A Pipeline for Building 3D Models using Depth Cameras

Avishek Chatterjee Suraj Jain Venu Madhav Govindu*

Department of Electrical Engineering
Indian Institute of Science
Bengaluru 560012 INDIA
{avishek suraj.jain venu}@ee.iisc.ernet.in

ABSTRACT

In this paper we describe a system for building geometrically consistent 3D models using structured-light depth cameras. While the commercial availability of such devices, i.e. Kinect, has made obtaining depth images easy, the data tends to be corrupted with high levels of noise. In order to work with such noise levels, our approach decouples the problem of scan alignment from that of merging the aligned scans. The alignment problem is solved by using two methods tailored to handle the effects of depth image noise and erroneous alignment estimation. The noisy depth images are smoothed by means of an adaptive bilateral filter that explicitly accounts for the sensitivity of the depth estimation by the scanner. Our robust method overcomes failures due to individual pairwise ICP errors and gives alignments that are accurate and consistent. Finally, the aligned scans are merged using a standard procedure based on the signed distance function representation to build a full 3D model of the object of interest. We demonstrate the performance of our system by building complete 3D models of objects of different physical sizes, ranging from cast-metal busts to a complete model of a small room as well as that of a complex scale model of an aircraft.

1. INTRODUCTION

Building accurate 3D models has been of interest in the fields of computer vision. Contemporary advancements in feature matching methods as well as optimisation techniques have lead to the development of structure-from-motion (SfM) systems that can handle thousands of camera images [14]. While such SfM systems have benefited from good feature representations like SIFT [10] as well as the availability of computationally efficient implementations of bundle adjustment [9] and dense multiview stereo [6], the overall computational requirements as well as implementational challenges are significant. The recent availability of consumer-range RGB-D cameras, i.e. Kinect has vastly simplified the acquisition of 3D scene structure. Using an infra-red structured-light stereo sensor, Kinect provides a dense 3D measurement of the scene as a 640×480 depth image at a frame rate of 30Hz. The easy availability of such a 3D depth camera has opened up many exciting possibilities including human pose estimation, tracking, human-computer interfaces etc.

In this paper, we are concerned with the use of Kinect to build complete 3D models. While Kinect provides depth images easily, these depth images have high noise levels that need to be overcome to be able to build the requisite 3D models. An example depth image rendered as a 3D scan is shown as an input in Fig. 1. While such depth images are satisfactory for tasks such as pose estimation, gesture recognition etc., in their raw form they are inadequate to build an accurate 3D model. Since any depth image from a single viewing direction can only cover a part of a 3D surface, building a complete representation entails observing the scene from multiple viewpoints. In turn, to build a global 3D surface representation we need to align or register the individual depth images (equivalently, 3D scans) in a single co-ordinate frame of reference. Such alignment requires the specification of 6 degrees of freedom of the Euclidean motion (3D rotation and translation) of every scan with respect to a global frame of reference. We denote the 4×4 matrix for Euclidean alignment as $\mathbf{M} \in SE(3)$ where $SE(3)$ is the Special Euclidean group. The canonical solution for the problem of aligning a pair of scans is the iterative closest point (ICP) method [2]. The ICP method is a greedy, iterative algorithm that is guaranteed to converge to a local minima. An ICP iteration consists of two steps : (a) after applying the current relative motion estimate to align the two scans, for every point in the first scan, we associate the closest point on the second scan as its match (correspondence step); (b) given these matching points, we update the estimate of the relative 3D motion between the scans (motion step). An overwhelming majority of 3D modeling systems use the ICP and its modern variants [13] to achieve alignment between individual scans.

Recent methods that use depth scanners and of relevance to our work include 3D alignment and super-resolution of time-of-flight depth camera data [4], methods that use both image and depth data for mapping [8] and the real-time approach of KinectFusion [11]. Since all of these methods use the Kinect or similar scanners, they necessarily need a strategy to overcome the effect of noise in the depth images. In [4]

*Corresponding author.

ICVGIP '12, December 16-19, 2012, Mumbai, India

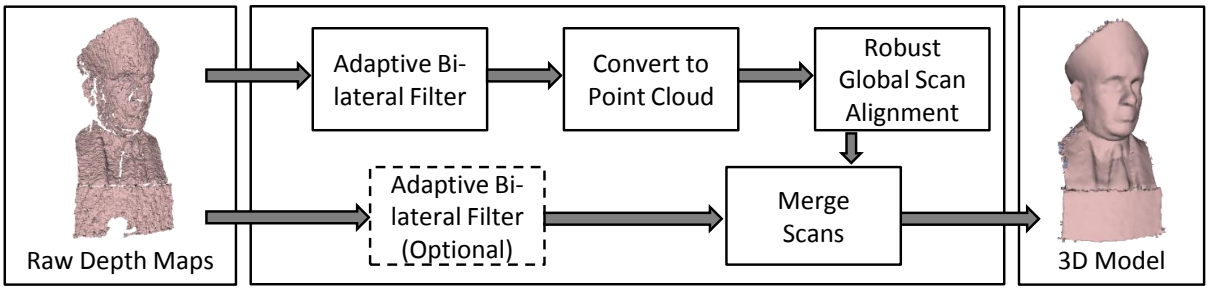


Figure 1: A schematic representation of our computational pipeline that takes many noisy raw Kinect scans as input (left image) and builds a full 3D model out of them (right image).

that uses a time-of-flight scanner, scans that are taken from viewpoints that are proximate to each other are grouped together. Within each such group, the depth images are aligned in a 2D sense by estimating optical flow of each depth image to a central one. In the process the depth images are denoised due to averaging and are super-resolved before being converted into a point cloud for alignment. In [8], the authors obtain an initial guess for ICP using Kinect's image camera. In the state-of-the-art method of KinectFusion [11], the depth images acquired at video frame rate are handled in a simultaneous-localisation-and-mapping (SLAM) framework. As the Kinect is slowly moved in the environment to be sensed, each individual depth image is merged to the global 3D model, i.e. the 3D scene representation is built on the fly. While the alignment of each scan with the global model built up to that point in time is carried out using ICP, [11] also develops a complicated method to accurately merge individual scans one-at-a-time to the global surface representation. The results of [11] are excellent but being based on using a large number of scans acquired at video frame rate, their method is both sophisticated as well as demanding on computational and memory resources. We also note that while [4] is a batch process, the methods of [8, 11] build the 3D models on the fly.

2. OUR APPROACH

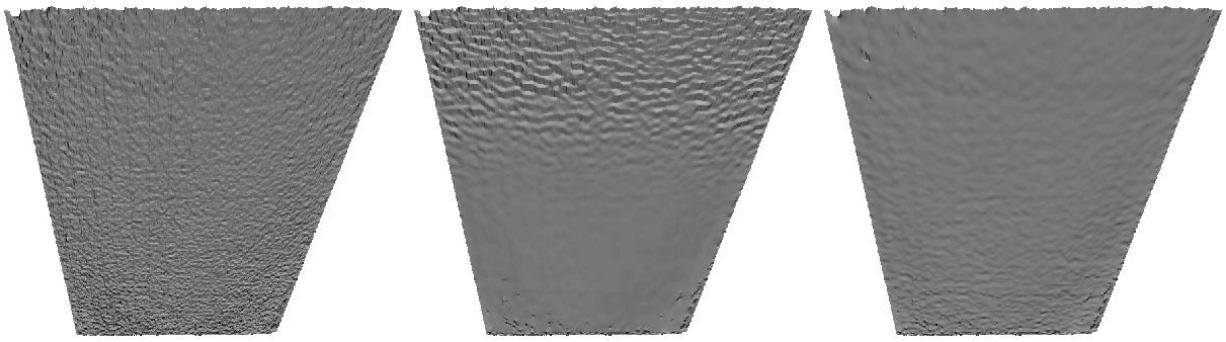
In all of the methods mentioned above, the computational load is significant since the full 3D model is built out of scans acquired at video frame rate. The large number of scans are required for two significant reasons that affect the performance of these systems. Firstly, since the depth images are noisy, a large number of scans are used to average out the effect of noise. Secondly, most methods require that adjacent scans be taken from viewpoints close to each other, as otherwise ICP would fail to converge. The requirement of close viewpoints means that the scanner has to move slowly in the scene, resulting in a large number of scans that need to be acquired and processed. In our work, we aim to build a full 3D surface representations at a lower computational load and with simple methods for merging aligned scans. As a result we adopt a batch approach that works with far fewer scans than in [11], and we neither require that the depth camera be moved slowly nor use all the depth data available at video frame rate. It is sufficient to use a few scans acquired from diverse viewpoints that adequately cover the surface of interest.

While we use a few scans to build our model, we too need to address both the issues of depth image noise as well as the convergence of scan alignment. While smoothing the depth images can reduce the effect of noise, it will also result in distorting the equivalent surface representation, especially across depth discontinuities. In our approach we address these issues by separating out the problem of scan alignment from that of merging the aligned scans. A schematic representation of our computational pipeline is shown in Fig. 1. We use adaptively smoothed depth images to estimate the global scan alignment. Subsequently, we use these estimates to align the scans and merge them. When we have many overlapping scans, we need not apply any smoothing to the raw scans since the merging procedure effectively averages out the noise. When we do not have enough overlapping scan data, we apply a moderate amount of smoothing to the noisy raw data. In other words, we use an adaptively smoothed depth representation to solve for 3D alignment. Having obtained an accurate estimate of the alignment, we use it to merge either the raw scans or smoothed versions as we may desire.

In Sec. 3 we use a sensitivity analysis of the Kinect depth images to develop an adaptive bilateral filter to suppress depth image noise. We demonstrate the superiority of our approach compared to the standard bilateral filter. Since our scans are acquired from viewpoints that are far apart from each other, we require an ICP-based method that can successfully converge with such data. In Sec. 4 we detail our use of a global alignment method that simultaneously solves for alignment of all the scans in a batch. The use of global, simultaneous alignment results in a far greater range of convergence than pairwise ICP. This approach also naturally takes into accounts available constraints such as loop closure. Moreover, we use a robust approach to global ICP-based alignment that can effectively handle errors that may occur in the case of pairwise scan alignment using ICP. In Sec. 5 we describe our approach to merging the scans to build global 3D representations of the scene. Finally, in Sec. 6, we present results of 3D models of different physical scales built using our approach.

3. SMOOTHING OF DEPTH IMAGES

In this section we present our modification of the standard bilateral filter and demonstrate its effectiveness in denoising the depth images acquired by Kinect. However, before proceeding further we briefly discuss the geometric relationship



(a) Raw Scan (b) Bilateral Filtered (c) Adaptive Bilateral Filtered

Figure 2: (a) shows the noisy raw scan of a planar surface (corridor floor) (b) shows the result of applying the standard bilateral filter to this scan. Notice that while the lower region that is closer to the scanner is smoothed, the upper third that is far away is not adequately smoothed. (c) This problem is mitigated by the use of our adaptive bilateral filter. In this figure, the plane has been rotated for ease of viewing so that the distant part of the plane is closer to us and at the top.

between depth images and their 3D counterparts, i.e. 3D scans. Scans are representations in the form of point clouds or meshes where each point (equivalently vertex in a mesh) is denoted by its 3D co-ordinates $\mathbf{P} = (X, Y, Z)^T$. We can establish the relationships of scans with depth images using standard projective geometry. If we consider a camera with calibration $\mathbf{K}$ (where $\mathbf{K}$ is a 3×3 upper-triangular matrix with 5 degrees of freedom), then the 2D image location $\mathbf{p}$ of the projection of 3D point $\mathbf{P}$ onto the camera plane is given as

$$\mathbf{p} = \mathbf{KP} \Rightarrow \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \lambda \mathbf{K} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{1}$$

where (x, y) specifies the location of the projection of $\mathbf{P}$ on to the image plane and λ is a unknown projective scale factor. Depth images are view-centric 2D representations of a surface where the intensity assigned to the pixel at $\mathbf{p}$ is the depth of the 3D point $\mathbf{P}$, i.e. $Z(x, y) = \mathbf{P}(3)$, for (x, y) given by Eqn. 1. Since $\mathbf{K}$ is invertible, we can use Eqn. 1 to convert a 3D scan into a depth image and vice-versa. In the remainder of this paper, we shall use the terms depth image and scan interchangeably. Here it remains to specify our method of estimating the calibration matrix $\mathbf{K}$ of the infra-red camera in Kinect.

We carry out the calibration of the infra-red camera using the standard procedure based on estimating homographies by observing a checker-board pattern [1]. To observe the checker-board under an infra-red source of lighting, we utilise the infra-red projector on the Kinect. In its normal mode of operation, the Kinect's infra-red projector emits a fixed pseudo-random pattern of dots. By placing a translucent sheet of paper in front of the projector, we convert this fixed pattern of dots into a diffuse source of lighting and then directly acquire the unprocessed infra-red camera images. In our experiments, we estimated the infra-red camera's focal length as $f = 587$ pixels which translates to a field of view of about 57°. We also note from the manufacturer's specification that the baseline distance between the centers of projection of the infra-red projector and camera is $B = 75mm$. Finally, we observed that the principal point is very close to the camera plane center and that the non-linear distortion of the infra-red camera is negligible.

Adaptive Bilateral Filtering

Since the raw depth images are highly noisy, converting them into 3D scan representations is unsatisfactory as ICP would fail to work with such noisy scans. Moreover, 3D scans are harder to smooth. Therefore, as in [11], we choose to smooth the depth image representation while taking care to preserve details, especially edge discontinuities. This objective is usually achieved by means of a bilateral filter that in addition to a Gaussian spatial smoothing kernel also includes a range weighting term that explicitly accounts for the intensity difference between the central pixel and other pixels in the support of the smoothing kernel [15]. Thus, for a noisy depth image $Z(\mathbf{p})$ where $\mathbf{p} = (x, y)$ is the pixel location, the denoised output of the bilateral filter is given by

$$\sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} w_s(\mathbf{q} - \mathbf{p}) w_d(Z(\mathbf{q}) - Z(\mathbf{p})) Z(\mathbf{q}) \tag{2}$$

where w_s and w_d are normalised Gaussian functions for spatial and range weighting with standard deviations of σ_s and σ_d respectively and $\mathcal{N}(\mathbf{p})$ is the neighbourhood of $\mathbf{p}$. In other words,

$$w_s(\mathbf{x}) \propto e^{-\frac{\mathbf{x}^T\mathbf{x}}{2{\sigma_s}^2}} \tag{3}$$

$$w_d(y) \propto e^{-\frac{y^2}{2{\sigma_d}^2}} \tag{4}$$

and the weighting masks are normalised to have a total sum of 1 over $\mathcal{N}(\mathbf{p})$. It will be noted that compared to the standard Gaussian smoothing filter, the bilateral filter explicitly accounts for the intensity difference between the central pixel $\mathbf{p}$ and its neighbour $\mathbf{q}$ thereby suppressing the influence of pixels that are spatially close to $\mathbf{p}$ but are different in intensity. The result is a smoothing filter that also preserves edges in an image. In Fig. 2(a) we show the 3D representation of a depth image of a planar surface (corridor floor) and in Fig. 2(b), we show the result of applying

a bilateral filter to it. Although bilateral filtering smooths an image while preserving discontinuities, we note that unlike intensity images, depth images obtained from a scanner have specific properties that can lead to better denoising than that of the conventional bilateral filter.

Depth scanners like the Kinect are based on using structured-light stereo to estimate stereo disparity (D) that is inversely related to depth (Z), i.e. $D = \frac{fB}{Z}$ where B is the baseline distance between the projector and camera centers and f is the focal length of the camera. By taking derivative of the disparity equation with respect to depth we have

$$\frac{\partial D}{\partial Z} = -\frac{fB}{Z^2} \quad \Rightarrow \quad \frac{\partial Z}{\partial D} = -\frac{Z^2}{fB} \tag{5}$$

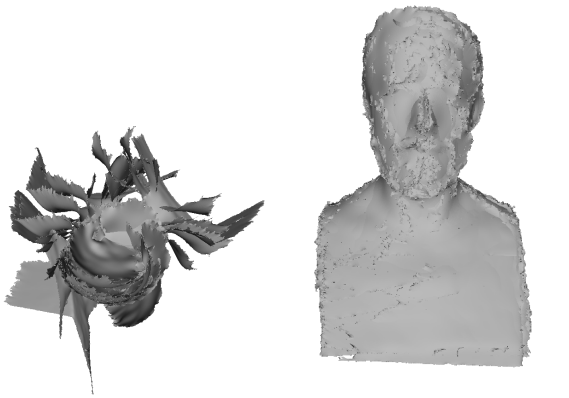
This implies that the sensitivity (equivalently precision) of estimating depth using a stereo projector-camera pair falls off according to an inverse square law. Now, if we consider two depth values of 2 and 5 feet ($600mm$ and $1500mm$ respectively) we have the sensitivity values of

$$\left.\frac{\partial Z}{\partial D}\right|_{Z=600mm} = -\frac{600^2}{75 \times 587} = -8.2\, mm/pixel$$
$$\left.\frac{\partial Z}{\partial D}\right|_{Z=1500mm} = -\frac{1500^2}{75 \times 587} = -51.1\, mm/pixel$$

i.e., an error of 1 pixel in disparity estimation translates to a depth error of $8.2mm$ and $51.1mm$ at 2 and 5 feet respectively. Apart from the variation in sensitivity, we may also note that since the projector bundle of rays are divergent, the surface sampling density is lower for objects that are far from the scanner. In combination, these two factors suggest that the variance of depth estimates (i.e. intensity of depth image pixels) of far away objects is more than surfaces that are close to the scanner. We utilise this characteristic of the Kinect (or other scanners) depth images to suitably modify the bilateral filter.

In the standard bilateral filter, the standard deviation of the spatial smoothing mask is given as σ_s and that in the intensity range by σ_d. Since our analysis shows that the depth estimate sensitivity is dependent on the depth itself, we modify σ_d to vary as Z^2, i.e. $\sigma_d = kZ^2$, where k is a constant. As a result, our modified *adaptive bilateral filter* is the same as that of Eqn. 2 with the modification that instead of σ_d being a constant, for each pixel $\mathbf{p}$, we have $\sigma_d = kZ^2(\mathbf{p})$. In Fig. 2(c) we show the result of applying our adaptive bilateral filter to smoothen the raw data obtained from scanning the floor of a building corridor in Fig. 2(a). While for a given σ_d the standard bilateral filter in Fig. 2(b) works well for points on the floor that are close to the scanner, we can see that for points that are further away (i.e. the upper part of the scan that has been rotated for viewing) the standard bilateral filter's smoothing is inadequate. In contrast, as can be seen in Fig. 2(c) our adaptive method gives superior results for all depths since the smoothing in the range kernel takes into account the specific manner in which the depth image is generated.

4. GLOBAL MULTIVIEW SCAN ALIGNMENT



(a) Alignment Failure (b) Successful Alignment

Figure 3: Importance of robust averaging : (a) shows a case where the averaging method of [12] fails due to outliers; (b) shows that our robust modification succeeds here.

Most scanning systems use a variant of the basic ICP method to carry out the registration of individual scans and build a global 3D model. In [8], ICP is used with an initialisation obtained from the image camera data. In [11], each new scan is registered using ICP to the global model built upto that point using all previous scans. Subsequently the global model is updated by merging the new scan with it. However, this approach of aligning one scan to another (i.e. pairwise) has some limitations, especially with noisy scans. In particular, pairwise ICP fails to converge without a good initialisation, forcing many systems to move the scanner slowly through the scene. The result is a large number of scans that make a significant demand on computational capacities. In contrast with such methods, we use far fewer scans taken from widely different viewpoints so as to reduce the overall computational load. However, pairwise ICP would fail to converge for such data that needs a large motion to be registered. We overcome this limitation of the conventional ICP method by adapting the results of [12] that present an extended approach that accounts for the global motion between *all* scans at the same time. As described below, the method of [12] increases the convergence range when compared to standard ICP while it also improves the accuracy of alignment as it simultaneously takes all available pairwise alignment constraints into account. The method of [12] may fail to work with noisy Kinect data which might create erroneous pairwise scan relationships. In our work, we have modified this technique by incorporating a robust estimation stage to make the method of [12] work in the presence of outliers. In the remainder of this section, we will briefly describe the global scan alignment method of [12] as well as explain our modification to make it robust.

Motion Averaged ICP

In [12], the authors present a multiview extension of the ICP algorithm that simultaneously aligns all scans in a common frame of reference. For a set of N scans, we can define an N-vertex viewgraph G, where each vertex denotes a scan. Further if two scans i and j have a sufficient amount of over-

lap to allow alignment, i.e. we can estimate $\mathbf{M}_{ij} \in SE(3)$ which is the relative motion between i and j, we add an edge between vertices i and j in G. Let the set of all such edges be denoted E. If the set of relative motions E contains a spanning tree S, we can always construct an alignment between the individual scans and place all of them in a common global frame of reference, i.e. solve the global alignment problem. Further, all the remaining edges $\{E\} \setminus \{S\}$ provide additional constraints on the motions between the individual scans. The method of [12] uses all the $|E|$ relative motions constraints to solve for a global motion model. In the process, it averages all the relative motions in E thereby improving the overall accuracy of the alignment estimate. In particular, loop closure constraints that are often available in scan data are naturally incorporated and utilised in the averaging method of [12]. Without loss of generality, we fix the frame of reference to the first scan. Therefore the global motion estimate required to align all scans is given as $\mathbf{M}_{global} = \{\mathbf{I}, \mathbf{M}_1, \cdots, \mathbf{M}_N\}$. Starting with an initial guess of $\mathbf{M}_{global}$, the 'motion averaged ICP' method of [12] can be briefly summarised as follows. During each iteration we carry out three steps

(a) for every scan pair $(i, j) \in E$ we carry out the ICP correspondence step between scans i and j

(b) subsequently, for each edge $(i, j) \in E$, the motion model $\mathbf{M}_{ij}$ is re-estimated using the correspondences established, following which

(c) given all the $|E|$ relative motions between pairs of scans $\{\mathbf{M}_{ij} | \forall (i, j) \in E\}$ a motion averaging step is carried out to create a global motion model.

These correspondence, motion and averaging steps constitute one iteration and are repeated till convergence of the solution. Here, given many relative motion estimates $\mathbf{M}_{ij}$, the motion averaging step solves for $\mathbf{M}_{global}$ by minimising a cost function :

$$\sum_{(i,j) \in E} d^2(\mathbf{M}_{ij}, \mathbf{M}_j {\mathbf{M}_i}^{-1}) \tag{5}$$

where $d(.,.)$ is the intrinsic Riemannian distance on $SE(3)$.

In general, this global alignment method works better than aligning a pair of scans at a time as it takes all available constraints into account. In the process, the errors in individual motion estimates for each edge in E are averaged out, resulting in a more accurate global solution. In particular, if E contains a cycle, such a 'loop closure' provides very strong constraints on the global motions and greatly reduces the drift of the solution that may be observed when scans are individually aligned either with other scans or with a global motion model built on the fly. This is particularly the case when the motions between the scans are large, i.e. scans are acquired from widely differing viewpoints.

Although the global alignment method of [12] improves on the standard ICP routine, it assumes that none of the relative motion estimates pertaining to scan pairs in E are outliers, i.e. no pairwise alignment is grossly wrong. In other words, the method of [12] fits a global motion model to the relative motion estimates by minimising their discrepancy in a least squares sense which makes it susceptible to outliers. Despite applying the above-mentioned adaptive bilateral smoothing to depth images, there are instances when the data acquired using Kinect generates outliers in the set of relative motions $\{\mathbf{M}_{ij} | \forall (i, j) \in E\}$. Such outliers can arise when (a) the correspondence step for a pair of scans fails due to noise or locks on to a false set of correspondences, (b) when the region of overlap between the two scans is small or has very few discriminating 3D features, or (c) when the dominant part of the depth image is devoid of unambiguous features, say a flat wall of a room. We parenthetically note here that these problems are mitigated to an extent by using the point-to-plane distance instead of the point-to-point distance in the ICP correspondence step. However, due to a combination of these factors, we may have one or more pairs (i, j) that result in erroneous motion estimates in which case the 'motion averaged ICP' of [12] fails as it uniformly distributes the overall error which can be large in the presence of outliers. To overcome this problem, in the 'averaging step' described above, we use a modified robust motion averaging step that is based on the proposal in [7] and works in the presence of outliers. In Fig. 3(a) we show the case where global alignment using the method of [12] fails whereas our modified robust motion averaged ICP works despite the presence of outliers, as can be seen in Fig. 3(b). Indeed, in our experiments we have found that this robust modification is crucial to make the overall global alignment method work for many scenarios since Kinect data is not of high quality.

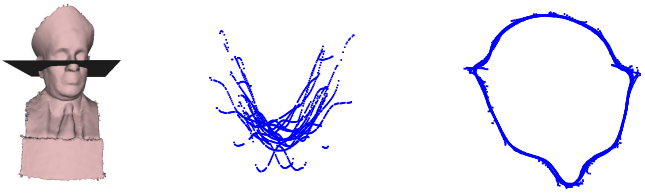
5. MERGING OF REGISTERED SCANS

After alignment, due to motion estimation errors, the presence of noise and other sources of inaccuracy, the overlapping regions of the scan surfaces will not be perfectly aligned. Therefore, we need to estimate a single 3D surface representation that best fits the individual, aligned scans. To achieve this objective we use the volumetric merging method of [5]. This method begins by laying a volumetric grid to cover all the individual aligned 3D scans. At each grid point, the signed distance of individual 3D scans along the respective viewing direction is computed. A weighted average of these signed distances is assigned to the grid points, resulting in a discrete scalar field that encompasses the volume containing the scans. The scalar field gives us an estimate of the average distance of a grid point from all the scans. By using a marching cubes algorithm, an iso-surface is extracted for this scalar field for the iso-value or distance of zero. This iso-surface can be shown to be the maximum likelihood estimate of the actual surface. In our system, we use the publicly available implementation of this method in the Meshlab package [3]. Finally, if required, we further smooth the final merged 3D surface representation by applying non-local smoothing [16]. In the case of large scale 3D models, we use the Laplacian mesh smoothing method for efficiency.

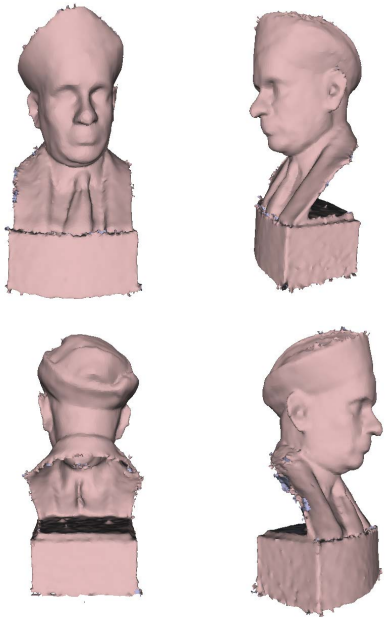
6. RESULTS

In this section we present some reconstructed 3D models that demonstrate the performance of our system. We note here that the 3D models have different shape characteristics and also illustrate the capacity of our system to operate

(a) Final Model with Estimated Scanner Locations



(b) Cross-sections before and after alignment



(c) Different views of our 3D model

Figure 4: (a) shows our final 3D model along with the estimated positions and orientations of the individual Kinect scans; (b) shows the same 3D model rendered from different viewpoints.

at different physical scales of the scene being reconstructed. Throughout, for the adaptive bilateral filtering used prior to alignment we set $k = 0.045/m$, i.e. $\sigma_d = 0.045Z^2$ for Z in meters. Thus, for $Z = 1m$, σ_d is $0.045m$ or $45mm$.[1] Fig. 4 shows different aspects of our reconstruction of a cast-metal bust that has a height of about 50cm. As illustrated in Fig. 1, although the input raw scans obtained by Kinect are noisy, our system is able to align the scans accurately lead to a 3D reconstruction of good quality. Fig. 4(a) shows the reconstructed model as well as the estimated locations and orientations of the 21 raw scans used as input. It will be noted that the location of the scanners demonstrates the ability of our system to successfully carry out alignment even when the individual scans are taken from viewpoints that are far apart. Fig. 4(b) indicates a cross-section plane on the model and shows the corresponding cross-section points on individual scans before and after alignment. As can be seen, our system is able to accurately align the individual scans. In Fig. 4(c) this accuracy is further illustrated by rendering the final reconstructed model from different viewpoints. Since we have significant overlap between the scans in this case, the merging procedure of Sec. 5 is applied to the raw scans, thereby ensuring that small-scale features are preserved.

In our second experiment we build a 3D model of an entire room that has a floor size of $2.6m \times 3.2m$ by using 30 raw scans that cover the four walls of the room. Fig. 5(a) and (b) show a novel plan view of our reconstruction of the room with and without texture respectively. It will be noted from the fidelity of the model that our method is able to accurately align the individual scans to construct this large scale model. In particular, we draw the reader's attention to the accurate recovery of the floor plan. Also notice, for instance, the orthogonality between the wall and the door in the upper-right corner of the plan view. We wish to emphasise here that although the room is *two orders* of magnitude larger than the previous example of the bust, we are able to reconstruct it with very few scans, i.e. unlike tracking methods such as [11], we do not need a large number of scans for successful alignment. Although the individual scans have minimal overlap between them, by using our robust, multiview alignment method of Sec. 4 we are able to solve the alignment problem. Our pipeline here is the same as in the previous example, except that we smooth the scans before using them in the merging algorithm as, unlike those in Fig. 4, the individual scans do not have significant overlap except at the boundaries. Therefore many parts of the final

[1]Note that σ_d applies to depth measured as intensity range in the depth image.

(a) Plan view with texture

(b) Plan view without texture

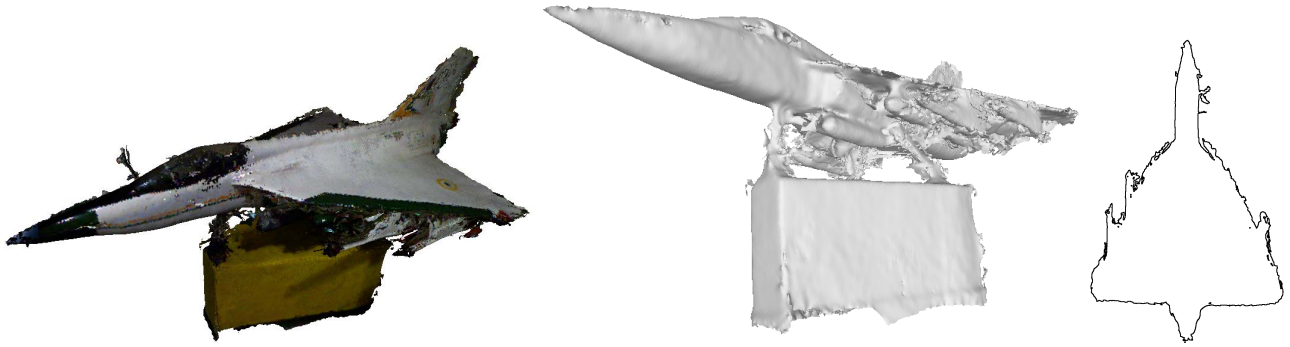
(c) Novel textured views from different viewpoints

Figure 5: Reconstructed 3D model of a room. Note the accuracy of the plan view as well as the textured novel views.

model have 3D information from only a single scan and need smoothing to reduce the effect of sensor noise. In Fig. 5(c) we further illustrate the accuracy of our model by showing four different novel views of our final, texture-mapped reconstruction.[2]

As the final experiment in this paper, in Fig. 6 we present our reconstruction of a $3.2m$ long scale model of a combat aircraft. Our reconstruction is built out of 37 scans and is shown from different viewpoints with and without texture in Fig. 6(a) and (b) respectively. It will be noted that the individual scans are obtained by walking around the model and all the partial scans are accurately aligned to result in a correct reconstruction. In Fig. 6(c) the accuracy of our full reconstruction is further illustrated by a plan outline view of our reconstruction of the aircraft. It will be noted that all 37 scans need to align correctly for this 'silhouette' to be correct. Reconstructing this model is a significant challenge that showcases the different strengths of our system. In particular, solving the global alignment is a difficult challenge for two reasons. Firstly, the underbelly manifold of the aircraft is very complex and can cause alignment errors due to repeated structures that can induce gross errors in the correspondence step. In such a scenario, pairwise ICP failures cannot be addressed. In contrast, since our approach treats all pairwise scan relationships in a multiview framework, the availability of all multiview relationships allows our robust method to naturally detect erroneous relationships and remove their influence on the overall solution. Secondly, since the nose of the aircraft is very narrow, it results in very little overlap between adjacent scans. Similarly, the flatness of the wings can create ICP errors as they lack features. This is also the case with the axial symmetry of the model which

[2]The textureless, white patches are holes in our model that arise when parts of the scene are occluded from individual scanning viewpoints. It may be noted that for scenes of such 3D complexity, occlusions are common especially when we use very few scans.



(a) Textured Model (b) Without Texture (c) Plan View of Estimate

Figure 6: Reconstructed 3D model of a complex aeroplane model is shown with and without texture mapping.

can induce completely erroneous alignments that are wrong but locally appear to be correct. For instance, one part of a wing can be aligned with the other wing by means of a large rotation. In the presence of these ambiguities, we estimate an initialisation for the robust multiview alignment method by utilizing the adjacency of the scans in the order in which they were taken. Using this initialisation, we successfully solve the full alignment problem for this challenging dataset.

7. CONCLUSION

In this paper we have presented a pipeline that can reliably combine noisy Kinect scans into geometrically consistent 3D models. We have introduced and demonstrated the effectiveness of our adaptive bilateral filter that mitigates the effect of noise by explicitly accounting for the sensitivity of depth estimation in structured-light scanners. A robust multiview scan alignment method provides accurate estimates which in turn leads to successful building of complex 3D object and scene models. The capabilities of our computational pipeline is demonstrated through examples of 3D reconstructions of objects and scenes of different shapes and physical scales.

8. REFERENCES

[1] J. Y. Bouguet. Camera calibration toolbox for Matlab, 2008.

[2] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image Vision Comput.*, 10(3):145–155, Apr. 1992.

[3] P. Cignoni, M. Corsini, and G. Ranzuglia. Meshlab: an open-source 3d mesh processing system. *ERCIM News*, pages 45–46, April 2008.

[4] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *CVPR'10*, pages 1173–1180, 2010.

[5] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '96, pages 303–312, New York, NY, USA, 1996. ACM.

[6] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[7] R. I. Hartley, K. Aftab, and J. Trumpf. L1 rotation averaging using the weiszfeld algorithm. In *CVPR*, pages 3041–3048, 2011.

[8] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgbd mapping: Using depth cameras for dense 3d modeling of indoor environments. In *In RGB-D: Advanced Reasoning with Depth Cameras Workshop in conjunction with RSS*, 2010.

[9] M. A. Lourakis and A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[11] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 7(10):127–136, 2011.

[12] A. Pooja and V. M. Govindu. A multi-view extension of the icp algorithm. In *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '10, pages 235–242, 2010.

[13] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3DIM01*, pages 145–152, 2001.

[14] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.

[15] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 839–, Washington, DC, USA, 1998. IEEE Computer Society.

[16] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. Smoothing by example: Mesh denoising by averaging with similarity based weights. In *In Proceedings of the IEEE International Conference on Shape Modeling and Applications (2006)*, pages 38–44. IEEE, 2006.