial surface is given by the ToF depth. This is further refined by optimizing an energy function including ToF, stereo, silhouette terms and a Laplacian prior. Optimization is done with the L-BFGS optimizer [49].

Fischer et al. 2011 [20] Fischer et al. extended Semi-Global Matching Stereo by the approach of Hirschmüller [30] to account for ToF-Stereo data. The algorithm works in disparity space. The energy being minimized is given as

$$E(d) = \sum_i C_{data}(d_i) + \sum_{j \in N_i} \gamma_1 \chi_{\{|d_i - d_j| = 1\}} + \sum_{j \in N_i} \gamma_2 \chi_{\{\{|d_i - d_j| > 1\}}, \tag{14}$$

where d is the disparity, χ_A is 1 when the condition A is true and 0 otherwise and $0 < \gamma_1 < \gamma_2$ are the user specified parameters. The data term in Eq. 14 is defined via

$$C_{data}(d) := \begin{cases} C_{ToF}(d) & \text{if theToF data is valid,} \\ C_{BT}(d) & \text{otherwise,} \end{cases} \tag{15}$$

where C_{ToF} is a truncated reverse Gaussian centered around the ToF disparity estimate. Note that *either* ToF *or* stereo data are used but not both at the same time. The ToF data is invalidated, if the photo consistency score for the ToF disparity is below a certain threshold. The regularizer does not penalize small spatial variations in disparity. As no additional term is added to the functional the optimization step remains the same as in [30] and is done in 16 different 1D directions. As a preprocessing step outliers in the ToF depth image are removed via wavefront propagation.

DalMutto et al. 2012 [16] Based on locally consistent stereo [44] the technique uses a segmentation of the RGB image to guide a bilateral filter for ToF data upsampling. The algorithm takes two depth hypothesis from a stereo algorithm (semi global matching) and ToF respectively which are calculated independently. Each pixel then propagates both depth hypothesis independently according to [44] to surrounding pixel based on color similarity, spatial proximity and photo consistency. Every pixel then has a number of 'votes' casted by neighboring pixels. From these hypothesis the one with the highest plausibility is finally chosen.

Gandhi et al. 2012 [22] The basic idea here is to combine the reprojection and interpolation step of the ToF depth map on the reference frame with a stereo matching procedure. The proposed technique is based on [11], with the difference that , reprojected ToF pixels are used as input instead of sparsely matched feature points. The reprojected ToF points are used as initial seeds for a region growing stereo algorithm. The seeds are first put in a priority queue based on their photo consistency score. Next, the seed with the highest priority is removed from the queue and the corresponding disparity drawn into a final disparity map. The neighbors of the pixel that has just been finalized are then added to the priority queue using the depth estimate with the best stereo score, found by searching around the interpolated ToF depth estimate. This process is repeated until all pixels in the final depth map are drawn, thus implicitly discarding ToF measurements with a bad photo consistency score.

6 Conclusion

We presented an overview over current ToF stereo fusion techniques as well as a guide to setting up such a system. Furthermore, we discussed the importance of high quality depth maps in multimedia applications due to the requirements that applications such as matting, view synthesis or CG effects impose on depth map quality. Still, more effort has to be put into assessing the actual benefits of the ToF stereo fusion over either method alone in a more systematic fashion. We considered various approaches to evaluate these methods and proposed new experiments that should be included in a future evaluation. Finally, we note that currently not all possible depth modalities available from such a system are actually being made use of for fusion purposes. Systems in the future may use the additional modalities to achieve a better accuracy.

7 Acknowledgements

This work is part of a joint research project with the Filmakademie Baden-Württemberg, Institute of Animation. It is co-funded by the Intel Visual Computing Institute and under grant 2-4225.16/380 of the ministry of economy Baden-Württemberg as well as further partners Unexpected, Pixomondo, ScreenPlane, Bewegte Bilder and Tridelity . The content is under sole responsibility of the authors. The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 256941, Reality CG.

References

[1] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 334–341. IEEE, 2012.

[2] B. Bartczak and R. Koch. Dense depth maps from low resolution time-of-flight depth and high resolution color views. *Advances in Visual Computing*, pages 228–239, 2009.

[3] C. Beder, B. Bartczak, and R. Koch. A combined approach for estimating patchlets from pmd depth images and stereo intensity images. In *Pattern Recognition*, pages 11–20. Springer, 2007.

[4] G. Bilodeau, A. Torabi, and F. Morin. Visible and infrared image registration using trajectories and composite foreground images. *Image and Vision Computing*, 29(1):41–50, 2011.

[5] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(4):401–406, 1998.

[6] J.-Y. Bouguet. Camera calibration toolbox for matlab, 2004.

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

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision-ECCV 2004*, pages 25–36. Springer, 2004.

[10] V. Castaneda, D. Mateus, and N. Navab. Stereo time-of-flight. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1684–1691. IEEE, 2011.

[11] J. Cech and R. Sara. Efficient sampling of disparity space for fast and accurate matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[12] J. Chen, S. Paris, J. Wang, W. Matusik, M. Cohen, and F. Durand. The video mesh: A data structure for image-based three-dimensional video editing. In *Proc. of the International Conference on Computional Photography (ICCP)*, 2011.

[13] W.-C. Chiu, U. Blanke, and M. Fritz. Improving the kinect by cross-modal stereo. In *British Machine Vision Conf. BMVA*, pages 116–1, 2011.

[14] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. A probabilistic approach to tof and stereo data fusion. *3DPVT, Paris, France*, 2, 2010.

[15] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. A probabilistic approach to tof and stereo data fusion. In *3DPVT*, Paris, France, May 2010.

[16] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. Cortelazzo. Locally consistent tof and stereo data fusion. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 598–607. Springer, 2012.

[17] F. Devernay and P. Beardsley. Stereoscopic Cinema. In R. Ronfard and G. Taubin, editors, *Image and Geometry Processing for 3-D Cinematography*, volume 5 of *Geometry and Computing*, pages 11–51. Springer Berlin Heidelberg, 2010.

[18] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 23, 2004.

[19] Y. Fei, M. Yu, F. Shao, and G. Jiang. A color correction algorithm of multi-view video based on depth segmentation. In *Computer Science and Computational Technology, 2008. ISCSCT '08. International Symposium on*, volume 2, pages 206–209, 2008.

[20] J. Fischer, G. Arbeiter, and A. Verl. Combination of time-of-flight depth and stereo using semiglobal optimization. In *Int. Conf. on Robotics and Automation (ICRA)*, pages 3548–3553. IEEE, 2011.

[21] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

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

[23] S. E. Ghobadi, O. E. Loepprich, O. Lottnera, F. Ahmadov, K. Hartmann, W. Weihs, and O. Loffeld. Analysis of the personnel safety in a man-machine-cooperation using 2d/3d images. In *Proceedings of the EURON/IARP International Workshop on Robotics for Risky Interventions and Surveillance of the Environment*, Benicassim, Spain, January 2008.

[24] M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, editors. *A State-of-the-Art Survey on Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications.* LNCS. Springer, 2013. to appear.

[25] L. Guan, M. Pollefeys, et al. A unified approach to calibrate a network of camcorders and tof cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008.

[26] S. Gudmundsson, H. Aanaes, and R. Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3d estimation. *IJISTA*, 5(3):425–433, 2008.

[27] U. Hahne and M. Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. *IJISTA*, 5(3):325–333, 2008.

[28] U. Hahne and M. Alexa. Depth imaging by combining time-of-flight and on-demand stereo. *Dynamic 3D Imaging*, pages 70–83, 2009.

[29] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, second edition, 2004.

[30] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 30(2):328–341, 2008.

[31] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Proc. CVPR*, pages 1–8. IEEE, 2007.

[32] B. K. Horn. Closed-form solution of absolute orientation using unit quaternions. *JOSA A*, 4(4):629–642, 1987.

[33] B. K. Horn and M. J. Brooks. *Shape from shading.* MIT press, 1989.

[34] T. Hrkać, Z. Kalafatić, and J. Krapac. Infrared-visual image registration based on corners and hausdorff distance. In *Image Analysis*, pages 383–392. Springer, 2007.

[35] B. Huhle, S. Fleck, and A. Schilling. Integrating 3d time-of-flight camera data and high resolution images for 3dtv applications. In *Proc. 3DTV Conf.* IEEE, 2007.

[36] A. T. Ihler and D. A. Mcallester. Particle belief propagation. In *International Conference on Artificial Intelligence and Statistics*, pages 256–263, 2009.

[37] Y. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3d reconstruction. In *ICCV Workshops*, pages 1542–1549. IEEE, 2009.

[38] Y. M. Kim, D. Chan, C. Theobalt, and S. Thrun. Design and calibration of a multi-view tof sensor fusion system. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–7. IEEE, 2008.

[39] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 508–515. IEEE, 2001.

[40] K. Kuhnert and M. Stommel. Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction. In *Int. Conf. on Intelligent Robots and Systems*, pages 4780–4785. IEEE, 2006.

[41] D. Lefloch, R. Nair, F. Lenzen, H. Schäfer, L. Streeter, M. J. Cree, R. Koch, and A. Kolb. *Technical Foundation and Calibration Methods for Time-of-Flight Cameras*, chapter 1. In Grzegorzek et al. [24], 2013. to appear.

[42] F. Lenzen, K. I. Kim, H. Schäfer, R. Nair, S. Meister, F. Becker, C. S. Garbe, and C. Theobalt. *Denoising Strategies for Time-of-Flight Data*, chapter 2. In Grzegorzek et al. [24], 2013. to appear.

[43] W.-Y. Lo, J. van Baar, C. Knaus, M. Zwicker, and M. Gross. Stereoscopic 3d copy & paste. *ACM Trans. Graph.*, 29(6):147:1–147:10, 2010.

[44] S. Mattoccia. A locally global approach to stereo correspondence. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1763–1770. IEEE, 2009.

[45] L. Mcmillan and S. Gortler. Image-based rendering: A new interface between computer vision and computer graphics. *SIGGRAPH Comput. Graph.*, 33:61–64, 2000.

[46] R. Nair, F. Lenzen, S. Meister, H. Schäfer, C. S. Garbe, and D. Kondermann. High accuracy tof and stereo sensor fusion at interactive rates. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 1–11. Springer, 2012.

[47] R. Nair, S. Meister, M. Lambers, M. Balda, H. Hoffmann, A. Kolb, D. Kondermann, and B. Jähne. *Ground Truth for Evaluating Time of Flight Imaging*, chapter 4. In Grzegorzek et al. [24], 2013. to appear.

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

[49] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 25:773–782, 1980.

[50] J. Park, H. Kim, Y. Tai, M. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *IEEE Proc. ICCV*, 2011.

[51] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph.*, 23(3):664–672, 2004.

[52] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. J. Brostow. Capturing time-of-flight data with confidence. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 945–952. IEEE, 2011.

[53] K. Ruhl, F. Klose, C. Lipski, and M. Magnor. Integrating approximate depth data into dense image correspondence estimation. In *Proc. European Conference on Visual Media Production (CVMP) 2012*, Aug. 2012.

[54] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002.

[55] I. Schiller, C. Beder, and R. Koch. Calibration of a pmd-camera using a planar calibration pattern together with a multi-camera setup. *The international archives of the photogrammetry, remote sensing and spatial information sciences*, 37:297–302, 2008.

[56] Y. Song, C. A. Glasbey, G. W. van der Heijden, G. Polder, and J. A. Dieleman. Combining stereo and time-of-flight images with application to automatic plant phenotyping. In *Image Analysis*, pages 467–478. Springer, 2011.

[57] M. Sturmer, J. Penne, and J. Hornegger. Standardization of intensity-values acquired by time-of-flight-cameras. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.

[58] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):787–800, 2003.

[59] K. Templin, P. Didyk, T. Ritschel, K. Myszkowski, and H.-P. Seidel. Highlight microdisparity for improved gloss depiction. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 31(4), 2012.

[60] A. Toet, L. J. Van Ruyven, and J. M. Valeton. Merging thermal and visual images by a contrast pyramid. *Optical Engineering*, 28(7):287789–287789, 1989.

[61] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool. A hierarchical symmetric stereo algorithm using dynamic programming. *International Journal of Computer Vision*, 47(1-3):275–285, 2002.

[62] L. Wilkes. The role of ocula in stereo post production. *The Foundry, Whitepaper*, 2009.

[63] C. Wu. Visualsfm: A visual structure from motion system, 2011.

[64] Q. Yang, K.-H. Tan, B. Culbertson, and J. Apostolopoulos. Fusion of active and passive sensors for fast 3d capture. In *MMSP*, 2010.

[65] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Pattern recognition: 29th DAGM symposium*, volume 29, pages 214–223, 2007.

[66] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–367. IEEE, 2003.

[67] J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):899–909, 2010.

[68] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Proc. CVPR*, pages 1–8. IEEE, 2008.

[69] J. Zhu, L. Wang, R. Yang, J. Davis, et al. Reliability fusion of time-of-flight depth and stereo for high quality depth maps. *TPAMI*, (99):1–1, 2011.

[70] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. A. J. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3):600–608, 2004.