# Algorithms for 3D Shape Scanning with a Depth Camera

Yan Cui[1], Sebastian Schuon[3], Sebastian Thrun[2], Didier Stricker[1], Christian Theobalt[3]
[1]Augmented Vision, DFKI [2]Stanford University [3]MPI Informatik
Yan.Cui@dfki.de, sebastian.schuon@stylight.de, thrun@cs.stanford.edu, Didier.Stricker@dfki.de, theobalt@mpi-inf.mpg.de

**Abstract**—We describe a method for 3D object scanning by aligning depth scans that were taken from around an object with a Time-of-Flight (ToF) camera. These ToF cameras can measure depth scans at video rate. Due to comparably simple technology they bear potential for economical production in big volumes. Our easy-to-use, cost-effective scanning solution which is based on such a sensor could make 3D scanning technology more accessible to everyday users. The algorithmic challenge we face is that the sensor's level of random noise is substantial and there is a non-trivial systematic bias. In this paper we show the surprising result that 3D scans of reasonable quality can also be obtained with a sensor of such low data quality. Established filtering and scan alignment techniques from the literature fail to achieve this goal. In contrast, our algorithm is based on a new combination of a 3D superresolution method with a probabilistic scan alignment approach that explicitly takes into account the sensor's noise characteristics.

**Index Terms**—Superresolution, Global alignment, Rigid transformation, Non-rigid transformation, 3D scanning, Time-of-Flight, Kinect

## 1 Introduction

Recent years have seen an enormous trend to employ new technology for visual media creation and presentation. New description languages and more powerful electronic devices make it easier to display 3D content in interfaces or create entire virtual worlds that are accessible through a web browser. All these technologies create a demand for high quality 3D models of real world objects that can be displayed in these 3D environments. Up until now creation of such models was a task performed by graphics specialists, who either employ specialized design software or use 3D scanning technology to capture a real object, Sect. 2. Today's scanning devices contain special-purpose sensors which are expensive and complicated to operate. This has prevented the spread of such technology, and thus few everyday people have access to such technology which would enable them to capture their own 3D models.

In this work we present a novel approach to 3D scanning, that employs a recently more popular new type of sensor for scanning, a so-called Time-of-Flight (ToF) camera [26], [24]. These sensors recover scene depth at video rate by measuring the travel time of infrared light, and have been around for a little while. So far, they remained expensive. However, when produced in large numbers, these sensors should be very affordable and as easy-to-operate as a normal camera. Recently Microsoft shipped the Kinect camera [31] to the mass market, which serves a similar purpose using an infrared-based active triangulation approach. This shows what potential lies in such sensors. Given the expected spread of Time-of-Flight or similar depth cameras and of algorithms like ours, we believe many new exciting applications of 3D models will arise: in games users could add their own object into the virtual world, or in online shopping people can easily digitalized their body and try new clothes on in a virtual fitting room.

Even though a Time-of-Flight camera returns 3D data directly, using such a device to produce high-quality scans is a challenge in itself. Most notably these devices were not designed as high quality 3D scanners, but rather for object detection and part of natural user interfaces. We therefore need to tackle their high noise level, low X/Y resolution and significant systematic measurement bias [1]. The results of the Kinect exhibit similar problems in data quality, even though the sensor resolution is slightly higher. If we overcome these challenges, we gain a scanning technology that can work in real-time, does not require complex scanning apparatus, and does not interfere with the scene in the visual spectrum (which would enable simultaneous capture of shape and appearance).

In the following sections we present an algorithm to create closed 3D models with a single ToF camera only. Our procedure works by either rotating the camera around the object or vice-versa by rotating the object in front of the camera. Our pipeline then creates closed 360 degree models of a static object whose reconstruction quality and resolution is way above what one could expect when looking at a single frame from a ToF sensor. See Fig. 1b for a sample input and note the quality that is achieved in the final model, Fig. 1c.

This paper builds on and extends our previous research on ToF-based 3D scanning [42], [10]. The biggest algorithmic challenge we face when putting this idea into practice is also the reason why ToF cameras have not yet taken over the 3D scanning market: ToF sensors have a very low X/Y resolution, an adverse random noise behavior, and a notable systematic measurement bias [1].

Digital Object Indentifier 10.1109/TPAMI.2012.190 

After a first look at the data quality of a single ToF depth scan, e.g., Fig. 1b, one may be tempted to not even try to use such a camera for shape scanning. However, in this paper we show that an appropriate combination of ToF specific resolution enhancement and scan alignment enables us to combine ToF scans taken from around an object into closed 3D shape models of reasonable quality, Fig. 1c.

In short, the new contributions in our algorithm (Sect. 3) are: 1) inherent ToF specific noise suppression; 2) a new ToF superresolution method that extends LidarBoost [42]; (3) a probabilistic scan alignment method that extends [10] by explicitly performing 3D loop closure that yields closed 3D models while handling the inherent systematic bias on-the-fly. We will show that all these steps are necessary to create models of high quality, whereas straight-forward traditional scan alignment and reconstruction methods fail. We present the algorithm's result on a variety of models captured with both Time-of-Flight cameras and the Kinect sensor, along with ground truth comparison against laser scanned data (Sect. 6).

## 2 Related Work

Most commercial systems for 3D shape scanning are based on structured light projection, laser stripe projection, or passive image-based reconstruction, please refer to [27], [15] for a recent overview. To build a complete 3D model, several scans taken from different viewpoints (usually under very controlled motion) are aligned. In contrast to ToF cameras, the above mentioned sensors provide rather clean data of relatively low random noise and systematic error. On such data, local rigid alignment techniques, such as Iterative Closest Points (ICP) and its variants [6] or global rigid alignment techniques, e.g. [5], [4], [17] can be used to register the scans against each other. Finally, a scan merging procedure, such as [12] can be applied to build a single 3D mesh. Hand-held scanners based on the above technologies have been proposed where the camera can be freely moved around an object (or vice versa), e.g. [40]. Given the ToF specifics the algorithms above render non-feasible. Nevertheless our work supports both hand-held scanning and scanning under controlled motion, e.g. with a turntable. A related idea to build a relatively simple 3D scanner has been proposed by Bouguet et al. [7] who measure 3D shape by recording a shadow cast by a rod moved over the object. However, freely moving the scanner around the object is not easy with this approach.

An alternative to active shape scanning are passive image-based approaches, such as stereo [41] or variants of shape-from-silhouette reconstruction [28], [30]. An extensive review of related work in this area would go beyond the scope of this article, but we would nonetheless like to highlight some recent developments in this domain. Stereo methods have been successfully applied to many reconstruction tasks. Newcombe et al. demonstrated a real-time system for SLAM-based camera tracking and stereo-based 3D scene reconstruction [35]. And recently they presented an approach that allows camera tracking even under sparse texture scene by matching dense depth maps [34]. However, there is a limit to the amount of detail that can be recovered. Similarly, silhouette-based approaches fail to capture concavities in objects, but a combination of shape-from-silhouette and photo-consistency-based refinement can help to overcome this issue [25]. An alternative to solely relying on images is to additionally take images under controlled illumination and to use shading or reflectance cues to further enhance the amount of reconstructed detail, such as in photometric stereo [43] and shape-from-shading [46]. Shading-based geometry reconstruction or stereo refinement under general uncontrolled illumination is also feasible [2], [45].

Time-of-Flight cameras [26], [24] have a variety of advantages over the above mentioned technologies (see Sect. 3 for details). Nevertheless, they have rarely been employed as sensors for 3D object scanning. This is mainly due to the challenging noise and bias characteristics [1] which renders direct application of established filtering and alignment approaches infeasible. Characterization of ToF sensor noise and ToF camera calibration is therefore a very active field of research [29], [14].

Some previous work tries to attack ToF camera deficiencies by combining them with normal color cameras. For instance, it is feasible to combine a stereo camera system with a Time-of-Flight sensor to obtain better reconstruction, also in sparsely textured areas [47], [3], [16]. Also in a multi-view setting, fusion of multi-view stereo and multi-view ToF recordings leads to more faithful reconstructions [22]. Recently, Reynolds et al. presented a machine learning approach to build a confidence measure for the quality of Time-of-Flight measurements that may be employed in many reconstruction applications [37].

In this paper we show that reliable shape capture is feasible with a single ToF camera alone. One component of our approach is a Time-of-Flight superresolution approach. We capitalize on recent developments in this domain [36], [42] and performed extensive experimental evaluation of the employed superresolution energy functionals to be minimized. Inspired by the image processing counterparts [44] and depth map smooth processing by a variational approach [38], we derived a new regularization term that leads to improved superresolution results compared to previous approaches, such as LidarBoost [42]. Related to ToF superresolution is the method by Kil et al. [21] for 3D superresolution with a laser scanner. The former approaches are designed for the more challenging noise characteristics of ToF cameras.

In addition, we present an extended scan alignment procedure with loop closure constraint tailored to ToF data. In our algorithm the systematic camera error is implicitly compensated during alignment. Some previous work already deals with global non-rigid shape align-


(a) Color Image (b) Raw Scan (c) Proposed Method (d) Laser Scan (e) Error Plot (f) Error Range

Fig. 1. ToF scan, antique head (a); our algorithm computes a 3D model of reasonable quality (c) despite severe errors in the raw ToF data (b). Reconstruction error (e) compared to a laser scan (d) shows that under no circumstance the error was larger than $2.5$ cm, while for most of the surface it was below $1.0$ cm. (note: raw aligned scans, no hole filling done)

ment [8], but not under consideration of ToF specifics. Related to our work is also research on surface reconstruction from noisy, but already aligned, scans [19], [13]. Our approach differs in that it extends previous work on probabilistic non-rigid alignment of pairs of scans [33] into a global method. Suitable rigid and non-rigid scan alignment is achieved by explicitly incorporating ToF specific noise characteristics, based on the work [10].

Recently, with KinectFusion a hand-held scanning approach with a Kinect camera has been presented [18]. It employs a combination of camera pose estimation, a fast variant of iterative closest-point-like scan alignment, and volumetric shape representation for scan integration. A fast GPU-supported implementation yields real-time performance. As stated before, the Microsoft Kinect sensor does not suffer from systematic distortions in the same way as ToF cameras, and thus faster alignment is possible. Direct application of the KinectFusion approach to ToF sensors would not be feasible. Nonetheless, we also show results of our method with the Kinect sensor, which illustrate that the denoising and superresolution aspects of our approach are also beneficial in that setting.

## 3 Our Algorithm

Our goal is to build a 3D shape scanner based on a Time-of-Flight camera that can be used in hand-held and turntable scanning mode [10]. Furthermore we demonstrate that our system works with the Kinect depth camera, which is an active-infrared stereo system [11]. As testbed for our experiments we use a MESA Swissranger SR4000 ToF camera. In a nutshell, it emits infrared light into the scene and at each pixel measures the return time of the reflected light from which it determines the depth of the pixel. More about the phase-shift based internal measurement principle of the SR4000 can be found, for instance, in [26], [24].

Depth cameras have a variety of conceptual advantages over previously used sensor techniques for shape scanning: they capture full frame depth at video rate and provide depth data with negligible latency as well as they do not need to subsequently scan scene points for a single depth map (like a laser scanner). As these cameras do not interfere with the scene in the visual spectrum, texture can be recorded simultaneously if an additional color camera is available. Current depth cameras are based on comparably simple CMOS technology, this enables low-cost manufacturing in large numbers. Additionally Time-of-Flight depth cameras do not rely on time multiplexing like structured light scanners (even though internally the ToF camera performs several measurements; for normal motion at normal speed this effect is negligible) and the measurement quality is largely independent from scene texture.

During acquisition the camera is moved by hand in an arc covering $360^\circ$ around the object. The object should stay roughly in the center of the field of view and the distance of the camera to the object is kept approximately constant, Fig 3. This way, a sequence of $i = 1, \ldots, N_c$ depth maps $D_i$ is obtained from $N_c$ subsequent positions along the camera path, each of which can be described by a tuple of the exponential map $M$ in conformal geometric algebra [32].

Unfortunately, the previously mentioned advantages of depth cameras come at the price of very low X/Y sensor resolution ($N_x = 176 \times N_y = 144$ pixels for the SR4000), random depth noise with significant standard deviation as an characteristic of ToF cameras with a substantial systematic measurement bias that distorts the depth maps [1], [23], see Figs. 1b, 4a and 9 for examples of raw ToF depth scans. Our new approach presented in this paper enables us to combine and align these rather low quality depth scans. The result is a 3D model of substantially higher quality than a single depth scan would suggest. It's algorithmic core and the main novelty is a combination of 3D depth superresolution with a new depth-specific probabilistic method for simultaneous rigid and non-rigid alignment of multiple depth scans. Our algorithm comprises the following steps, Fig. 2:

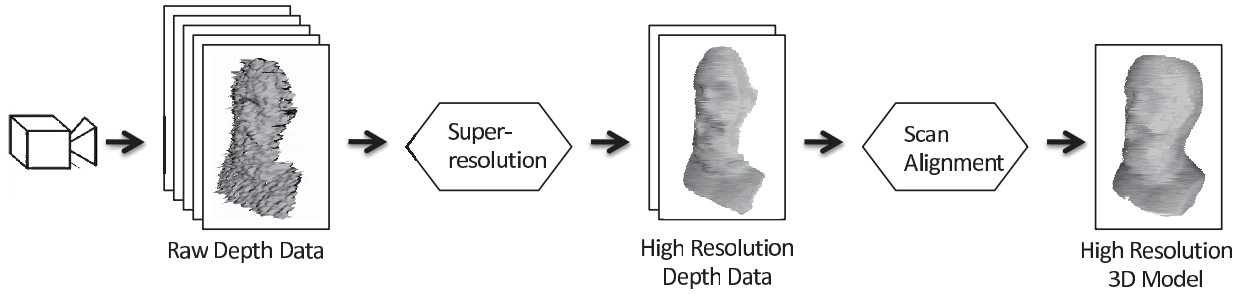



Fig. 2. Outline of our processing pipeline.

1) Superresolution: We subdivide the set of $N_c$ depth images into a set of $\ell = 1, \ldots, K$ chunks of depth images, Fig. 3. Each chunk $\mathcal{C}_\ell = (L_{\rho(\ell)}, \ldots, L_{\rho(\ell)+C})$ comprises $C$ subsequent depth images starting from a frame index $\rho(\ell)$. To each chunk of depth images a superresolution approach is applied, yielding $K$ new depth maps $H_\ell$ with much higher X/Y resolution, Section 4.

2) Scan Alignment: The $H_\ell$ are converted to 3D geometry $Y_\ell$ and aligned by means of a probabilistic alignment approach that not only recovers the rigid scan alignment parameters, but also compensates for non-rigid bias-induced deformations, Sects. 5.1 and 5.2. We will show that the explicit handling of this non-rigid component is essential for scanning with ToF cameras. A closed model will be reconstructed by explicit loop closure as detailed in Sect. 5.2.2. Finally, a closed surface is reconstructed from the point cloud [20].

## 4 Superresolution

To each chunk of frames $\mathcal{C}_\ell$, we apply a ToF superresolution approach which yields a high-resolution depth map aligned to the center frame of the chunk, similar to the LidarBoost approach [42]. In the following we briefly describe the core concepts of LidarBoost which our new superresolution approach inherits, and refer the reader to [42] for more detail. We then augment the LidarBoost concepts with a new regularization framework that yields better results.

First, all depth maps in the chunk are aligned to the center frame using optical flow, by computing a global displacement between each two frames. This is sufficiently accurate since the maximum viewpoint displacements throughout the entire chunk are typically one to two depth pixels. LidarBoost extracts a high-resolution denoised center depth map $H_\ell$, Fig. 2, from the aligned low resolution depth maps by solving an optimization problem of the form:

$$\min_{H_\ell} E_{\text{data}}(L_{\rho(\ell)}, \ldots, L_{\rho(\ell)+C}, H_\ell) + \lambda E_{\text{reg}}(H_\ell)\ . \quad (1)$$

Here, $L_{\rho(\ell)}, \ldots, L_{\rho(\ell)+C}$ are the low resolution depth maps of a given chunk, which will be upsampled to a resolution of $\beta N_X \times \beta N_Y$ and aligned to the center depth frame. The upsampling factor $\beta = 4$ is constant in both directions. We use nearest neighbor sampling from the low resolution images, which is preferable over any type of interpolated sampling. Interpolation implicitly introduces unwanted blurring that leads to a less accurate reconstruction of high-frequency shape details in the superresolved result. $E_{\text{data}}$ measures the agreement of $H_\ell$ with the aligned low resolution maps; unreliable depth pixels with low amplitude are discarded. $E_{\text{reg}}$ is a feature-preserving smoothing regularizer tailored to ToF data. The $H_\ell$ are straightforwardly converted into 3D point clouds $Y_\ell = \{y_j \mid j = 1, \ldots, \beta N_X \beta N_Y\}$ by reprojection into space using the ToF camera's intrinsic parameters (calibrated off-line).

Our new superresolution approach is based on a similar energy function and uses the same definition for $E_{\text{data}}$ as LidarBoost:

$$E_{data} = \sum_{k=1}^{C} \left\| W_k * \left( H_l - L_{\rho(l)+k} \right) \right\|^2 , \quad (2)$$

where $*$ denotes element-wise multiplication, $W_k \in R^{\beta N_X \times \beta N_Y}$ is a banded matrix that encodes the positions of $L_{\rho(l)+k}$ which one samples from during resampling on the highresolution target grid. $\|\cdot\|^2$ denotes the square sum of each element in the matrix, which is a $\ell_2-$norm. Previous depth superresolution methods, as well as many image superresolution methods, employ a $\ell_1-$norm. While a $\ell_1-$norm forces the depth value at a certain high-resolution grid point towards the median of registered low-resolution samples, an $\ell_2-$norm yields their mean. For very noisy data, the median is certainly reasonable since it rejects outliers. In contrast, the mean yields a smoother surface reconstruction, since the averaging cancels out recording noise. From our experience

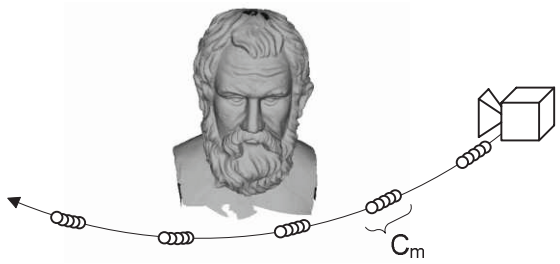
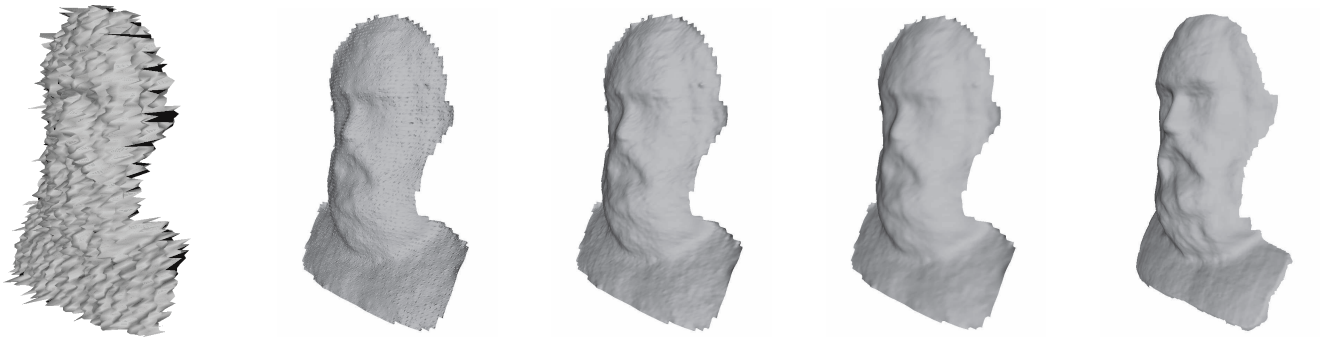



Fig. 3. A typical camera path: The dotted segments are the frame chunks $C_\ell$ from which superresolved depth scans are computed.



(a) Raw data (b) Linear operator (LidarBoost) (c) Total Variation (TV) (d) Isotropic non-linear (e) Anisotropic nonlinear

Fig. 4. Results with different regularization terms used in the superresolution algorithm.

using ToF data and our method, it is more beneficial to capitalize from the smoothing effect of a $\ell_2-$norm.

Our approach differs in the employed regularizer. To improve the superresolution performance, we experimented with four different versions of regularizer functions $\Psi$ (Eq. 3) that are inspired by their image processing counterparts [44]: linear, square nonlinear, isotropic nonlinear and anisotropic nonlinear, Tab. 1. Before we explain them in more detail, we define a gradient operator in the depth image domain similar to [42]:

$$E_{\text{reg}}(H_\ell) = \Psi(|\nabla H_l|^2) \tag{3}$$

$$\nabla H_{u,v} = \begin{pmatrix} G_{u,v}(0,1) \\ G_{u,v}(1,0) \\ \vdots \\ G_{u,v}(m,n) \end{pmatrix} \tag{4}$$

$$G_{u,v}(m,n) = \frac{X(u,v) - X(u+m, v+n)}{\sqrt{m^2+n^2}} \tag{5}$$

In contrast to [42], we evaluate finite differences in eight directions. Therefore, we set $m = -2 \cdots 2$, and $n = -2 \cdots 2$, and there are 24 neighbors for each pixel. Based on this definition, the different regularizers compare as follows (see also Tab. 1):

1) Linear Regularization (original LidarBoost), i.e., $E_{\text{reg}} = |\nabla H_l|^2$: The simplest form of $E_{\text{reg}}$ is a linear regularizer. However, with that regularization, a lot of spikes and artifacts in actually smooth regions appear, as Fig. 4b shows.
2) Total Variation (TV) Regularization, i.e., $E_{\text{reg}} = 2\,|\nabla H_\ell|$: As opposed to the linear case, the result exhibits no erroneous spikes and is smoother in regions that are smooth in reality. At the same time, important shape detail is better preserved, as Fig. 4c. However, some smaller shape detail is not reconstructed, or appears shallower than in reality.
3) Nonlinear Isotropic Regularization, i.e., $E_{\text{reg}} = 2\lambda^2 / g(|\nabla H_l|^2) - 2\lambda^2$: This regularizer prevents occasional mislocalization and blurring of edges that could still happen with LidarBoost. The regularizer allows isotropic nonlinear diffusion and introduces feedback into the process by adapting the diffusivity $g$ (Eq. 6) to the gradient of the superresolution result. Here $g$ is some decreasing function, in our case the Charbonnier diffusivity with $\lambda = 2.5$.

$$g(s^2) = \frac{1}{\sqrt{1 + s^2/\lambda^2}} \tag{6}$$

The reconstruction result (Fig. 4d) features a smooth surface with less reconstruction noise. Overall, the result is still similar to LidarBoost.
4) Nonlinear Anisotropic Regularizer, i.e., $E_{\text{reg}} = trG\left(\nabla H_\ell \nabla {H_\ell}^T\right)$: diffusion filters face difficulties when confronted with noisy edges. For instance, they cannot enhance coherent flow-like structures (e.g. the beard in the Head model). Anisotropic nonlinear regularization performs better here. Anisotropic nonlinear filtering employs a diffusion tensor instead of scalar-valued diffusivity.

$$G = (v_1 \dots v_{lm-1}) diag(\lambda_1 \dots \lambda_{lm-1}) (v_1^T \dots v_{lm-1}^T)^T \tag{7}$$

This way, it takes into account the direction of local structures $\nabla H_{u,v} \nabla H_{u,v}{}'$ (in Eq. 4), which are specified by their eigenvectors and eigenvalues. The function $G$ is defined as in Eq. 7, the eigenvectors $v_1, v_2 \dots v_{lm-1}$ and their eigenvalues $\lambda_1, \lambda_2 \dots \lambda_{lm-1}$ determine the diffusion tensor. If an eigenvalue is bigger than a threshold, one considers this direction as collinear with the edge direction. In this case, the eigenvalue is set $\lambda = 1$ for this direction. Otherwise, set $\lambda = g(|\nabla H_l|^2)$ for the direction across the edge, with a diffusivity $g$ as defined in Eq. 6. See also [44] for a background on these considerations. As Fig. 4e shows, the anisotropic non-linear diffusion regularizer leads to smooth reconstructions in truly smooth areas of the model. At the same time it preserves actual detail structure better than the alternative regularizers, and also brings out fine scale shape detail more

clearly (c.f. the eyes and the beard part shown in the head result). The average Euclidean errors between the Head model ground truth scan and the results with different regularizers are shown in Tab. 1. Not only visually but also quantitatively the anisotropic nonlinear regularizer leads to the most favorable results. We thus henceforth use the anisotropic nonlinear regularizer to get all the superresolved 3D point clouds.

Our implementation uses the Euler-Lagrange equations to transform the optimization problem into a linear equation system, which we solve in turn using the Gauss-Seidl method. The first derivative of $E_{\text{reg}}$ in Euler-Lagrange equations are shown in Table 2. Run times for our C++ implementation are also shown in Table 2. Note the significant improvement in runtime, with about 30 seconds compared to the earlier LidarBoost implementation that took up to two hours with a general solver for disciplined convex optimization [42].

## 5 Scan Alignment

### 5.1 Systematic Bias

While the random noise is effectively reduced by the superresolution approach, the ToF data's systematic bias leads to non-rigid ToF scan distortions and therefore needs special attention. Let $x_i$, $i = 1, \ldots, N_x \times N_y$ be a 3D point measured by the depth camera (i.e. the point in space after reprojection of the depth pixel), and $V_i$ be the direction of the camera ray towards the point. Then the bias can be modeled as a systematic offset $d_i$ along this ray which makes the camera measure the value $x_i = \tilde{x}_i + V_i d_i$ rather than the true 3D point $\tilde{x}_i$ (see Figure 5a). Previous studies have shown that the depth bias is pixel dependent and dependent on many factors, including the camera's integration time, scene reflectance, surface orientation, and distance [1], [23]. Therefore, the systematic bias cannot be handled by offline calibration beforehand.

Accounting for all such dependencies in our framework would render the problem intractable. We therefore make a few simplifying assumptions. First, in practice the bias dependency on reflectance, surface orientation, and integration (since it stays constant for a scan) can be neglected. Second, the depth range covered by the scanned object is usually limited and the distance of the camera to the object remains fairly constant. Therefore, we ignore the depth dependency of the bias. Finally, when averaging hundreds of depth frames of a flat wall, the resulting 3D model typically shows a radially symmetric deviation from the plane, with increasing curvature (bias) the further one is away from the depth image's center. We therefore assume that all depth pixels with the same radial distance from the image center have the same bias, and the bias increases with the radius. Since the superresolved depth maps $H_\ell$ are computed from closely spaced low resolution ToF scans, we assume the above bias characteristics also applies to them.

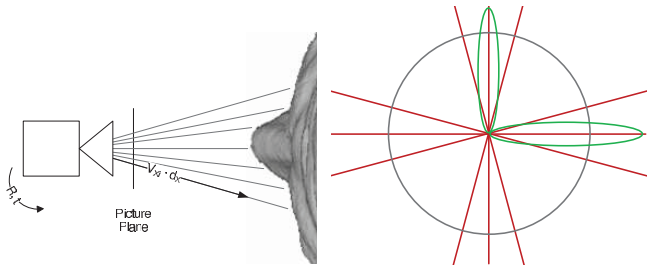



Fig. 5. (a) During multi-scan alignment, the motion of the points in the scan is parameterized by a rigid component $M$ and a non-rigid warp along the viewing ray direction (representing measurement bias). (b) Point sampling strategy during alignment (in pixel domain). The red line determines locations where the sparse sampling strategy looks at the reference set $X$; the green region is the sampling region for the $Y_\ell$.

The set of radially symmetric bias values therefore is parameterized as $(d_1, \ldots, d_{O(H_\ell)})$ with $O(H_\ell)$ being half the number of pixels on the diagonal of $H_\ell$.

### 5.2 Probabilistic Simultaneous Scan Alignment

The first step towards the reconstruction of a complete ToF-based 3D model is a probabilistic global alignment step between two scans that solves for the rigid alignment $M_\ell$ of all high-resolution 3D point clouds $Y_\ell$, as well as the non-rigid systematic bias values $(d_1, \ldots, d_{O(H_\ell)})$. Please note that, as opposed to prior work [10], we parameterize general rigid body transforms not as full matrices but with the twist and exponential map representation for rigid body transformations, as derived form conformal geometric algebra [32], [39], [9]. This representation is more compact and enables convenient linearization during the later optimization steps. A general rigid body transform is represented by a motor, and can be approximated as:

$$X' = MX\tilde{M} = E + e_\infty(x - l \cdot x - m) \qquad (8)$$

where $E$ is the identity matrix, $M$ is rotor, $\tilde{M}$ is its reverse, $e_\infty$ stands for the point at infinity, $l \in \Re^3$ is the rotation part, $m \in \Re^3$ is the translation part in conformal geometric algebra [32].

Rather than pre-compensating the bias, as in [23], we explicitly model the set of bias variables as unknowns of the alignment procedure. This enables us to accommodate for the potential scene-dependency of the bias to a certain degree while keeping the number of variables in reasonable bounds.

Our algorithm is inspired by the non-rigid registration approach for pairs of scans described in [33]. We extend their ideas to our setting and develop an approach for simultaneous rigid and non-rigid multi-scan alignment that incorporates knowledge about the ToF bias characteristics. ToF scan registration is formulated as a maximum-likelihood estimation problem. We choose

| Smoothing strategy | Differential operator $E_{\text{reg}}$ | First derivative in Euler-Lagrange | AEE (Head model) |
|---|---|---|---|
| Linear (original LidarBoost) | $\lvert\nabla H_l\rvert^2$ | $2\Delta H_l$ | 4.8325 |
| Total Variation (TV) | $2\lvert\nabla H_\ell\rvert$ | $2div\left(\left(1/\lvert\nabla H_\ell\rvert\right)\nabla H_\ell\right)$ | 3.7342 |
| Isotropic Nonlinear | $2\lambda^2/g(\lvert\nabla H_l\rvert^2) - 2\lambda^2$ | $2div\left(g\left(\lvert\nabla H_l\rvert^2\right)\nabla H_l\right)$ | 2.8933 |
| Anisotropic Nonlinear | $trG\left(\nabla H_\ell \nabla H_\ell{}^T\right)$ | $2div\left(G\left(\nabla H_l \nabla H_\ell{}^T\right)\nabla H_l\right)$ | 2.4331 |

TABLE 1
Different alternatives for regularization terms in the superresolution method.

any one of the high-resolution 3D point clouds as the reference 3D point set, henceforth termed $X = \{x_i \mid i = 1, \ldots, N_f\}$. For ease of explanation, in the following we describe the alignment process for a single point cloud $Y$. Thereafter, we show how this process is applied to all scans simultaneously, enabling full 360 degree reconstruction with implicit loop closure enforcement.

#### 5.2.1 Pairwise Alignment

On each point in $Y$, a multi-variate Gaussian is centered. All Gaussians share the same isotropic covariance matrix $\sigma^2 I$, $I$ being a $3 \times 3$ identity matrix and $\sigma^2$ the variance in all directions. Hence the whole point set can be considered a Gaussian Mixture Model (GMM) with density:

$$p(x) = \sum_{m=1}^{N_g} \frac{1}{N_g} p(x|m) \quad with \quad x|m \propto N(y_m, \sigma^2 I)\ . \tag{9}$$

Alignment of $Y$ to $X$ is performed by maximizing the likelihood function. In our setting, the motion of $Y$ towards $X$ is parameterized by the rigid motion component $M$ as Sect. 5.2, as well as the bias motion component, i.e., a translation of each point $y_j$ along the local viewing ray direction $V_j$ by the bias factor $d_j$, Sect. 5.1. Please note that while the rigid component for $X$ remains fixed, the bias-induced non-rigid deformation is also applied to $X$ when searching for optimal alignment. The maximum likelihood solution for $M$ and $d_1, \ldots, d_{O(H_\ell)}$ is found by minimizing the negative log-likelihood which yields the following energy functional:

$$\begin{aligned} &E(M, d_1, \ldots d_O) = \\ &-\sum_{n=1}^{N_f} \log \sum_{m=1}^{N_g} \exp\left(-\tfrac{1}{2}\left\|\frac{x_n + V_{x_n} d_n - M(y_m + V_{y_m} d_m)\widetilde{M}}{\sigma}\right\|^2\right) \\ &+\lambda \left\|(d_1 - d_2, \ldots, d_{O-1} - d_O)\right\|^2\ . \end{aligned} \tag{10}$$

The variance $\sigma^2$ of the mixture components is estimated using

$$\sigma^2 = \frac{1}{N_f N_g} \sum_{n=1}^{N_f} \sum_{m=1}^{N_g} \left\|x_n - y_m\right\|^2\ . \tag{11}$$

Note that our energy functional contains a regularization term weighted by $\lambda$ ($\lambda = 3$ in all our experiments) which ensures a smooth distributions of the bias values. In accordance with Sect. 5.1 the term favors monotonous bias distributions with increasing radius from the image center.

We use an iterative Expectation Maximization (EM) like procedure to find a maximum likelihood solution of Eq. (10). During the E-step the best alignment parameters from the previous iteration are used to compute an estimate of the posterior $p^{old}(m|x_n)$ of mixture components by using Bayes theorem. During the M-step, new alignment parameter values are found by minimizing the negative log-likelihood function, or more specifically, its upper bound $Q$ which evaluates to:

$$\begin{aligned} &Q(M, d_1, \ldots, d_O) = \\ &\sum_{n=1}^{N_f} \sum_{m=1}^{N_g} P^{old}(m|x_n) \frac{\left\|x_n + V_{x_n} d_n - M(y_m + V_{y_m} d_m)\widetilde{M}\right\|^2}{2\sigma^2} \\ &+\lambda \left\|(d_1 - d_2, \ldots, d_{O-1} - d_O)\right\|^2\ . \end{aligned} \tag{12}$$

The above EM procedure converges to a local minimum of the negative log-likelihood function. Please note that the variances $\sigma^2$ are continuously recomputed which is similar to an annealing procedure in which the support of the Gaussians is reduced when point sets get closer.

#### 5.2.2 Global Alignment

To archive a closed model, all frames from a scan need to be aligned globally. A well-known approach is using Iterative-Closest-Point (ICP). Doing so does not yield satisfactory results, as Fig. 6 shows. Here the point clouds colored in brown were aligned with ICP only. Clear alignment errors appear, due to two reasons - subsequent frames are only aligned pairwise which leads to an error accumulation as well as non-rigid distortions in the ToF scans are not accounted for. To enable closed model reconstruction and still allow for compensation of bias-induced distortions, we developed the following strategy:

To start with, an ICP based alignment of all superresolved scans is used to roughly identify for each scan a set of corresponding scans. This yields scans that contain data in close spatial proximity and thus should ideally overlap. Given this set of corresponding scans, we can now globally optimize the alignment using our previously described alignment approach that also explicitly compensates the non-rigid transformations due to the systematic bias. This is a clear step forward in comparison with our previous work [10], where subsequent superresolved scans were only pairwise aligned and thus could not guarantee to achieve loop-closed models.

As the global alignment problem is a joint optimization of many pair-wise alignments, the resulting energy

function incorporates equations that take form of Eq. 10. The energy function for global alignment thus takes form

$$E(M_1,\ldots,M_k,D_1,\ldots,D_k) = \\ -\sum_{f,g}\sum_{n=1}^{N_f}\log\sum_{m=1}^{N_g}\exp\frac{\left\|M_f(x_n+V_{x_n}d_n)\widetilde{M_f}-M_g(y_m+V_{y_m}d_m)\widetilde{M_g}\right\|}{-2\sigma^2} \\ +\lambda\sum_{j=1}^{K}\left\|(d_{j,1}-d_{j,2},\ldots,d_{j,O-1}-d_{j,O})\right\|^2 . \tag{13}$$

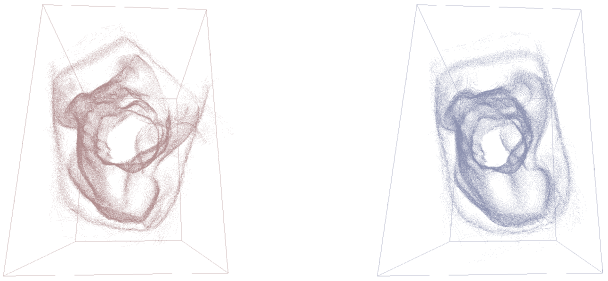
where $M_j$ is the rigid alignment, and $D_j$ the set of bias values for each frame. There are total $K$ frames, for each corresponding frames $f$ and $g$, add the cost to the energy function. Experimentally we determined that it is best to solve for all $M_j$ as well as $D_j$ in batch, using the same two-step EM like procedure explained before.

To render our problem tractable we had to introduce some simplifications: First, for each pair of corresponding frames we only compute the alignment using a reduced number of points from the scans, namely those points that are spatially the closest as determined by the previous local ICP step. In our experiments we determined a number of 20 points from that region to be sufficient. For the number of iteration we limit ourselves to three, and we use the first order Taylor series expansion of exponential maps to express the motors, as explained earlier in this section. Furthermore the non-rigid bias values in $Q$ are not evaluate for all 3D points in the respective scans, but only for a subset of samples from two corresponding point sets. Fig. 5b illustrates this sampling pattern in the pixel (depth image) domain. Here the red lines are the depth pixels (3D points) of the current frame which we limited ourselves to, and the green elliptical regions refer to points in the corresponding frame which are considered. Finally, for camera paths covering a larger viewpoint range, we perform several global alignments to several reference scans, such that sufficient overlap is guaranteed.

Fig. 8 illustrates that all the steps we propose are indeed relevant, only the joint consideration of both rigid and non-rigid alignment leads to final 3D models of good quality.

## 6 Results

We have tested our approach with six different test objects: an antique head (height 31 cm), a buddha statue (height 59 cm), a sculpture of two angels (height 34 cm), a lion statue (height 18 cm), a human statue (height 56 cm), a standing angel (height 24 cm), Fig. 1 and Fig. 9. For each object we also captured a ground truth model with a Minolta Vivid 3D scanner. We have chosen these objects due to their reasonable size for hand-held scanning and due to a lot of fine surface detail below the physical ToF resolution. Of each object we captured approximately thousand frames taken from a path covering 360° around the object. The distance of the camera to the object was about 100 cm in all cases, well within the total range of 7 m of the Swissranger.



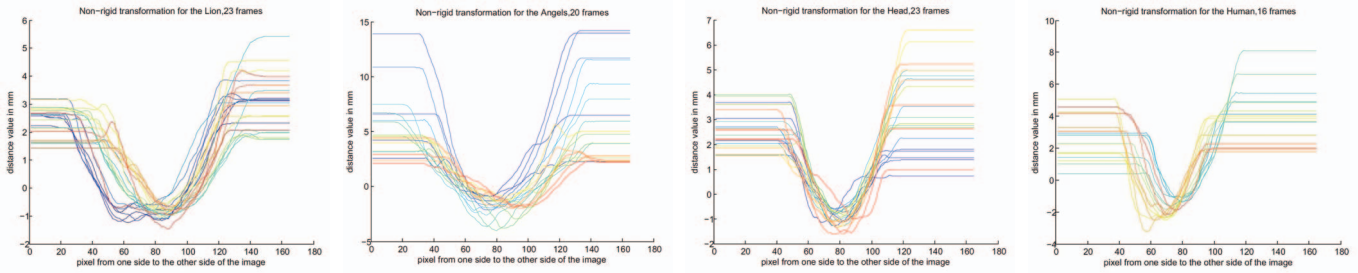
(a) Lion with ICP only (b) Lion with global align

Fig. 6. Point clouds of lion aligned with ICP only (a) and with our global rigid and non-rigid alignment procedure (b). In the first case, obvious misalignments between the scans exist, which are due to non-rigid distortions and missing loop closure constraints.

The objects were scanned with a hand-held setup where the camera was manually moved around the object. And the object is placed in a plain, to separtae it from the foreground by distance thresholding. The integration time of the Swissranger 4000 was set to 35 ms. The data sets and results are available for download.

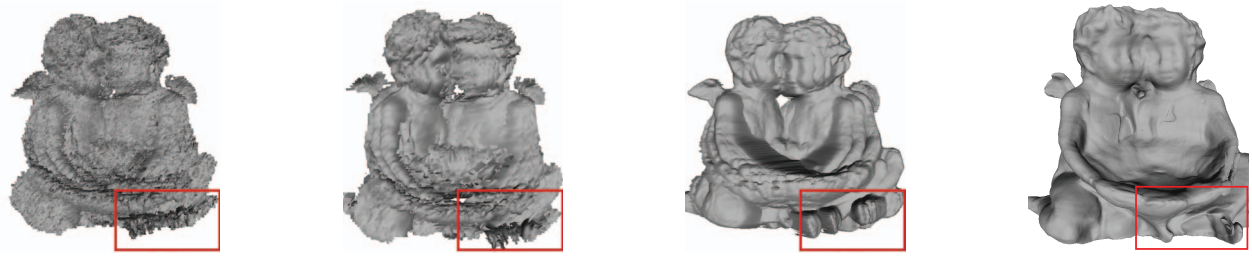
For all scenes we determined key frames (as Table 2), i.e., centers of chunks of frames for which superresolved scans were reconstructed. Subsequent key frames are typically 100 original depth frames apart, such that a typical key frame sequence would be $5, 105, \ldots, 1105, 1195 \ldots$. The chunk size was always $C = 10$ frames. All of the 3D renderings of our results in this paper were created by converting the individual 3D scans (low-resolution and high-resolution) to meshes using Poisson surface reconstruction [20]. Run times for reconstructions were measured on the a standard PC with an Intel(R) CPU 2.67GHz and 12GB RAM memory, and are shown for each object in Table 2.

Overall, the results of our method show that decent 3D models can also be captured with apparently low quality sensors. In a single ToF scan (Fig. 2 and Fig. 9) one can hardly even recognize the overall type of an object, let alone fine detail. Please note that in some raw depth scans depth pixels with low return amplitude have been discarded, as they are considered unreliable by the superresolution approach. The superresolution approach creates high-resolution denoised depth scans for certain viewpoints (the chunks), as shown in Fig. 4. By this means, the resolution and visible detail in each such scan is already dramatically improved. Also, as shown earlier, the new regularization term in the superresolution framework leads to clearly improved superresolved depth maps, as compared to the original LidarBoost approach. But each map only covers a part of the object and suffers from nonlinear distortion due to the camera bias. Our final alignment registers the scans into a 3D model of reasonable quality, Fig. 9 and Fig. 1. A lot of finer scale detail that lies close to the resolution limit of the ToF camera comes out in the final results, such as




(a) $d$ distribution for Lion (b) $d$ distribution for Angels (c) $d$ distribution for Head (d) $d$ distribution for Human

Fig. 7. The estimated non-rigid bias components plotted for a number of globally aligned frames for four different models. Each plot is the bias distribution along one scan line in the superresolved depth map. Depth pixels for which no value exists are clamped to the nearest bias values (constant segments of the curves left and right).



(a) ICP with 70 Raw Inputs (b) ICP with 7 Averaged Raw Inputs (c) ICP with 7 Superresolved Inputs (d) Proposed Method

Fig. 8. All steps in our pipeline are important. (a) Aligning 7 chunks of frames (10 frames each) with ICP fails (severe noise, wrong alignment - e.g. around feet (red)). (b) Averaging 10 frames from the center view of each chunk center reduces noise, but does not provide more resolution and ICP alignment fails. (c) Superresolution of the chunks boosts resolution, but ICP fails due to non-rigid distortions. (d) Our method produces clear detail and correct alignment. (Note: no hole-filling was done; triangles at occlusion edges were filtered).

the petals of the flower in the angels sculpture, or the hair detail of the head. In addition, our global alignment procedure enables us now to build closed 360 degree models from sets of superresolved depth scans.

Fig 7 show the bias distribution against the pixel distance from one side of the depth map to the other side for four objects. The bias distribution is fairly stable across objects, and follows our assumption of monotonous increase with radius. There are slight differences in the bias values found for different frame pairs to be aligned. This can be explained by differences in camera to scene alignment, for instance the varying distance of the camera to the object. There are also slight scene dependencies in the bias distributions, which our algorithm can accommodate for in certain bounds.

Of course, the resolution and shape quality of our models cannot rival that obtainable with a laser scanner, Fig. 9 and Fig. 1, and we never claimed that. Nonetheless, our models are of sufficient quality for many applications were sub-millimeter accuracy is not required, a result that was at first unexpected looking at the raw ToF quality.

**Validation** We also quantitatively measured the reconstruction quality against the laser scanned ground truth. As one can see in the color-coded reconstruction error renderings in Fig. 9 and Fig. 1 our models compare very favorably. With the error range bar in Fig. 1f we note that in most areas the error is below $1.0$ cm, and there are only a few outliers. In the figures, areas where there is no ground truth geometry in the laser scan (e.g. holes due to occlusion) are rendered in grey.

Throughout our experiment we used the same parameters for the superresolution pipeline as the LidarBoost ($\lambda_{SR} = 1,5 \times 5$ regularizer region, see [42]) algorithm and our alignment procedure ($\lambda = 3$), but with anisotropic nonlinear smooth term. This shows that our method is rather stable and does not require scene dependent parameter tuning.

**Limitations** Our approach is subject to a few limitations: the ToF camera fails to capture good data for certain surface materials, like highly specular objects. However, other scanners suffer from similar limitations. While scan acquisition is straightforward and fast, the runtime of scan reconstruction is notable. However, our algorithm lends itself very much to parallelization.

Despite these limitation, we were able to demonstrate that good quality 3D shape scanning without manual correction is also feasible with low quality sensors.


Fig. 9. Results of our method - in each row: test object; single ToF depth scan; our reconstructed model front and behind; Color-coded distance error against laser scan.

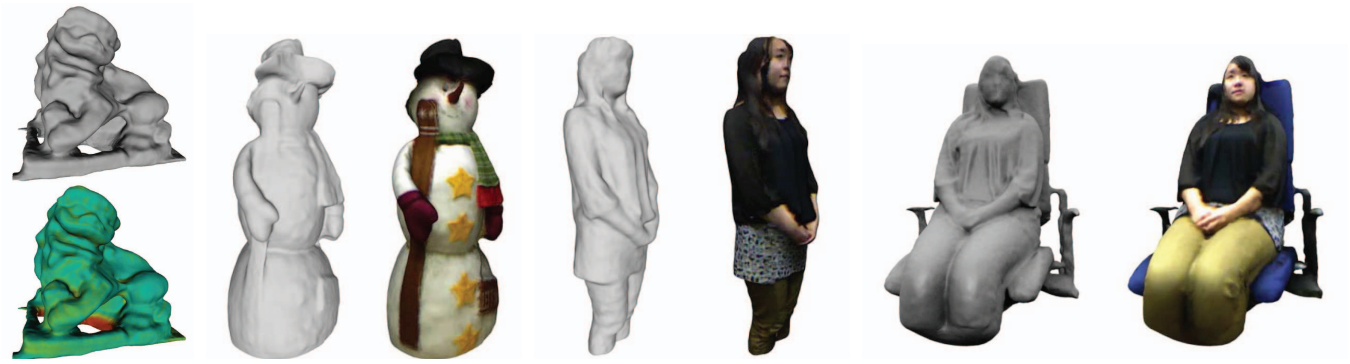
### 6.1 Results with a Kinect camera

We also tested our approach with a Kinect camera, which is not a Time-of-Flight system but an active stereo sensor. In addition to a ToF camera, the Kinect also captures a color image of a scene. As stated before, the Kinect data do not suffer from the same non-rigid distortions as ToF cameras, which is why more efficient real-time alignment approaches are also feasible [18]. However, our processing pipeline still bears advantages also when processing Kinect data. The resolution of the Kinect camera is limited (depth: $320 \times 240$, color: $6400 \times 480$ @30 frames/sec) and the data exhibits random noise, which is why a superresolution approach like ours enables a boost of reconstruction quality. Considering the slightly different characteristics of the Kinect camera, we modified our processing pipeline in two ways:

1) Color data can be used during superresolution. During superresolution, sub-pixel scan alignment can be computed based on optical flow in the color image rather than $x, y$ two displacement flow on the depth maps as in the original LidarBoost method. The quality of flow alignment can be used

| | Nr. key frames | Superresolution | Initial rigid alignment | Non-rigid alignment | Poisson recon | Combined |
|---|---|---|---|---|---|---|
| ToF-Lion | 23 | 28sec | 110sec | 61sec | 68sec | 267sec (4.45min) |
| ToF-Angels | 20 | 24sec | 102sec | 58sec | 79sec | 263sec (4.38min) |
| ToF-Head | 23 | 27sec | 115sec | 66sec | 68sec | 276sec (4.6 min) |
| ToF-Mannequin | 16 | 20sec | 93sec | 49sec | 65sec | 227sec (3.78min) |
| ToF-Angel 2 | 20 | 25sec | 109sec | 55sec | 83sec | 272sec (4.53min) |
| Kinect-Lion | 23 | 30sec | 111sec | 51sec | 64sec | 256sec (4.27min) |
| Kinect-Snowman | 35 | 31sec | 121sec | 61sec | 68sec | 281sec (4.68min) |
| Kinect-Humanstanding | 30 | 28sec | 118sec | 66sec | 66sec | 278sec (4.63min) |
| Kinect-Humansitting | 32 | 29sec | 119sec | 71sec | 77sec | 296sec (4.93min) |

TABLE 2
Run times and other parameters used for creating the results in this paper.



(a) (b) (c) (d)

Fig. 10. Results of Kinect scanning, Lion (Mesh and Error plot), Snowman, Human Standing, Human Sitting results of Mesh and Textured model.

to additionally weight the contributions of different depth maps in the superresolution data term. Additionally when we scaning non-rigid object (e.g. the human standing and human sitting model), we can improve depth scan alignment by running optical flow in the color images with pixel-based displacements.

2) the non-rigid distortions are neglected during global scan alignment.

We show results for four different scenes, a lion model, snowman statue, a person and a person sitting on a chair. For the first two scenes, the Kinect was moved manually around the scene, in the latter two case the human was moved in front of a static camera. Lens distortion has been corrected with a commonly available calibration matrix for Kinect. As before, 10 subsequent frames are combined into one superresolved depth map. The numbers of key frames (i.e. high-res depth maps) used, as well as the processing times are shown in Table 2. As one can see in Fig. 10, the results in all three cases are of very good quality, and exhibit a high amount of detail. The Lion model result yields an error below 3mm for 90% of the points. Texturing the results with the RGB data is also possible. As before, final surface models were generated with Poisson surface reconstruction. The results show that also with the Kinect camera, which has slightly different sensor characteristics, high-quality 3D model reconstruction is feasible with our proposed method.

## 7 Conclusion

In this paper we demonstrated that 3D shape models of static objects can also be acquired with a Time-of-Flight sensor that, at first glance, seems completely inappropriate for the task. The key in making this possible is the effective combination of 3D superresolution with a new probabilistic multi-scan alignment algorithm tailored to ToF cameras. In future, we plan to investigate approaches for real-time shape scanning, as well as incorporation of more sophisticated noise models into the reconstruction framework. Another interesting line of research is to give the user feedback, if enough data has been collected to compute the object.

## References

[1] D. Anderson, H. Herman, and A. Kelly. Experimental characterization of commercial flash ladar devices. In *Proc. of ICST*, 2005.
[2] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *In IEEE CVPR*, pages 374–381, 2001.
[3] C. Beder, B. Bartczak, and R. Koch. A combined approach for estimating patchlets from pmd depth images and stereo intensity images. In *Proc. DAGM*, pages 11–20, 2007.
[4] R. Benjemaa and F. Schmitt. A solution for the registration of multiple 3d point sets using unit quaternions. In *Proc. ECCV '98 II*, pages 34–50, 1998.
[5] R. Bergevin, M. Soucy, H. Gagnon, and D. Laurendeau. Towards a general multi-view registration technique. *IEEE PAMI*, 18(5):540–547, 1996.
[6] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE PAMI.*, 14(2):239–256, 1992.
[7] J.-Y. Bouguet and P. Perona. 3d photography on your desk. In *Proc. ICCV*, page 43. IEEE, 1998.
[8] B. Brown and S. Rusinkiewicz. Global non-rigid alignment of 3-D scans. *ACM (Proc. SIGGRAPH)*, 26(3), Aug. 2007.

[9] Y. Cui and D. Hildenbrand. Pose estimation based on geometric algebra. 2009.
[10] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010). - Pt. 2*, pages 1173–1180, San Francisco, USA, 2010. IEEE.
[11] Y. Cui and D. Stricker. 3d shape scanning with a kinect. In *ACM SIGGRAPH Posters*, 2011.
[12] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proc. SIGGRAPH*, pages 303–312. ACM, 1996.
[13] J. R. Diebel, S. Thrun, and M. Brünig. A bayesian method for probable surface reconstruction and decimation. *ACM TOG*, 25(1):39–59, 2006.
[14] M. Erz and B. Jhne. Radiometric and Spectrometric Calibrations, and Distance Noise Measurement of TOF Cameras. In R. Koch and A. Kolb, editors, *3rd Workshop on Dynamic 3-D Imaging*, volume 5742 of *Lecture Notes in Computer Science*, pages 28–41. Springer, 2009.
[15] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *CVPR '07*, 2007.
[16] U. Hahne and M. Alexa. Depth imaging by combining time-of-flight and on-demand stereo. In *Dyn3D*, pages 70–83, 2009.
[17] Q.-X. Huang, S. Flöry, N. Gelfand, M. Hofer, and H. Pottmann. Reassembling fractured objects by geometric matching. *ACM TOG.*, 25(3):569–578, 2006.
[18] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proc. of ACM UIST*, 2011.
[19] P. Jenke, M. Wand, M. Bokeloh, A. Schilling, and W. Straer. Bayesian point cloud reconstruction. *Computer Graphics Forum*, 25(3):379–388, September 2006.
[20] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.
[21] Y. J. Kil, B. Medereos, and N. Amenta. Laser scanner superresolution. In *Point-based Graphics*, 2006.
[22] Y. Kim, T. C., J. Diebel, J. Kosecka, and B. Mikusic. Multi-view image and tof sensor fusion for dense 3d reconstruction. In *Proc. of 3DIM 2009*, 2009.
[23] Y. Kim, D. Chan, C. Theobalt, and S. Thrun. Design and calibration of a multi-view tof sensor fusion system. In *Proc. CVPR Worksh. TOF-CV*, pages 1–7, 2008.
[24] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-Flight Sensors in Computer Graphics. *EUROGRAPHICS STAR report*, 2009.
[25] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. Technical report, Rochester, NY, USA, 1998.
[26] R. Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. *Diss., University of Siegen*, 2000.
[27] D. Lanman and G. Taubin. Build your own 3d scanner: 3d photograhy for beginners. In *SIGGRAPH courses*, pages 1–87. ACM, 2009.
[28] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE PAMI*, 16(2):150–162, 1994.
[29] M. Lindner, I. Schiller, A. Kolb, and R. Koch. Time-of-flight sensor calibration for accurate range sensing. *Computer Vision and Image Understanding*, 114(12):1318 – 1328, 2010.
[30] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, pages 369–374, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.
[31] Microsoft. http://www.xbox.com/en-US/kinect.
[32] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC, 1 edition, March 1994.
[33] A. Myronenko, X. Song, and M. Carreira-Perpinan. Non-rigid point set registration: Coherent Point Drift. *NIPS*, 19:1009, 2007.
[34] R. Newcombe, S. Lovegrove, and A. Davison. Dtam: Dense tracking and mapping in real-time. In *ICCV*, 2011.
[35] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *CVPR*, pages 1498–1505. IEEE, 2010.
[36] A. Rajagopalan, A. Bhavsar, F. Wallhoff, and G. Rigoll. Resolution enhancement of pmd range maps. *Pattern Recognition*, pages 304–313, 2008.
[37] M. Reynolds, J. Doboš, L. Peel, T. Weyrich, and G. J. Brostow. Capturing time-of-flight data with confidence. In *CVPR*, 2011.
[38] L. Robert and R. DERICHE. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In *ECCV*, 1996.
[39] B. Rosenhahn. *Pose estimation revisited*. PhD thesis, Universität Kiel, September 2003.
[40] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. In *Proc. SIGGRAPH*, pages 438–446. ACM, 2002.
[41] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002.
[42] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. *Proc. CVPR*, 2009.
[43] G. Vogiatzis and C. Hernández. Practical 3d reconstruction based on photometric stereo. In *Computer Vision: Detection, Recognition and Reconstruction*, pages 313–345. 2010.
[44] J. Weickert and H. E. Hagen. *Visualization and Processing of Tensor Fields*. Springer, Berlin, 2006.
[45] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, 2011. IEEE.
[46] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Tans. PAMI*, 21(8):690–706, 1999.
[47] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. *Proc. CVPR*, 0:1–8, 2008.

**Yan Cui** received his M.S. degree from Visual Computing group leaded by Prof. Dr. Weickert in Saarland university (2009). Studying and researching the topics of differential equations, optical flow and feature descriptor in Image Processing and Computer Vision.

Now Yan Cui is PhD student in the Augmented Vision group at the German Research Center for Artificial Intelligence in the context of the CAPTURE Project. The CAPTURE project is related to large 3D scene reconstruction with spherical and high dynamic range images. He is also very interested in 3D reconstruction and tracking human pose with Time-of-Flight and Kinect camera in real time.

**Sebastian Schuon** is Chief-Technology-Officer and Co-Founder at STYLIGHT GmbH. Here he helps to bring innovations from Computer Vision to the fashion eCommerce space. He grew to company to more than 20 employees and was listed with STYLIGHT among the Top100 European Startups by TechCrunch. His research interest is in all camera related fields, more recently in bringing innovations from Computer Vision to end users.

Sebastian was a Fulbright Scholar to Stanford University, from where he holds also a Masters degree. Furthermore he received degrees from Technische Universitaet Muenchen as well as Center for Digital Technology and Management.

**Sebastian Thrun** Sebastian Thrun is a research professor of computer science at Stanford University and a vice president/fellow at Google. At age 39, he was elected into the National Academy of Engineering and the German Academy of Sciences Leopoldina. Thrun has won international acclaim by winning the DARPA Grand Challenge, and by leading the Google Self-Driving Car project. He has been teaching the largest online class on computer science ever taught, with a 160,000 students signed up. Fast Company named Thrun the 5th most creative person in business in the World, and the Max Planck Society bestowed its research price (750k Euro) on Sebastian. Thrun pursues research in robotics, artificial intelligence, and human computer interaction. He has co-authored over 350 scientific articles and 11 books. He is one of the most cited authors in the field of artificial intelligence.

**Didier Stricker** is Professor at the University of Kaiserslautern and Scientific Director at the German Research Center for Artificial Intelligence (DFKI GmbH) in Kaiserslautern where he leads the new research department Augmented Vision.

From 2002 to June 2008 Didier Stricker lead the department Virtual and Augmented Reality at the Fraunhofer Institute for Computer Graphics (Fraunhofer IGD) in Darmstadt, Germany. In this function he initiated and participated to many national and international projects in the areas of computer vision and virtual and augmented reality. In 2006 he received the Innovation Prize of the German Society of Computer Science. Didier Stricker serves as reviewer for different European or National research organizations, and is a regular reviewer for the most important journals and conferences in the areas of VR/AR and Computer Vision.

**Christian Theobalt** is the head of the research group ”Graphics, Vision, and Video” at the Max-Planck-Institut Informatik and a Professor of Computer Science at Saarland University, Saarbruecken, Germany. From 2007 until 2009 he was a Visiting Assistant Professor in the Department of Computer Science at Stanford University. He received his MSc degree in Artificial Intelligence from the University of Edinburgh, Scotland, and his Diplom (MS) degree in Computer Science from Saarland University, in 2000 and 2001 respectively. From 2001 to 2005 he was a researcher and PhD candidate in Hans-Peter Seidel's Computer Graphics Group at MPI Informatik. In 2005, he received his PhD (Dr.-Ing.) from Saarland University and MPI.

His core research interest are problems that lie on the boundary between the fields of Computer Vision and Computer Graphics, such as dynamic 3D scene reconstruction and marker-less motion capture, computer animation, appearance and reflectance modeling, machine learning for graphics and vision, new sensors for 3D acquisition, advanced video processing, as well as image- and physically-based rendering.

For his work, he received several awards including the Otto Hahn Medal of the Max-Planck Society in 2007, and the EUROGRAPHICS Young Researcher Award in 2009.

**Acknowledgement:** Yan Cui has been jointly supervised by Didier Stricker and Christian Theobalt for this project.