A Brief Survey of Image-Based Depth Upsampling

Dmitry Chetverikov[1,2], Iván Eichhardt[1,2], and Zsolt Jankó[1]

[1] MTA SZTAKI, Hungary
csetverikov@sztaki.mta.hu
[2] Eötvös Loránd University, Hungary

Abstract. Recently, there has been remarkable growth of interest in the development and applications of Time-of-Flight (ToF) depth cameras. However, despite the permanent improvement of their characteristics, the practical applicability of ToF cameras is still limited by low resolution and quality of depth measurements. This has motivated many researchers to combine ToF cameras with other sensors in order to enhance and upsample depth images. In this paper, we compare ToF cameras to three image-based techniques for depth recovery, discuss the upsampling problem and survey the approaches that couple ToF depth images with high-resolution optical images. Other classes of upsampling methods are also mentioned.

1 Introduction

Image-based 3D reconstruction of static [73, 81, 31] and dynamic [85] objects and scenes is a core problem of computer vision. In the early years of computer vision, it was believed that visual information is sufficient for a computer to solve the problem, as humans can perceive dynamic 3D scenes based on their vision. However, humans do not need to build precise 3D models of an environment to be able to act in the environment, while numerous applications of computer vision require precise 3D reconstruction.

Today, different sensors and approaches are often combined to achieve the goal of building a detailed, geometrically correct and properly textured 3D or 4D (spatio-temporal) model of an object or a scene. Visual and non-visual sensor data are fused to cope with varying illumination, surface properties [37], motion and occlusion. This requires good calibration and registration of the modalities such as color images, laser-measured data (LIDAR, hand-held scanners, Kinect), or Time-of-Flight (ToF) depth cameras. The output is typically a point cloud, a depth image, or a depth image with a color value assigned to each pixel (RGBD).

A calibrated stereo rig is a widespread, classical device to acquire depth information based on **visual data** [73]. Since its baseline, i.e, the distance between the two cameras, is usually narrow, the resulting depth resolution is limited. Wide-baseline multiview stereo [81] can provide a better depth resolution at the expense of more frequent occlusions and partial loss of spatial data. A collection of different-size, uncalibrated images of an object, or a video, can also be used for 3D reconstruction. However, this requires point correspondence, or tracking, across images/frames, which is not always possible.

Photometric stereo [31] applies a camera and several light sources to acquire the surface normals. The normal vectors are integrated to reconstruct the surface. The method

provides fine surface details but suffers from less robust global geometry [61]. The latter is better captured by stereo methods which can be combined with photometric stereo [61] to obtain precise local and global geometry.

Shape acquisition systems using structured light [72, 16] contain one or two cameras and a projector that casts a specific, fixed or programmable, pattern onto the shape surface. Systems with programmable light pattern can achieve high precision of surface measurement.

The approaches to image-based 3D reconstruction listed above are the most widely used in practice. A number of other approaches to 'Shape-from-X' exist [84, 86], such as Structure-from-Motion, Shape-from-Texture, Shape-from-Shading and Shape-from-Focus. These approaches are usually less precise and robust. They can be applied when high precision is not required, or as additional shape cues in combination with other methods.

Among the **non-visual** sensors, the popular Kinect [101] can be used for real-time dense 3D reconstruction, tracking and interaction [38, 62]. The device combines a color camera with a depth sensor projecting invisible structural light. Currently, its resolution and precision are limited, but still sufficient for applications in game industry and human-computer interaction (HCI).

Different LIDAR devices [92] have numerous applications in various areas including robot vision, autonomous vehicles, traffic monitoring, as well as scanning and 3D reconstruction of indoor and outdoor scenes, buildings and complete residential areas. They deliver point clouds with a measure of surface reflectivity assigned to each point.

Last but not least, ToF depth cameras [18, 29] acquire low-resolution, registered depth and intensity images at the rates suitable for real-time robot vision, navigation, obstacle avoidance, game industry and HCI. This paper is devoted to a specific but critical aspect of ToF image processing, namely, depth image upsampling. The upsampling can be performed in different ways. We give a brief survey of the methods that combine a low-resolution ToF depth image with a registered high-resolution optical image in order to refine the depth resolution, typically by a factor of 5 to 10.

The rest of the paper is structured as follows. In section 2, we discuss the specifics of an important class of ToF cameras and compare their features to the features of three main image-based methods. Section 3 is the core of our survey, while section 4 provides conclusion and outlook.

2 Time-of-Flight cameras

A recent survey [18] offers a comprehensive summary of the operation principles, advantages and limitations of ToF cameras. The survey focuses on lock-in ToF cameras which are widely used in numerous applications, while the other category of ToF cameras, the pulse-based, is still rarely used. Our survey is also devoted to lock-in ToF cameras; for simplicity we will omit the term 'lock-in'.

ToF cameras [68, 24] are small, compact, low-weight, low-consumption devices that emit infrared light and measure the time-of-flight to the observed object for calculating the distance to the object, usually called the depth. Contrary to LIDAR devices, ToF cameras have no mobile parts, and they capture depth images rather than point clouds.

In addition to depth, ToF cameras deliver registered intensity images of the same size and reliability values of depth measurements.

The main disadvantages of ToF cameras are their low resolution and significant acquisition noise. Although both resolution and quality are gradually improving, they are inherently limited by chip size and small active illumination energy, respectively. The highest currently available ToF camera resolution is QVGA (320×240), with VGA (640×480) being a target of future development.

Tab. 1 compares ToF cameras to three main image-based methods in terms of basic features. Stereo vision (SV) and structured light (SL) need to solve the correspondence, or matching, problem; the other two methods, photometric stereo (PS) and ToF, are correspondence-free. Of the four techniques, only ToF does not require extrinsic calibration. SV is a passive method, the rest use active illumination. This allows them to work with textureless surfaces when SV fails. On the other hand, distinct, strong textures facilitate the operation of SV but can deteriorate the performance of the active methods, especially when different textures cover the surface and its reflectance varies.

Table 1. Comparison of four techniques for depth measurement.

	stereo vision	photometric stereo	structured light	ToF camera
correspondence	yes	no	yes	no
extrinsic calibration	yes	yes	yes	no
active illumination	no	yes	yes	yes
weak texture perform.	weak	good	good	good
strong texture perform.	good	medium	medium	medium
low light performance	weak	good	good	good
bright light perform.	good	weak	medium/weak	medium
outdoor scene	yes	no	no	yes?
dynamic scene	yes	no	yes	yes
image resolution	camera depend.	camera depend.	camera depend.	low
depth accuracy	mm to cm	mm	μm to cm	mm to cm

The active methods operate well in low lighting conditions, when scene illumination is poor. Not surprisingly, passive stereo fails when visibility is low. The situation reverses for bright lighting that can prevent the operation of PS and reduce the performance of SL and ToF. In particular, bright lighting can increase ambient light noise in ToF [18] if ambient light contains the same wavelength as camera light. (A more recent report [51] claims that bright lighting performance of ToF is good.) High-reflectivity surfaces can be a problem for all of the methods.

PS is efficient for neither outdoor nor dynamic scenes. SL can cope with time-varying surfaces, but currently it is not applied in outdoor conditions. Both SV and ToF can be used outdoor and applied to dynamic scenes, although the outdoor applicability of ToF cameras can be limited by their illumination energy and range [14, 9], as well as by ambient light. Image resolution of the first three techniques depends on the camera and can be high, contrary to ToF cameras whose resolution is low. Depth accuracy of

SV depends on the baseline and is comparable to that of ToF. The other two techniques, especially SL, can yield higher accuracy.

From the comparison of the four techniques, we observe that ToF cameras and passive stereo vision have complementary characteristics. As discussed below in section 3, this fact has motivated researchers to combine the two sources of depth data in order to enhance applicability, accuracy and robustness of 3D vision systems. Although ToF camera–stereo data fusion usually results in ToF depth image upsampling, in some cases this may be rather a by-product than the main goal of the fusion.

ToF cameras have numerous **applications**. The related surveys [19, 18] conclude that the most exploited feature of the cameras is their ability to operate without moving parts while providing depth maps at high frame rates. This capability greatly simplifies the solution of a critical task of 3D vision, the foreground-background separation. ToF cameras are exploited in robot vision [36] for navigation [91, 13, 88, 99] and 3D pose estimation and mapping [67, 56, 22].

Further important application areas are 3D reconstruction of objects and environments [10, 17, 3, 20, 46, 42], computer graphics [82, 69, 44] and 3D television [80, 78, 90]. (See [77] for a recent survey of depth sensing for 3DTV.) ToF cameras are applied in various tasks related to recognition and tracking of people [26, 4, 43] and parts of human body: hand [53, 60], head [23] and face [60, 71]. Alenya et al. [1] use color and ToF camera data to build 3D models of leaves for automated plant measurement. Additional applications are discussed in the recent book [24].

3 ToF depth image upsampling

Low resolution and low signal-to-noise ratio are the two main disadvantages of ToF depth imagery. The goal of depth image upsampling is to increase the resolution and simultaneously improve image quality, in particular, near depth edges where surface discontinuities tend to result in erroneous or lacking measurements [18]. In some applications, such as mixed reality and game industry, the depth edge areas are especially important because they determine occlusion and disocclusion of moving actors.

Approaches to depth upsampling form three main classes [15]. In this survey, we discuss image-guided upsampling when a high-resolution optical image registered with a low-resolution depth image is used to refine the depth. Image-guided upsampling was selected because it is more widespread than the other two classes of approaches, and sufficient experience had been gained in the area. However, for completeness we will now briefly discuss the other two classes, as well.

3.1 Upsampling with stereo and with multiple measurements

ToF–stereo fusion [59] combines ToF camera depth with multicamera stereo data. Hansard et al. [29] discuss the existing variants of this approach and provide a comparative evaluation of several methods. The important issue of registering the ToF camera and the stereo data is also addressed. By mapping ToF depth values to the disparities of a high-resolution camera pair, it is possible to simultaneously upsample the depth values and improve the quality of the disparities [25]. Kim et al. [42] address the problem

of sparsely textured surfaces and self-occlusions in stereo vision by fusing multicamera stereo data with multiview ToF sensor measurements. The method yields dense and detailed 3D models of scenes challenging for stereo alone while enhancing the ToF depth images. Zhu et al. [103, 102, 104] also explore the complementary features of ToF cameras and stereo in order to improve accuracy and robustness.

Yang et al. [96] present a setup that combines a ToF depth camera with three stereo cameras and report on GPU-based, fast stereo depth frame grabbing and real-time ToF depth upsampling. The system fails in large dark regions that cause troubles to both stereo and ToF cameras. Bartczak and Koch [2] combine multiple high-resolution color views with a ToF camera to obtain dense depths maps of a scene. Similar input data are used by Li et al. [49] who present a joint learning-based method exploiting differential features of the observed surface. Kang and Ho [39, 33] report on a system that contains multiple depth and color cameras.

Hahne and Alexa [27, 28] claim that combination of ToF camera and stereo vision can provide enhanced depth data even without precise calibration. Kuhnert and Stommel [46] fuse ToF depth data with stereo data for real-time indoor 3D environment reconstruction in mobile robotics. Further methods are discussed in the recent survey [59]. A drawback of ToF–stereo is that it still inherits critical problems of passive stereo vision: the correspondence problem, the problem of textureless surfaces, and the problem of occlusions.

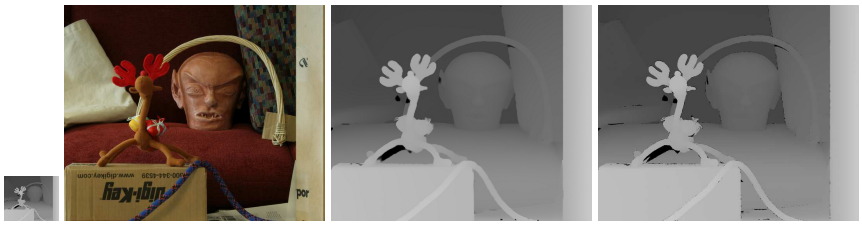
A natural way to improve resolution is to combine multiple measurements of an object. Fusing multiple ToF depth measurements into one image is sometimes referred to as **temporal and spatial upsampling** [15]. In the studies [76, 8], the authors acquire multiple depth images of a static scene from different viewpoints and merge them into a single depth map of higher resolution. An advantage of such approaches is that it does not need a sensor of another type. Working with depth images only allows one to avoid the so called ‘texture copying problem’ that will be discussed later in relation to image-guided upsampling. A limitation of the methods [76, 8] is that only static objects can be measured.

Mac Aodha et al. [55] use a training dataset of high-resolution depth images for patch-based upsampling of a low-resolution depth image. Although theoretically attractive, the method is too time-consuming for most applications. A somewhat similar, patch-based approach was developed by Hornacek et al. [34] who exploit patchwise self-similarity of a scene and search for patch correspondences within the input depth image. The method [34] aims at single-image upsampling while the algorithm [55] needs a large collection of high-resolution exemplars to search in. A drawback of the method [34] is that it relies on patch correspondences which may be difficult to obtain, especially for less characteristic surface regions. Finally, Katz et al. [40] have recently patented a method for combined depth filtering and resolution refinement.

3.2 Problems of image-guided upsampling

Fig. 1 demonstrates an example of successful upsampling of a high-quality depth image of low resolution. The input depth and color images are from the Middlebury stereo dataset [74]. The original high-resolution depth image was acquired with structural light, then artificially downsampled to get the low-resolution image shown in Fig. 1.

Small parts of depth data (dark regions) are lost. The upsampled depth is smooth and very similar to the original high-resolution data used as the ground truth. In the Middlebury data, depth discontinuities match well the corresponding edges of the color image. This dataset is often used for quantitative comparative evaluation of image-guided upsampling techniques.



input depth and color images upsampled depth ground-truth depth

Fig. 1. Middlebury input data, upsampled depth and the ground truth.

For real-world data and applications, the problem of depth upsampling is more complicated than for the high-quality Middlebury data. Fig. 2 illustrates the negative features of depth images captured by ToF cameras[1]. The original depth resolution is very low compared to that of the color image. When resized to the size of the color image, the depth image clearly shows its deficiencies: a part of the data is lost due to low resolution; some shapes, e.g., the heads, are distorted. Despite the calibration, the contours of the depth image do not always coincide with those of the color image. There are erroneous and lacking measurements along the depth edges, in the dark region on the top, and in the background between the chair and the poster.

To use a high-resolution image for depth upsampling, one needs to relate image features to depth features. A basic assumption exploited by most upsampling methods is that **image edges are related to depth edges**, that is, to surfaces discontinuities. It is usually assumed [11, 21, 54, 64, 50, 15] that smooth depth regions exhibit themselves as smooth intensity, or color, regions, while depth edges underlie intensity edges. Clearly, this assumption is violated in the regions of high-contrast texture and on the border of a strong shadow.

Some studies [94, 83] relax the assumption of depth-intensity edge coincidence in order to circumvent the problems discussed below and avoid the resulting artefacts. However, depth edges are in any case a sensitive issue. Since image features are the only data available for upsampling, one has to find a balance between the edge coincidence assumption and other priors. This balance is data-dependent, which may necessitate adaptive parameter tuning of an upsampling algorithm.

Precise camera **calibration** is crucial for the applications that require good-quality depth images, in general, and accurate depth discontinuities, in particular. Techniques and engineering tools used to calibrate ToF cameras and enhance their quality are discussed in numerous studies [29, 65, 68, 32, 52, 48]. Procedures for joint calibration of a

[1] Data courtesy of Zinemath Zrt [105].

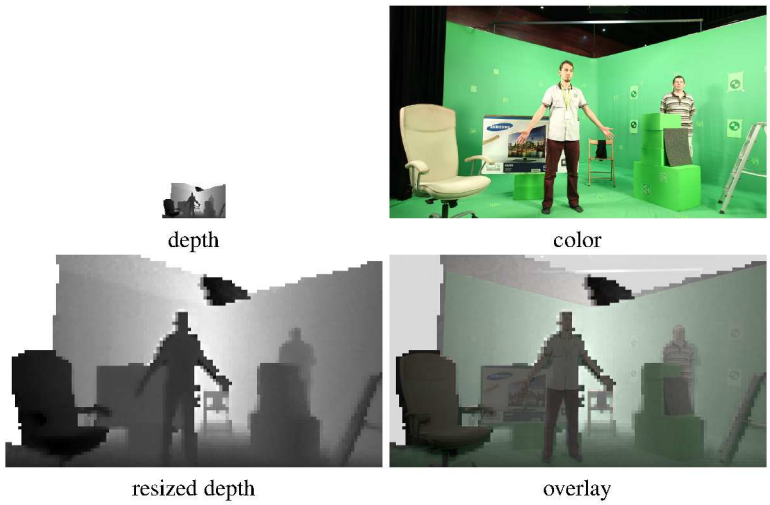


Fig. 2. Depth and color data captured in a studio. Upper row: original depth and color images. Lower row: depth image resized to color image size (left) and overlaid (right).

ToF camera and an intensity camera are described in [64, 15]. Many studies apply the well-known calibration method [100].

Inaccurate registration of depth and intensity images due to imprecise calibration results in deterioration of upsampled depth. Fig. 3 illustrates the effect of inaccurate registration on depth upsampling. The discrepancy between the depth and intensity images is caused by a relative shift of two pixels, in one case, and ten pixels, in the other. As the shift grows, the depth borders become blurred and coarse.

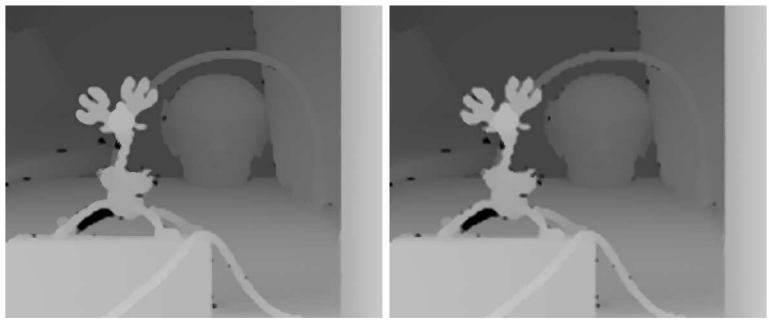


Fig. 3. The effect of imprecise calibration on depth upsampling. The discrepancy between the input depth and color images is 2 and 10 pixels, respectively.

Avoiding depth image blur to preserve contrast depth edges is a major issue of upsampling methods. Because of the depth-intensity edge coincidence assumption, this

issue is related to the so-called **texture copying**, or texture transfer, problem. Contrast image textures tend to 'imprint' onto the upsampled depth image, as illustrated in Fig. 4 where textured regions cause visible perturbation in the refined depth. This disturbing phenomenon and possible remedies are discussed in [94, 83].

color image upsampled depth ground-truth depth

Fig. 4. The texture transfer problem in depth upsampling.

3.3 Depth upsampling with single image

Image-guided ToF depth upsampling can be based on a single image, or use video. For a single image, upsampling methods in their operation principles can be loosely grouped into the following classes:

- methods using different versions of **multilateral filtering** [45, 97, 5, 70, 21, 94];
- methods based on **Markov Random Fields** [11, 54, 7];
- methods applying **optimization** [64, 7, 15, 50];
- methods using **Non-Local Means** (NL-Means) filtering [35, 64];
- methods based on **segmentation** [87, 83];
- other methods, e.g., using a Bayesian approach [50].

The classes may overlap since a method may combine several techniques. For example, MRF-based approaches often lead to optimization and may apply filtering techniques, as well.

Techniques using video are based on similar principles, but they may exploit video redundancy and additional constraints such as motion coherence, also called temporal consistency. We will discuss video-based approaches separately.

Upsampling methods have to combine two different kinds of spatial data, the ToF depth and the intensity, or color. When video is available, the temporal dimension should also be taken into account. Upsampling techniques based on filtering in spatial or spatio-temporal domain are usually variants and extensions of the original **bilateral filter** [89]. The bilateral filter applies two Gaussian kernels, a spatial (or domain) one and a range one. The spatial kernel weighs the distance from the filter center, while the range kernel weighs the absolute difference between the image value in the center and the value in a point of the window. The bilateral filter can be efficiently implemented

in constant and real time [66, 95] which makes its practical application especially attractive. The reader is referred to the book [63] for a detailed discussion of bilateral filtering.

The idea of bilateral filtering has been extended in different ways. The joint (or cross) bilateral filters apply the range filter to a second image (guidance image) rather than to the original one. These filters have been successfully used in a wide range of tasks including **joint bilateral upsampling** (JBU) of depth images [45]. Further attempts to combine different criteria and enhance the result of upsampling led to the use **multilateral**, rather than bilateral, filters.

Yang et al. [97] applied the joint bilateral filter to a cost volume that measures the distance between the potential depth candidates and the ToF depth image resized to the color image size. The filter enforces the consistence of the cost values and the color values. The upsampling problem is formulated as adaptive cost aggregation. To improve the robustness of the method [97] and its performance at depth edges, the authors later added the weighted median filter and proposed a multilateral framework [94]. The use of the median filter can also diminish the effect of texture copying. (See [98] for a tutorial on weighted median filtering.) The improved method [94] was implemented on a GPU to build a real-time high-resolution depth capturing system.

Chan et al. [5] proposed an upsampling scheme based on the blended, composite joint bilateral filter that locally adapts to the noise level and the smoothness of the depth function. Depending on the local context, the composite filter switches between the standard bilateral upsampling filter and an edge-preserving smoothing depth filter independent from color data. Such solution can potentially reduce artefacts like texture copying. Riemens et al. [70] presented a multi-step (multiresolution) implementation of JBU that doubles the depth resolution at each step. Finally, Garcia et al. [21] enhanced the joint bilateral upsampling by taking into account the low reliability of depth values near depth edges.

The early paper [11] describes an application of the **Markov Random Fields** (MRF) to depth upsampling using a high-resolution color image. The two-layer MRF is defined via the quadratic difference between the measured and the estimated depth, a depth smoothing prior, and the weighting factors that relate image edges to depth edges. This formulation leads to a least square optimization problem which is solved by the conjugate algorithm. Lu et al. [54] use a linear cost term (truncated absolute difference) since the quadratic cost is less robust to outliers. Their formulation of the MRF-based depth upsampling problem includes adaptive elements and is solved by the loopy belief propagation. Choi et al. [7] use quadratic terms in the proposed MRF energy and apply both discrete and continuous optimization in a multiresolution framework.

A number of approaches apply an **optimization** algorithm to an upsampling cost function not related to an MRF. Such cost functions often contain terms similar to those used by the MRF-based methods. Ferstl et al. [15] define an energy function that combines a standard quadratic depth data term with a regularizing Total Generalized Variation (TGV) term and an anisotropic diffusion term that relates image gradients to depth gradients. The primal-dual optimization algorithm is used to minimize the energy functional.

Park et al. [64] apply an MRF to detect and remove outliers in depth data prior to upsampling. However, their optimization approach to upsampling does not rely on Markov Random Fields. The functional formulated in [64] includes **Non-Local Means** (NLM) regularizing term that helps preserve local structure and fine details in presence of significant noise. (See the recent survey [57] for a discussion of the NLM filter.).

The method proposed by Huhle [35] et al. also detects outliers and uses the color NLM filter. However, their approach is based on filtering rather than optimization. The paper [35] discusses the interdependence between surface texturing and smoothing. The authors point out that the correspondence of depth and image pixels may change due to the displacement of the reconstructed point.

Segmentation of color and depth images can be used for upsampling either separately [87] or in combination with other tools. Tallon et al. [87] propose an upsampling and noise reduction method based on joint segmentation of depth and intensity into regions of homogeneous color and depth. Conditional mode estimation is used to detect and correct regions with inconsistent features. Soh et al. [83] point out that the image-depth edge coincidence assumption may occasionally be invalid. They oversegment the color image to obtain image super-pixels and use them for depth edge refinement. Then a MAP-MRF framework is used to further enhance the depth.

Li et al. [50] developed a Bayesian approach to depth image upsampling that takes intrinsic camera errors into consideration. The method simulates uncertainty of depth and color measurements by a Gaussian and a spatial-anisotropic kernel, respectively. The scene is assumed to be piecewise planar. RANSAC is used to select inliers for each plane model. An objective function combining depth and color data terms is introduced and optimized to obtain the refined depth.

Most of the above mentioned studies compare the proposed method to existing techniques. Often, images from the Middlebury stereo dataset [74] containing the ground truth depth are used for quantitative comparison. The recent evaluation study [47] uses images from [74] as well as manually labelled ToF camera and color data. The study compares a number of image-guided upsampling methods including bilateral filters, MRF optimization and the cost volume-based technique [97].

3.4 Video-based depth upsampling

In this section, we briefly discuss the depth upsampling methods that use video rather than a single image. As already mentioned, the two categories of methods are based on the same assumptions and principles, but the video-based techniques may apply additional constraints.

To obtain depth video, Choi et al. [6] apply motion-compensated frame interpolation and the composite Joint Bilateral Upsampling procedure [5]. Dolson et al. [12] consider dynamic scenes and do not use the assumption of identical frame rate of the two video streams. They present a Gaussian framework for multidimensional extension of 2D bilateral filter in space and time. Fast GPU implementation is discussed.

Xian et al. [93] consider synchronized depth and image video cameras and propose upsampling solution implemented on GPU in real time on the frame-by-frame basis without temporal processing. Their multilateral filter is inspired by the composite Joint Bilateral Upsampling procedure [5]. Kim et al. [41] propose a depth video upsampling

method that also operates on the frame-by-frame basis. They use adaptive bilateral filter taking into account the low SNR of ToF camera data. The problem of texture copying is addressed.

Richardt et al. [69] consider the task of video-based depth upsampling in the context of computer graphics applications, such as video relighting, geometry-based abstraction and stylization, and rendering. The depth data are first pre-processed to remove typical artefacts. Then a dual-joint-bilateral filter is applied to upsample the depth. Finally, a spatio-temporal filter is used that blends spatial and temporal components. The blending parameter specifies the degree of depth propagation from the previous time step to the current time step using motion compensation.

Min et al. [58] propose weighted mode filtering based on a joint histogram. Temporal coherence of depth video is achieved by extending the method to neighboring frames. Optical flow supported by a patch-based flow reliability measure is used for motion estimation and compensation. Schwarz et al. [78–80] view the depth upsampling process as a weighted energy optimization problem constrained by temporal consistency.

Finally, Vosters et al. [90] evaluate and compare several efficient video depth upsampling methods in terms of depth accuracy and interpolation quality, in the context of 3DTV. They also provide an analysis of computational complexity and runtime for GPU implementations of the methods.

4 Conclusion

The main purpose of this brief survey was to provide an introduction to the depth upsampling problem and give short descriptions of approaches. In our opinion, this problem is of interest beyond the area of ToF camera data processing since sensor data fusion becomes more and more popular. For example, studies in image-based **point cloud upsampling** [30, 75] apply tools similar or identical to those used by the depth upsampling methods.

We believe that in near future ToF cameras will undergo fast changes in the direction of higher resolution, increasing range, better robustness and improved image quality. As a consequence, their application areas will extend and grow, leading to much more frequent use and lower prices. We also believe that the trend of coupling ToF cameras with other complementary sensors will persist resulting in growing demand for studies in depth data fusion with other kinds of data.

For the image processing community to be able to meet this demand, the critical issue is that of evaluation and comparative testing of the proposed methods. Currently, many studies assume ideally calibrated data and provide tests on the Middlebury stereo dataset [74]. Such tests are not really indicative of the performance in real applications. A good benchmark of ToF data acquired in different real-world conditions is needed.

Acknowledgement

We are grateful to Zinemath Zrt for drawing our attention to the depth upsampling problem and providing test data.

References

1. G. Alenya, B. Dellen, and C. Torras. 3D modelling of leaves from color and ToF data for robotized plant measuring. In *IEEE Int. Conf. on Robotics and Automation*, pages 3408–3414, 2011.
2. B. Bartczak and R. Koch. Dense depth maps from low resolution time-of-flight depth and high resolution color views. In *Advances in Visual Computing*, pages 228–239. Springer, 2009.
3. C. Beder, B. Bartczak, and R. Koch. A comparison of PMD-cameras and stereo-vision for the task of surface reconstruction using patchlets. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
4. A. Bevilacqua, L. Di Stefano, and P. Azzari. People Tracking Using a Time-of-Flight Depth Sensor. In *Proc. Int. Conference on Video and Signal Based Surveillance*, page 89, 2006.
5. D. Chan, H. Buisman, C. Theobalt, and S. Thrun. A noise-aware filter for real-time depth upsampling. In *Proc. ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.
6. J. Choi, D. Min, B. Ham, and K. Sohn. Spatial and temporal up-conversion technique for depth video. In *Proc. Int. Conf. on Image Processing*, pages 3525–3528, 2009.
7. O. Choi, H. Lim, B. Kang, et al. Discrete and continuous optimizations for depth image super-resolution. In *Proc. IS&T/SPIE Electronic Imaging*, pages 82900C–82900C, 2012.
8. Y. Cui, S. Schuon, D. Chan, et al. 3D shape scanning with a time-of-flight camera. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1173–1180, 2010.
9. G. De Cubber, D. Doroftei, H. Sahli, and Y. Baudoin. Outdoor terrain traversability analysis for robot navigation using a time-of-flight camera. In *Proc. RGB-D Workshop on 3D Perception in Robotics*, 2011.
10. B. Dellen, R. Alenyà, Sergi Foix, S., and C. Torras. 3D object reconstruction from Swissranger sensor data using a spring-mass model. In *Proc. Int. Conf. on Comput. Vision Theory and Applications*, volume 2, pages 368–372, 2009.
11. J. Diebel and S. Thrun. An application of Markov random fields to range sensing. In *Proc. Advances in Neural Information Processing Systems*, pages 291–298, 2005.
12. J. Dolson, J. Baek, C. Plagemann, and S. Thrun. Upsampling range data in dynamic environments. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1141–1148, 2010.
13. P. Einramhof and M. Olufs, S. Vincze. Experimental evaluation of state of the art 3D-sensors for mobile robot navigation. In *Proc. Austrian Association for Pattern Recognition Workshop*, pages 153–160, 2007.
14. D. Falie and V. Buzuloiu. Wide range time of flight camera for outdoor surveillance. In *Proc. IEEE Symposium on Microwaves, Radar and Remote Sensing*, pages 79–82, 2008.
15. D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proc. Int. Conf. on Computer Vision*, pages 993–1000, 2013.
16. D. Fofi, T. Sliwa, and Y. Voisin. A comparative survey on invisible structured light. In *Electronic Imaging 2004*, pages 90–98. International Society for Optics and Photonics, 2004.
17. S. Foix, G. Alenya, J. Andrade-Cetto, and C. Torras. Object modeling using a ToF camera under an uncertainty reduction approach. In *Proc. Int. Conf. on Robotics and Automation*, pages 1306–1312, 2010.
18. S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (tof) cameras: a survey. *Sensors Journal*, 11(9):1917–1926, 2011.
19. S. Foix, R. Alenyà, and C. Torras. Exploitation of time-of-flight (ToF) cameras. Technical Report IRI-DT-10-07, IRI-UPC, 2010.

20. S. Fuchs and S. May. Calibration and registration for precise surface reconstruction with time-of-flight cameras. *International Journal of Intelligent Systems Technologies and Applications*, 5:274–284, 2008.
21. F. Garcia, B. Mirbach, B. Ottersten, et al. Pixel weighted average strategy for depth sensor data fusion. In *Proc. Int. Conf. on Image Processing*, pages 2805–2808, 2010.
22. P. Gemeiner, P. Jojic, and M. Vincze. Selecting good corners for structure and motion recovery using a time-of-flight camera. In *Int. Conf. on Intelligent Robots and Systems*, pages 5711–5716, 2009.
23. S.B. Gokturk and C. Tomasi. 3D head tracking based on recognition and interpolation using a time-of-flight depth sensor. In *Proc. Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 211–217, 2004.
24. M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb (Eds). *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013.
25. S.A. Guomundsson, H. Aanæs, and R. Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3D estimation. *Int. Journal of Intelligent Systems Technologies and Applications*, 5(3):425–433, 2008.
26. S.A. Guomundsson, R. Larsen, H. Aanæs, et al. ToF imaging in smart room environments towards improved people tracking. In *Proc. Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2008.
27. U. Hahne and M. Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. *Int. Journal of Intelligent Systems Technologies and Applications*, 5:325–333, 2008.
28. U. Hahne and M. Alexa. Depth imaging by combining time-of-flight and on-demand stereo. In *Dynamic 3D Imaging*, pages 70–83. Springer, 2009.
29. M. Hansard, S. Lee, O. Choi, and R. Horaud. *Time-of-flight cameras*. Springer, 2013.
30. A. Harrison and P. Newman. Image and sparse laser fusion for dense scene reconstruction. In *Field and Service Robotics*, pages 219–228. Springer, 2010.
31. S. Herbort and C. Wöhler. An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods. *3D Research*, 2(3):1–17, 2011.
32. C. Herrera, J. Kannala, J. Heikkilä, et al. Joint depth and color camera calibration with distortion correction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34:2058–2064, 2012.
33. Y.-S. Ho and Y.-S. Kang. Multi-view depth generation using multi-depth camera system. In *International Conference on 3D Systems and Application*, 2010.
34. M. Hornacek, C. Rhemann, M. Gelautz, and C. Rother. Depth Super Resolution by Rigid Body Self-Similarity in 3D. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1123–1130, 2013.
35. B. Huhle, T. Schairer, P. Jenke, and W. Straßer. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Computer Vision and Image Understanding*, 114:1336–1345, 2010.
36. S. Hussmann and T. Liepert. Robot vision system based on a 3D-ToF camera. In *Proc. Conf. on Instrumentation and Measurement Technology*, pages 1–5, 2007.
37. I. Ihrke, K.N. Kutulakos, M. Magnor, W. Heidrich, et al. State of the art in transparent and specular object reconstruction. In *EUROGRAPHICS 2008 State of the Art Reports*, 2008.
38. S. Izadi, D. Kim, O. Hilliges, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proc. ACM Symp. on User Interface Software and Technology*, pages 559–568, 2011.
39. Y.-S. Kang and Y.-S. Ho. High-quality multi-view depth generation using multiple color and depth cameras. In *IEEE Int. Conf. on Multimedia and Expo*, pages 1405–1410, 2010.
40. S. Katz, A. Adler, and G. Yahav. Combined depth filtering and super resolution. US Patent 8,660,362. `http://www.google.com/patents/US8660362`, 2014.

41. C. Kim, H. Yu, and G. Yang. Depth super resolution using bilateral filter. In *Proc. Int. Congress on Image and Signal Processing*, volume 2, pages 1067–1071, 2011.
42. Y.M. Kim, C. Theobalt, J. Diebel, et al. Multi-view image and ToF sensor fusion for dense 3D reconstruction. In *ICCV Workshops*, pages 1542–1549, 2009.
43. S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion for 3D human body tracking with an articulated 3D body model. In *Proc. Int. Conf. on Robotics and Automation*, pages 1686–1691, 2006.
44. A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29, pages 141–159, 2010.
45. J. Kopf, M.F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. In *ACM Transactions on Graphics*, volume 26, page 96, 2007.
46. K.-D. Kuhnert and M. Stommel. Fusion of stereo-camera and pmd-camera data for real-time suited precise 3D environment reconstruction. In *Proc. Int. Conf. on Intelligent Robots and Systems*, pages 4780–4785, 2006.
47. B. Langmann, K. Hartmann, and O. Loffeld. Comparison of depth super-resolution methods for 2D/3D images. *Int. Journal of Computer Information Systems and Industrial Management Applications*, 3:635–645, 2011.
48. D. Lefloch, R. Nair, F. Lenzen, et al. Technical Foundation and Calibration Methods for Time-of-Flight Cameras. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 3–24. Springer, 2013.
49. J. Li, Z. Lu, G. Zeng, et al. A Joint Learning-Based Method for Multi-view Depth Map Super Resolution. In *Proc. Asian Conference on Pattern Recognition*, pages 456–460, 2013.
50. J. Li, G. Zeng, R. Gan, et al. A Bayesian approach to uncertainty-based depth map super resolution. In *Proc. Asian Conf. on Computer Vision*, pages 205–216, 2012.
51. Larry Li. Time-of-Flight Camera – An Introduction. Technical Report SLOA190B, Texas Instruments, 2014. Available at `www.ti.com/lit/wp/sloa190b/sloa190b.pdf`.
52. M. Lindner, I. Schiller, A. Kolb, and R. Koch. Time-of-flight sensor calibration for accurate range sensing. *Computer Vision and Image Understanding*, 114:1318–1328, 2010.
53. X. Liu and K. Fujimura. Hand gesture recognition using depth data. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 529–534, 2004.
54. J. Lu, D. Min, R.S. Pahwa, and M.N. Do. A revisit to MRF-based depth map super-resolution and enhancement. In *Int. Conference on Acoustics, Speech and Signal Processing*, pages 985–988, 2011.
55. O. Mac Aodha, N.D.F. Campbell, A. Nair, and G.J. Brostow. Patch based synthesis for single depth image super-resolution. In *Proc. European Conf. on Computer Vision*, pages 71–84, 2012.
56. S. May, D. Droeschel, D. Holz, et al. 3D pose estimation and mapping with time-of-flight cameras. In *Proc. IROS Workshop on 3D Mapping*, 2008.
57. Peyman Milanfar. A tour of modern image filtering: new insights and methods, both practical and theoretical. *IEEE Signal Processing Magazine*, 30:106–128, 2013.
58. D. Min, J. Lu, and M.N. Do. Depth video enhancement based on weighted mode filtering. *IEEETIP*, 21:1176–1190, 2012.
59. R. Nair, K. Ruhl, F. Lenzen, et al. A Survey on Time-of-Flight Stereo Fusion. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 105–127. Springer, 2013.
60. H. Nanda and K. Fujimura. Visual tracking using depth data. In *Proc. Conf. on Computer Vision and Pattern Recognition Workshops*, 2004.
61. D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *ACM Transactions on Graphics*, volume 24, pages 536–543, 2005.

62. R.A. Newcombe, A.J. Davison, S. Izadi, et al. KinectFusion: Real-time dense surface mapping and tracking. In *Proc. IEEE Int. Symp. on Mixed and Augmented Reality*, pages 127–136, 2011.
63. S. Paris, P. Kornprobst, and F. Tumblin, J. andDurand. *Bilateral Filtering*. Now Publishers Inc., 2009.
64. J. Park, H. Kim, Y.-W. Tai, et al. High quality depth map upsampling for 3D-ToF cameras. In *Proc. Int. Conf. on Computer Vision*, pages 1623–1630, 2011.
65. N. Pfeifer, D. Lichti, J. Böhm, and W. Karel. 3D cameras: Errors, calibration and orientation. In *TOF Range-Imaging Cameras*, pages 117–138. Springer, 2013.
66. F. Porikli. Constant time O(1) bilateral filtering. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
67. A. Prusak, O. Melnychuk, H. Roth, and I. Schiller. Pose estimation and map building with a time-of-flight-camera for robot navigation. *Int. Journal of Intelligent Systems Technologies and Applications*, 5:355–364, 2008.
68. F. Remondino and D. Stoppa. *ToF range-imaging cameras*. Springer, 2013.
69. C. Richardt, C. Stoll, N. A Dodgson, et al. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. In *Computer Graphics Forum*, volume 31, pages 247–256, 2012.
70. A.K. Riemens, O.P. Gangwal, B. Barenbrug, and R.-P.M. Berretty. Multistep joint bilateral depth upsampling. In *IS&T/SPIE Electronic Imaging*, pages 72570M–72570M, 2009.
71. J.R. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez. Improving human face detection through ToF cameras for ambient intelligence applications. In *Ambient Intelligence-Software and Applications*, pages 125–132. Springer, 2011.
72. J. Salvi, S. Fernandez, T. Pribanic, and X. Llado. A state of the art in structured light patterns for surface profilometry. *Pattern Recognition*, 43:2666–2680, 2010.
73. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002.
74. D. Scharstein, R. Szeliski, H. Hirschmüller, et al. Middlebury stereo datasets. `http://vision.middlebury.edu/stereo/data/`, 2001–2014.
75. J.R. Schoenberg, A. Nathan, and M. Campbell. Segmentation of dense range information in complex urban scenes. In *Proc. Int. Conf. on Intelligent Robots and Systems*, pages 2033–2038. IEEE, 2010.
76. S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for ToF 3D shape scanning. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 343–350, 2009.
77. S. Schwarz, R. Olsson, and M. Sjostrom. Depth Sensing for 3DTV: A Survey. *IEEE MultiMedia*, 20:10–17, 2013.
78. S. Schwarz, M. Sjöström, and R. Olsson. Temporal consistent depth map upscaling for 3DTV. In *IS&T/SPIE Electronic Imaging*, pages 901302–901302, 2014.
79. S. Schwarz, M. Sjöström, and R. Olsson. Weighted Optimization Approach to Time-of-Flight Sensor Fusion. *IEEE Trans. Image Processing*, 23:214–225, 2014.
80. Sebastian Schwarz. *Gaining Depth: Time-of-Flight Sensor Fusion for Three-Dimensional Video Content Creation*. PhD thesis, Mittuniversitetet, Sweden, 2014.
81. S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006.
82. N. Snavely, C. L. Zitnick, S.B. Kang, and M. Cohen. Stylizing 2.5-D video. In *Proc. of 4th Int.l Symp. on Non-photorealistic Animation and Rendering*, pages 63–69. ACM, 2006.
83. Y. Soh, J.Y. Sim, C.S. Kim, and S.U. Lee. Superpixel-based depth image super-resolution. In *IS&T/SPIE Electronic Imaging*, pages 82900D–82900D. Int. Society for Optics and Photonics, 2012.

84. M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson, 2008.
85. E. Stoykova, A.A. Alatan, P. Benzie, et al. 3-D time-varying scene capture technologies – A survey. *IEEE Trans. on Circuits and Systems*, 17:1568–1586, 2007.
86. Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.
87. M. Tallón, S.D. Babacan, J. Mateos, et al. Upsampling and denoising of depth maps via joint-segmentation. In *Proc. of European Signal Processing Conference*, pages 245–249, 2012.
88. J.T. Thielemann, G.M. Breivik, and A. Berge. Pipeline landmark detection for autonomous robot navigation using time-of-flight imagery. In *Proc. Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1–7, 2008.
89. C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. Int. Conf. on Computer Vision*, pages 839–846, 1998.
90. L.P.J. Vosters, C. Varekamp, and G. de Haan. Evaluation of efficient high quality depth upsampling methods for 3DTV. In *IS&T/SPIE Electronic Imaging*, pages 865005–865005, 2013.
91. J.W. Weingarten, G. Gruener, and R. Siegwart. A state-of-the-art 3D sensor for robot navigation. In *Proc. Int. Conf. on Intelligent Robots and Systems*, volume 3, pages 2155–2160, 2004.
92. Wikipedia. Lidar. `http://en.wikipedia.org/wiki/Lidar`, 2014.
93. X. Xiang, G. Li, J. Tong, et al. Real-time spatial and depth upsampling for range data. *Transactions on Computational Science XII: Special Issue on Cyberworlds*, 6670:78, 2011.
94. Q. Yang, N. Ahuja, R. Yang, et al. Fusion of median and bilateral filtering for range image upsampling. *IEEE Trans. Image Processing*, 22:4841–4852, 2013.
95. Q. Yang, K.-H. Tan, and N. Ahuja. Real-time O(1) bilateral filtering. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 557–564, 2009.
96. Q. Yang, K.H. Tan, B. Culbertson, and J. Apostolopoulos. Fusion of active and passive sensors for fast 3D capture. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, pages 69–74, 2010.
97. Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
98. L. Yin, R. Yang, M. Gabbouj, and Y. Neuvo. Weighted median filters: a tutorial. *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, 43:157–192, 1996.
99. F. Yuan, A. Swadzba, R. Philippsen, et al. Laser-based navigation enhanced with 3D time-of-flight data. In *Proc. Int. Conf. on Robotics and Automation*, pages 2844–2850, 2009.
100. Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1330–1334, 2000.
101. Z. Zhang. Microsoft Kinect sensor and its effect. *IEEE MultiMedia*, 19:4–10, 2012.
102. J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32:899–909, 2010.
103. J. Zhu, L. Wang, R. Yang, and J.E. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
104. J. Zhu, L. Wang, R. Yang, et al. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(7):1400–1414, 2011.
105. Zinemath Zrt. The zLense platform. `http://www.zinemath.com/`, 2014.