3D Self-Portraits

Hao Li[1] Etienne Vouga[2] Anton Gudym[5] Linjie Luo[3] Jonathan T. Barron[4] Gleb Gusev[5]
[1]University of Southern California [2] Columbia University [3]Adobe Research [4]UC Berkeley [5]Artec Group

Figure 1: *With our system, users can scan themselves with a single 3D sensor by rotating the same pose for a few different views (typically eight, ∼45 degrees apart) to cover the full body. Our method robustly registers and merges different scans into a watertight surface with consistent texture in spite of shape changes during repositioning, and lighting differences between the scans. These surfaces are suitable for applications such as online avatars or 3D printing (the miniature shown here was printed using a ZPrinter 650.)*

Abstract

We develop an automatic pipeline that allows ordinary users to capture complete and fully textured 3D models of themselves in minutes, using only a single Kinect sensor, in the uncontrolled lighting environment of their own home. Our method requires neither a turntable nor a second operator, and is robust to the small deformations and changes of pose that inevitably arise during scanning. After the users rotate themselves with the same pose for a few scans from different views, our system stitches together the captured scans using multi-view non-rigid registration, and produces watertight final models. To ensure consistent texturing, we recover the underlying albedo from each scanned texture and generate seamless global textures using Poisson blending. Despite the minimal requirements we place on the hardware and users, our method is suitable for full body capture of challenging scenes that cannot be handled well using previous methods, such as those involving loose clothing, complex poses, and props.

CR Categories: I.3.3 [Computer Graphics]: Three-Dimensional Graphics and Realism—Digitizing and scanning

Keywords: 3D scanning, non-rigid registration, depth-sensor, human body, texture reconstruction

Links: DL PDF WEB VIDEO

1 Introduction

For many years, acquiring 3D models of real-world objects was a complex task relegated to experts using sophisticated equipment such as laser scanners, carefully calibrated stereo setups, or large arrays of lights and cameras. The recent rise of cheap, consumer-level 3D sensors, such as Microsoft's Kinect, is rapidly *democratizing* the process of 3D scanning: as these sensors become smaller, cheaper, more accurate and robust, they will continue to permeate the consumer market. Within a decade, 3D capability will likely become as standard built-in feature on laptops and home computers as ordinary video cameras are today.

Recent work on software systems for geometry processing have leaped forward to adapt to the revolution in 3D acquisition hardware. Using methods like Kinect Fusion [Newcombe et al. 2011], ordinary users with no domain knowledge can now generate scans of everyday objects with stunning detail and accuracy. However, with the users behind the 3D sensor, it is difficult to use these methods to capture the *3D self-portraits* of the users *on their own* analogous to photographic self-portraits. In this paper, we concern ourselves with the development of a flexible, robust and accurate capture system for 3D self-portraits using a single 3D sensor.

Figure 2: *3D printed miniatures of captured surfaces.*

There are many potential applications for such 3D self-portraits: combined with some algorithms for automatic skinning, these portraits could be used as personalized, *virtual avatars* in video games or video conferencing applications. Users could quickly scan and upload complete 3D portraits of themselves showing off new outfits and styles to social media sites, or create physical action figures of themselves by having the models 3D printed. Since a 3D portrait fully captures a user's measurements, it could be used to accurately preview the fit and drape of clothing ("virtual try-on") when shopping online. By scanning themselves regularly over a period of time, users could both visually and quantitatively track improvements in their health and fitness.

With a single 3D sensor and no other operators helping to move the sensor, users have to rotate themselves to scan all parts of their bodies. This naturally raises two problems. First, incidental changes of

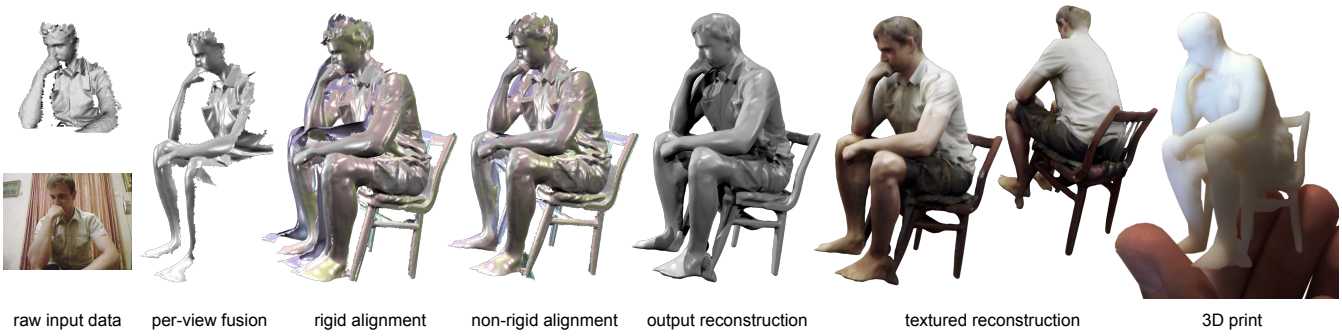


Figure 3: *Our reconstruction pipeline takes as input around 150 frames of raw depth maps and textures for each of eight captured views (Sec. 3.1). It fuses and segments these frames to yield per-view fused surfaces (Sec. 3.2). These per-view surfaces are first registered using rigid alignment (Sec. 3.3) and then refined with non-rigid alignment in a global optimization (Sec. 3.4). The aligned surfaces are merged into a final watertight surface (Sec. 3.5) with consistent texture (Sec. 3.6) which is suitable for 3D printing.*

users' poses can happen between the scans even with the best effort to keep the same pose; and second, non-uniform and uncontrolled lighting in the homes of typical users can cause significant appearance changes during rotation as captured by the sensor.

To address the first problem of pose change, we propose a multi-view non-rigid registration based on work by Li et al [2009] to robustly align all the scans and merge them into a globally consistent surface using Poisson surface reconstruction [Kazhdan et al. 2006]. To ensure that the final reconstructed models have consistent illuminated texture, we leverage the ideas of Barron et al [2012; 2013] and recover the underlying albedo for each scan. Poisson texture blending [Chuang et al. 2009] then adds smooth transitions between these albedo maps across scans.

Our system is the first autonomous capture system for 3D self-portraits with a lightweight acquisition setup (only one 3D sensor, in contrast to multi-view stereo methods [Seitz et al. 2006]) and great flexibility in the kinds of user poses and personalized garments and accessories that can be captured (see Figure 1 for examples). Although many exisiting methods based on human body shape or structure priors [Anguelov et al. 2005; Hasler et al. 2009; Weiss et al. 2011; Cui et al. 2012] work well for subjects standing in controlled poses and wearing unobtrusive clothing, they are not designed to handle more interesting poses with all varieties of personalized garments where these assumptions break down. Similarly, skeleton tracking [Shotton et al. 2011; Wei et al. 2012] is difficult for poses involving significant limb occlusions in partial body scans. Our system takes a *purely geometric* shape-based approach to body capture that makes minimal prior assumptions about the subject's shape and pose to ensure maximal flexibility.

User Experience. The user begins the scanning process by pressing the "start capture" button on her computer, then prepares the desired pose about one meter away from the depth sensor. A sound signals the beginning of the acquisition process and warns the user to stand as still as possible for about four seconds until the sensor has made a full scan of the subject from the current view. A second sound notifies the user that capture of the current view is complete, and the user is given five seconds to turn roughly 45 degrees clockwise and roughly reproduce the original pose. This process repeats about eight times, until a 360-degree capture of the subject is complete. This online portion of our algorithm, illustrated in Figure 1, takes about two minutes total. It is followed by about 12 minutes of offline processing requiring no user intervention. The details of both portions are described in the next section.

2 Related Work

We review the relevant recent advances in 3D shape reconstruction from captured depth and point cloud data.

Static Objects. To obtain the complete geometry of a static object, *rigid* reconstruction techniques assemble multiple partial 3D scans by matching their overlapping regions. When pairs of scans are very close to each other, alignment techniques based on the iterative closest point algorithm (ICP) [Rusinkiewicz and Levoy 2001] are generally preferred due to their efficiency and reliability. Real-time 3D reconstruction methods from continuous streams of depth maps [Rusinkiewicz et al. 2002; Newcombe et al. 2011] are all based on ICP since consecutive scans are temporally coherent. However, even when the registration between pairs of scans converges, it is likely that tiny errors are introduced due to noise and incompleteness of the acquisition data. The accumulation of errors leads to the well-known "loop closure" problem which has been addressed by the multi-view registration framework of Pulli [1999]. A method for handling the loop closure problem in a real-time 3D scanning setting has been developed by Weise and coworkers [2011]. While both approaches aim at diffusing the registration error across all recorded 3D scans, they only align scans of static objects. Using rigid registration techniques for human capture requires that the subject stays perfectly still; this requirement is only reasonable with the assistance of a turn-table or a second operator (to sweep the sensor around the subject while the she stands still.)

Deformable Objects. For 3D digitization at home, we wish to reconstruct an entire body by turning around a single static depth sensor. To align scans of deformable subjects, pairwise non-rigid registration techniques have been introduced to handle different types of deformations such as quasi-articulated motions [Chang and Zwicker 2008; Chang and Zwicker 2009], isometric deformations [Huang et al. 2008], and smooth rigidity-maximizing transformations [Li et al. 2008]. As in the rigid case, these techniques have been extended to reconstruct deformable objects from a stream of multiple input scans. Due to the potential complexity of the deformations, these methods generally require good coverage, e.g. multiple sensors surrounding the subject [Sharf et al. 2008; Vlasic et al. 2009; Li et al. 2012; Tong et al. 2012]. While [Tong et al. 2012] requires 3 depth sensors and a turntable, our pipeline only needs one static sensor and can handle arbitrary poses and props, since we do not involve human shape priors. Moreover, our non-rigid alignment is purely geometric and thus reliable for textureless subjects while [Tong et al. 2012] relies on texture features for matching.

The approach of Li and colleagues [2009] uses a crude approximation of the scanned object as a shape prior to obtain a more detailed reconstruction. Several dynamic shape reconstruction techniques [Mitra et al. 2007; Wand et al. 2009] do not require any shape prior, but assume the motion of the subject to be very small and smooth. Two techniques [Tevs et al. 2012; Bojsen-Hansen et al. 2011] were recently introduced to handle topology changes in the input data, but cannot avoid drift when longer sequences are involved. Brown and Rusinkiewicz [2007] presented a global multi-view registration technique to unwarp small-scale deformations introduced by calibration errors across multiple scans. The global alignment method of Chang and Zwicker [2011] can cope with larger deformations, but is designed to handle quasi-articulated motions and is less suitable for aligning garments. Based on a similar optimization framework, Cui and colleagues [2012] developed a pipeline to reconstruct a full human body from a single Kinect sensor. While the results are promising, their method is limited to subjects that perform a T-pose and wear relatively tight clothing, so that the articulation assumption remains valid.

3 Reconstruction Pipeline

Approach and Terminology Our method begins by scanning the user from a few views (usually around 8) so as to cover the full body (Sec. 3.1). The raw *frames* captured by the sensor are fused and segmented to reduce acquisition noise and separate the subject from the background (Sec: 3.2), resulting in one *view scan* per view. Since the user needs to rotate and repose for each different view, deformations and inconsistencies are inevitable. We develop a systematic approach to robustly align the separate view scans. We first perform rigid alignment between the scans of adjacent views (Sec. 3.3). We then refine the alignments with non-rigid optimization and find the dense correspondences between the adjacent scans. The correspondences are used to warp the scans in a global optimization (Sec. 3.4) and the aligned scans are merged into a watertight surface using the visual hull as the shape prior (Sec. 3.5). Finally we reconstruct a globally consistent texture for the merged surface that transitions smoothly across different views in spite of any illumination differences (Sec. 3.6). Figure 3 illustrates the steps of our system.

3.1 Scanning

The first step of our pipeline is to scan the user and collect geometry and texture data. As described in the introduction, the user maintains the desired pose in front of the 3D sensor for a few seconds while the sensor collects depth and texture frames. Then the user rotates about 45 degrees and reproduces the same pose for another scan from a new view. Our system scans the user about 8 to 9 times (a full spin) for full-body reconstruction of the desired pose. In total, the scanning takes about two minutes to finish. The resulting partial scans may contain deformations and inconsistencies, which we address robustly in Sec. 3.3 and Sec. 3.4.

We use the Kinect sensor as our main 3D sensor for scanning since it is affordable and widely available. However, it should be noted that our system is not restricted to any particular scanning technology and can work with all kinds of 3D sensors. The current generation of Kinect hardware requires a tradeoff between depth map quality and field of view: adequately resolving the fine details of a user's clothing and body requires them standing close enough that the entire body cannot be captured at once.

We therefore use the Kinect's *motorized tilt* feature to extend the effective scanning range while asking the user to stand relatively close (ca. one meter) to the sensor to maintain scanning resolution. The scanning for each view begins with the Kinect maximally elevated (27 degrees above horizontal), and during the capture we sweep the Kinect downwards in a 54-degree arc. We expect that

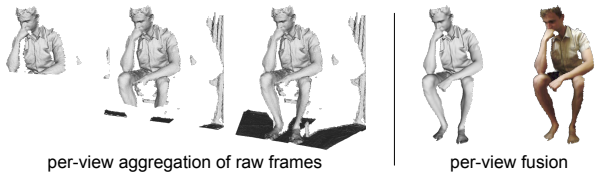


Figure 4: *The captured incomplete raw depth maps* (left). *Fused and segmented per-view surface and texture* (right).

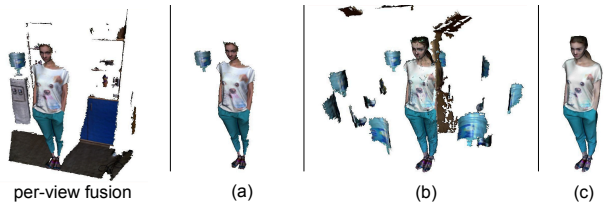


Figure 5: *Geometric debris are filtered for each view (a). After multiple views are aligned some fragments remain (b) and we extract the largest single connected mesh (c).*

the future generations of the 3D sensors with greater resolution and accuracy will obsolete this strategy and enable faster accurate scanning. While it is best to avoid motion during this sweeping process, we find that slight involuntary movements (e.g. breathing) do not affect the quality of the result.

3.2 Fusion and Segmentation

During the scanning for each view, we gather roughly 150 frames of raw depth maps and textures. Since the Kinect is in motion during this capture, registration of the raw depth map is required to aggregate a full per-view scan. For each frame, since we know an approximate volume enclosing the user being scanned, we clip all geometry outside of this volume as a simple initial segmentation of the body from the background, and then register the new frame to those previously captured using rigid ICP (correspondence search radius is 25 mm, maximal iterations number is 15, and all vertices are used for matching). We perform this registration in real time so that we can give the user visual feedback of scan progress.

We use Poisson surface reconstruction [Kazhdan et al. 2006] to merge all the registered depth maps into a single reconstructed view surface $\mathbf{S}$ and obtain the per-view fusion mesh of Figure 5. We then texture and clip $\mathbf{S}$: we compute the median frame's camera position and orientation $\mathbf{c}$, and rasterize all of the raw frames using this camera. We average the resulting image and use it as the texture of $\mathbf{S}$. Moreover, we delete any parts of $\mathbf{S}$ that do not overlap any of the raw frames, so that we do not introduce any spurious data during Poisson reconstruction. By analysing the distribution of normals, we also detect large horizontal planar regions (the floor) and remove them. As shown in Figure 4, this super-resolution reconstruction process yields surfaces $\mathbf{S}$ with substantially greater detail and less noise than the individual frames. We show quantitative validation of this process in Figure 12.

Our initial segmentation removes some of the background, but we still end up with a lot of outlier data in $\mathbf{S}$, such as parts of the floor, walls, and nearby objects. We therefore perform a second round of segmentation to remove this misclassified debris: we know that the user is rotating between each view, while the background is not, so we remove, before the alignment, portions of $\mathbf{S}$ that remain unchanged for more than 3 scans as shown for a single view in Figure 5 (a) and all views in Figure 5 (b). We also delete small disconnected debris from the main watertight mesh by extracting the largest single connected mesh as illustrated in Figure 5 (c).

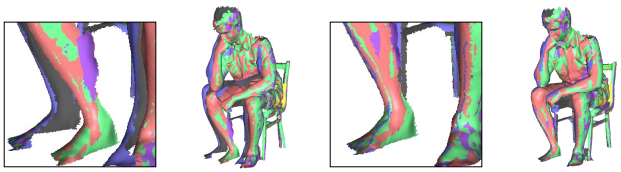


Figure 6: *Illustration of the loop closure problem. Aligning consecutive scans sequentially leads to large accumulated alignment error between the first and last scans near the subject's foot, due to inevitable movement during reposing* (left)*. The problem is solved by optimizing the multi-view non-rigid alignment globally* (right)*.*

3.3 Initial Rigid Alignment

Although we ask the user to rotate about 45 degrees between views, in practice the user turns between 30 and 60 degrees, resulting in $n \sim 8$ views. Before attempting to find correspondences for non-rigid registration, we preprocess the n views by roughly aligning them, first by assuming they differ exactly by a 45-degree rotation, and then refining the alignment using progressive rigid registration. We compute the axis of rotation by extracting the largest eigenvector from the covariance matrix of the point cloud formed by taking the union of the vertices of all the scans; this heuristic works well for scanning people (who are taller than they are wide), and is optional if prior information is known about the setup of the 3D sensor (e.g., that it is mounted parallel to the ground).

Registration is done in a cumulative fashion: we first perform rigid ICP to align $\mathbf{S}_1$ with $\mathbf{S}_0$, then align $\mathbf{S}_2$ with both $\mathbf{S}_0$ and $\mathbf{S}_1$, and so forth. Our rigid ICP algorithm is based on point-to-plane minimization [Rusinkiewicz and Levoy 2001] using closest point constraints. To improve robustness, we then repeat alignment in reverse, aligning $\mathbf{S}_{n-1}$ to $\mathbf{S}_n$, $\mathbf{S}_{n-2}$ to $\mathbf{S}_{n-1}$ and $\mathbf{S}_n$, and so forth. During rigid ICP, we ignore all potential correspondences where the corresponding points' normals differ by more than ten degrees: in this way, non-rigidly-deforming local features, like a moving arm, are ignored by ICP. We observed that this measure was necessary to prevent ICP from converging to poorly aligned local minima. Figure 3c shows a typical example of the views after rigid alignment.

Applying the rigid transformations to the meshes and their cameras gives us transformed, aligned meshes $\mathbf{S}_i^R$ and their transformed camera views $\mathbf{c}_i^R$. These will be the input to the next step in our algorithm, non-rigid registration.

3.4 Multi-View Non-Rigid Registration

Once the views have been rigidly aligned, we fuse them together using non-rigid registration. In our application, this registration is challenging for two reasons: first, large, reliable regions of overlap exist only between consecutive views; and second, the subjects inevitably shift their arms, head, and feet slightly while rotating between each view, even when trying to hold very still, and this error accumulates so that there is a significant non-rigid deformation between the first and last view. We attempted progressive fusion of the views, i.e. registering $\mathbf{S}_1^R$ with $\mathbf{S}_2^R$, registering that result with $\mathbf{S}_3^R$, and so on, but because of accumulation of deformations in the input and errors in the registration, this approach often gave poor results with loop closure errors (see Figure 6). We therefore split non-rigid registration and reconstruction into two steps, along the lines of Chang and Zwicker's approach to global registration of articulated models [2011]: first, we find *pairwise* correspondences between consecutive views, and then *globally* stitch together the view scans using those correspondences.

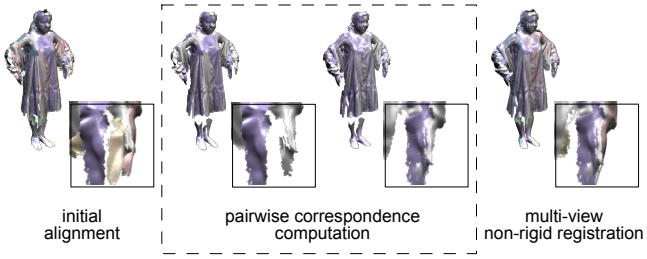


Figure 7: *The stages of registration between per-view scans. The first scan is rigidly aligned to the second, the first two are aligned to the third, and so on until all scans are rigidly aligned* (left)*. Paiwise non-rigid ICP is then used on consecutive pairs of scans* (middle) *to find dense correspondences between the scans; these are used for global multi-view non-rigid registration* (right)*.*

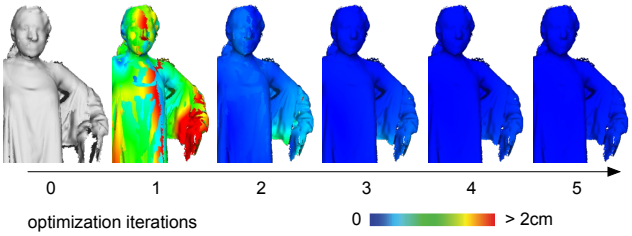


Figure 8: *We visualize the convergence motion of the view scans between consecutive iterations during multi-view non-rigid registration by projecting and color-coding the motion on the merged surface (typically converges after 5 iterations).*

Pairwise Correspondences For each rigidly aligned view scan $\mathbf{S}_i^R$, we use the robust non-rigid ICP algorithm of Li and colleagues [2009] to register it with the surface of the next view $\mathbf{S}_{i+1}^R$ (if $\mathbf{S}_i^R$ is the last view, we register it to the first). Since consecutive scans have reasonably large overlapping region of similar shape, this registration is robust (the most challenging registration is of the last scan to the first). We do not use the warped source surface directly for registration but rather sample its vertices and find those samples that lie within a certain distance ($\sim 5mm$) of the target surface. We store the sample, and its projection onto the target surface (represented by the barycentric coordinates on the target triangle), as a *correspondence pair* to be enforced during the global deformation. We use a sampling density of one sample per 15 mm, which results in roughly 1000–2000 correspondences for a typical pairs of view scans.

Global Deformation Once we have computed reliable correspondences between each pair of views, we deform all of the view scans globally by enforcing those correspondences while keeping the view scans as rigid as possible. We efficiently solve for this maximally-rigid deformation using a modification of Li et al's method [2009]: as in their approach, we represent the deformation of each view using a coarse deformation graph (we use a graph node density of one node per 50 mm), and minimize the energy

$$E = \sum_{i=1}^{n} \left[\alpha_{\text{rigid}} E_{\text{rigid}}(\mathbf{S}_i^R) + \alpha_{\text{smooth}} E_{\text{smooth}}(\mathbf{S}_i^R) \right] + \alpha_{\text{corr}} E_{\text{corr}},$$

where as in Li et al. [2009], the terms E_{rigid} and E_{smooth} measure the local rigidity of the deformation graph deformation, and the smoothness of the deformation, respectively. The new term E_{corr}

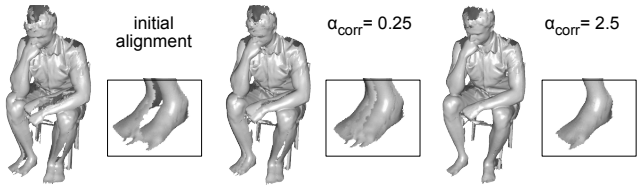
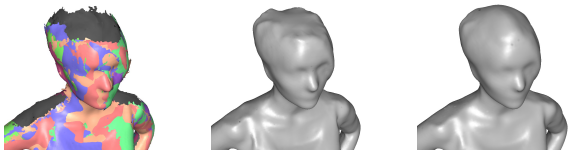


Figure 9: *E_{corr} uses the reliable pairwise correspondences from the previous step to avoid local minima during global alignment.*



(a) (b) (c)

Figure 10: *Merging with visual hull as a shape prior. The globally aligned scans contain a hole on top of the subject's head due to occlusion* (a). *Without using the visual hull, Poisson surface reconstruction flattens the hole* (b)*, whereas more pleasing results can be obtained using the visual hull as a shape prior* (c).

measures violation of the correspondences found previously:

$$E_{\text{corr}} = \frac{1}{|\mathbf{C}|} \sum_{(\mathbf{p}_1,\mathbf{p}_2)\in\mathbf{C}} \|\mathbf{p}_1 - \mathbf{p}_2\|^2$$

with $\mathbf{C}$ the set of all pairwise correspondences, and where for each correspondence, $\mathbf{p}_1$ and $\mathbf{p}_2$ are the deformed positions of the corresponding points on the surfaces.

We follow Li and coworkers [2009] and minimize E using Gauss-Newton iterations, where the linear system at each step of Gauss-Newton is solved using a sparse Cholesky factorization. Since the Jacobian of the energy is rank-deficient (rigid motions of all of the view scans does not change any of the energies, for example), we regularize the problem by adding a small (10^{-8}) multiple of the identity to the Gauss-Newton matrix. This regularization has the effect of slightly penalizing motion of the deformation graphs away from their initial configuration. We use CHOLMOD [Chen et al. 2008] to compute the Cholesky factorization, and have found that the parameters $\alpha_{\text{rigid}} = 500$, $\alpha_{\text{smooth}} = 2.0$, and $\alpha_{\text{corr}} = 2.5$ work well in practice. Gauss-Newton typically converges in under 10 iterations. Figures 3 and 7 show examples of the set of views before and after global non-rigid registration. Figure 9 illustrates the effect of E_{corr}. We found that it is often beneficial to repeat the entire non-rigid registration step several times, using the previous iteration's output instead of the rigid alignment as the starting point. There is usually no noticeable improvement after 5 such iterations. The results of this step of the algorithm are globally, non-rigidly warped view meshes $\mathbf{S}_i^N$. The convergence of the global optimization is visualized in Figure 8.

3.5 Watertight Merging

If the coverage of the warped view scans were complete, we could perform a watertight Poisson surface reconstruction [Kazhdan et al. 2006] to obtain a single consistent mesh from all the non-rigidly aligned meshes. Unfortunately, in practice there are large holes, for example on the top of the head and under the arms, where the 3D sensor is unable to gather any depth information. We therefore provide a prior to reduce discontinuous transitions between captured and missing regions, using a visual hull similar to the approach of Vlasic et al [2009].

(a) (b) (c)

Figure 11: *Mapping texture onto the merged surface directly from each scan leads to discontinuous transitions* (a). *Albedo recovery (SIRFS) offsets the lighting variations and yields consistent texture* (b). *Poisson blending further smooths the transitions and fill the texture in the occluded areas* (c).

For each warped view surface $\mathbf{S}_i^N$ we take its camera view $\mathbf{c}_i^R$ from after the initial rigid alignment (we assume that nonrigid alignment does not significantly alter the camera position or orientation) and use it to rasterize a depth map of $\mathbf{S}_i^N$. We then fit a bounding cuboid around the set of all warped meshes $\mathbf{S}_I^N$ and uniformly sample its volume (we use a density of one sample per 10 mm); for each sample, we look at all of the views and check if the ray from the sample to $\mathbf{c}_i^R$ intersects the view $\mathbf{S}_i^N$. We perform this check efficiently by projecting the sample onto the depth maps. If at least 70% of the views occlude the sample, we declare the sample inside the visual hull. In this way the visual hull calculation is robust against errors in segmentation (limbs that are missing because they are almost facing away from the camera, for instance) without introducing spurious volume.

We take the visual hull, delete all samples in its interior, and reconstruct a surface using marching cubes. This hull surface is included along with the surfaces of the warped views $\mathbf{S}_i^N$ in a final watertight Poisson surface reconstruction. The reconstruction is weighted, with the hull surface given a weight of 0.5 relative to the scanned surfaces, so that the hull surface has influence only in regions where scanned data is missing. Figure 10 illustrates the benefits of including the hull in the reconstruction.

3.6 Texturing

Since our capture is done by rotating the subject with respect to the sensor and the lighting environment, under typical non-uniform lighting conditions the appearance of the subject changes between views. The large lighting variations across the captured view scan textures make it challenging to compute a consistent global texture for the merged model. Simply averaging per-scan textures can lead to unpleasant results (see Figure 11).

To offset the lighting variations, we recover the underlying albedo from the per-scan texture using a simplification of the "Shape, Illumination, and Reflectance from Shading" (SIRFS) technique of Barron and Malik [Barron and Malik 2012; Barron and Malik 2013]. SIRFS is a unified optimization framework for recovering a depth map, reflectance (albedo) image, and global illumination from a single image, using shading cues. Because we have already acquired a high-quality depth map, we can assume depth is known, and solve a simplified version of SIRFS where only reflectance, shading, and illumination are unknown.

For each scan, we have a composite log-RGB texture image I and a composite depth map Z. With this, our simplified SIRFS problem is

$$\max_{R,L} \; P(R)P(L) \qquad \text{s.t.} \qquad I = R + S(Z, L),$$

where R is a log-reflectance image, and L is a spherical-harmonic model of illumination. $S(Z, L)$ is a rendering engine which linearizes Z into a set of surface normals, and produces a log-shading

image from those surface normals and L. $P(R)$ and $P(L)$ are priors on reflectance and illumination, respectively, whose likelihoods we maximize subject to the constraint that the log-image I is equal to a rendering of our model $R + S(Z, L)$. $P(R)$ is an elaborate "smoothness" prior on reflectance and $P(L)$ is the log-likelihood of a multivariate Gaussian distribution trained on spherical harmonic illuminations from natural environment maps (see [Barron and Malik 2012] for details). We solve this problem using the optimization technique of Barron and Malik [2012], where Z is fixed to its input value and only L and R are free parameters, giving us an estimate of log-reflectance R, which we exponentiate to produce a reflectance image. The reflectance image produced by SIRFS often has many errors, mostly due to shadows in the image, and fine-scale mistakes in the depth-map. These errors are usually in the form of new edges and gradients in the reflectance image which were not present in the original input image. To remove these mistakes, we adopt a simple heuristic: any edge in the reflectance image should have a corresponding edge in the input image. We construct two steerable pyramids, one from the image, and one from the reflectance. We then construct a composite pyramid, where for each coefficient in the pyramid, we take whichever coefficient in either the image-pyramid or the reflectance-pyramid has a smaller absolute magnitude. Once we have collapsed this composite pyramid, we have our improved reflectance image.

Despite our best efforts, non-diffusive reflectance, complex in-door illumination and inter-reflections can still introduce errors and inconsistencies unaccounted for in our reflectance computation. Additionally, occluded areas on the merged model need to be properly textured in the final output model. We therefore perform Poisson blending [Chuang et al. 2009] on the per-scan albedo maps yielding smooth texture transitions between the scans and within the occluded areas on the merged model. We find the textures near the boundary of each scan rather unreliable and thus we erode the texture of each scan by around $10mm$ from the boundary before performing Poisson blending.

4 Results

We validated our framework by scanning several subjects, including many in interesting and traditionally challenging poses. It should be stressed here that although our pipeline consists of many steps, each with several settings and parameters, all of these parameters were set *once* and were left untouched across all examples shown here.

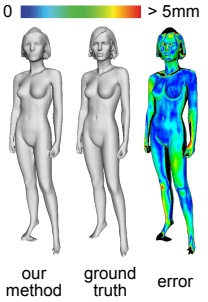


Figure 12: *A comparison of our reconstruction with a high-quality scan.*

Rigid Mannequin Validation As an initial test of our acquisition pipeline, we acquired a 3D model of a rigid mannequin (1.7 m tall) rotated using a turntable, and compared our results to a model captured using Artec's Eva, a high-performance structured light scanner with a 0.1 mm confidence interval and 0.5 mm spatial resolution. Although our reconstruction predictably loses some of the mannequin's fine details, it accurately captures the mannequin's shape and silhouette, with an average distance error of less than three millimeters (see Figure 12).

Challenging Examples Figure 15 shows several cases where our algorithm successfully captures subjects in poses that would have been impossible using previous methods. Some of the challenging features highlighted in these examples include:

- *Props* touching and occluding the subject (chair, guitar, etc.);
- *Loose-fitting, thick clothing* such as dresses and jackets;
- *Complex poses* far from the typical T-pose, including subjects sitting, standing on one leg, or covering their face. All of these poses introduce occlusions and intricate holes and concave regions that our method nevertheless successfully captures;
- *Multiple subjects* in the same scan;
- *Large deformation* that cannot be resolved using rigid alignment, such as the foot of the Thinker (first row) and head of the Guitarist (second row). All of the examples in the last row also include severe nonrigid deformation.

All of these features are critical obstructions to methods that rely on skeleton or body shape priors. Figure 13 highlights some cases where, unlike other state-of-the-art techniques that work on arbitrary geometry, our non-rigid alignment algorithm correctly registers the input view surfaces even for large deformations.

3D Printing The reconstructed models produced by our method are watertight, detailed, and well-suited for 3D printing. We chose three of our captured models and had them printed; Figure 2 shows a photograph of the resulting miniatures. They were printed using 120 grams of VeroWhite material on a Connex 500 3D printer; the total printing time for all three miniatures was six hours.

Importance of Each Pipeline Step In our experience, each step of our algorithm is essential, and cannot be omitted without severely degrading the quality of our results. Several examples demonstrate the benefits of the various portions of our method. From any of the examples in Figure 15, it is clear that rigid alignment is wholly insufficient to account for the inevitable deformations that occur as the subject rotates between the different views. Figure 6 shows the necessity of computing global nonrigid deformations of the view surfaces, rather than only nonrigidly registering the views pairwise, to avoid loop closure problems. Without a visual hull prior, missing geometry due to occlusion, poor sensor coverage, or segmentation errors degrades the shape during Poisson reconstruction (see Figure 10). Lastly, except under very controlled lighting conditions, albedo recovery is required to avoid obvious artifacts at texture seams (see Figure 11).

Algorithm Step	Section	Time (s)
Scanning with ICP registration	3.1	113
Poisson fusions (eight views)	3.2	130
Background segmentation	3.2	22
Rigid alignment	3.3	23
Nonrigid alignment	3.4	126
Albedo extraction	3.6	120*
Visual hull	3.5	14
Final watertight fusion	3.5	119
Poisson texture blending	3.6	180
Total time		**847**

Table 1: *Average performance timings (these timings are representative). All steps except the first are offline. * Albedo extraction is implemented in Matlab.*

Performance We measured the performance of our algorithms on a typical consumer desktop (CPU i7-930 2.8Ghz, 4 Cores, 12 GB RAM) and listed the average time spent during each phase of our algorithm for our data sets in Table 1; the final, watertight meshes contain about 200,000 vertices. Since all subjects were scanned within the same capture volume and used the same settings, these timings are representative. Without texture reconstruction, our algorithm needs about 2 GB of RAM; with reconstruction, 4 GB. We made no attempt to optimize memory consumption.

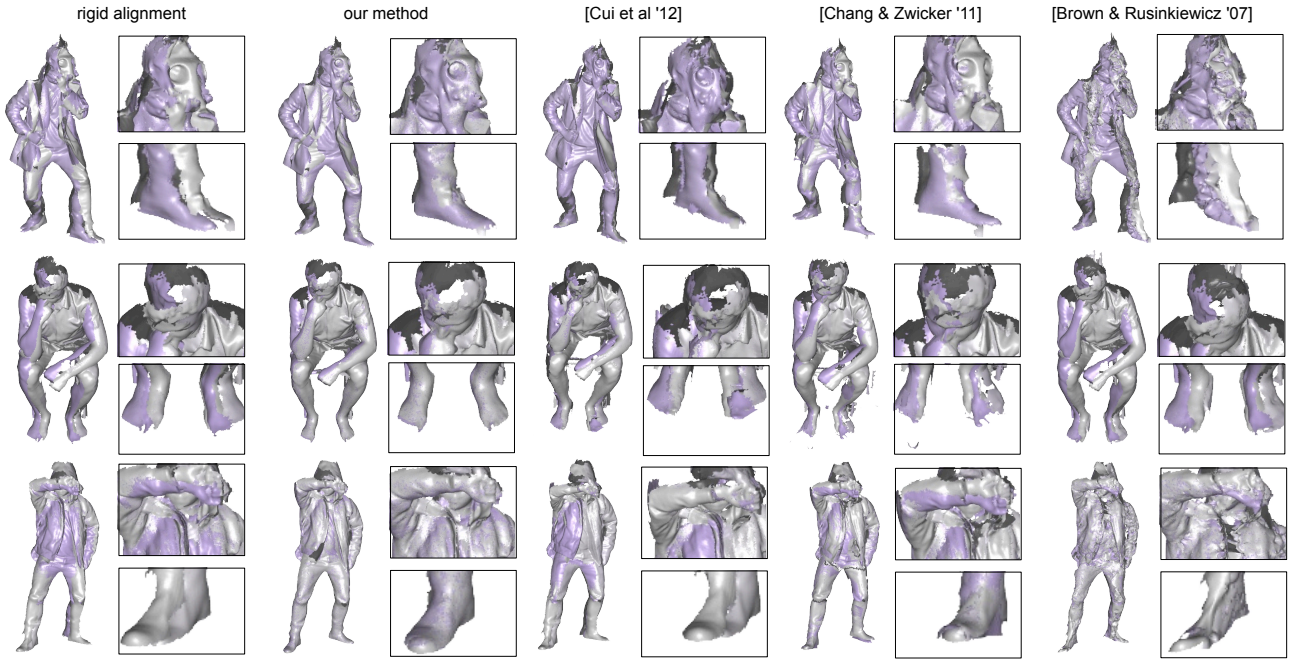


Figure 13: *Comparisons of final alignment results with the state-of-the-art methods [Cui et al. 2012; Chang and Zwicker 2011; Brown and Rusinkiewicz 2007] on a few examples. For each example we highlight two adjacent scans for visual comparison. While our method takes minutes to complete, the other techniques take several hours.*

Figure 14: *Two failure cases for our system: significant movement of subject's head (a) between scans may lead to the failure of global non-rigid alignment (b); and large illumination change (c) results in visible artifacts in the final texture (d).*

5 Limitations and Future Work

Although we have shown that our method is robust for a wide variety of clothing and poses, and captures high-quality and detailed models of the subjects despite the use of only a single low-quality sensor, there are also failure cases.

Our global non-rigid registration and warping algorithm requires that the subjects stand as still as possible. Some movement is unavoidable and our method usually correctly accounts for it, especially relative to previous methods, but particularly large deformation can cause registration to fail. Figure 14, left, shows one example where excessive tilting of the head results in poor non-rigid alignment. Since it is discouraging to the user to pose for two minutes only to have the system fail, and since increasing the robustness of registration would allow even more ambitious, complex poses and "action shots", and would allow us to scan uncooperative subjects like babies or pets, it is one of our highest priorities for future research. One promising approach would be to search *all pairs* of view scans for correspondences, instead of only considering consecutive views; the primary challenge of such an approach is that non-consecutive views have less overlap, and so non-rigid ICP is less reliable; filtering out noisy and erroneous correspondences becomes much more important. Similarly, excessive occlusion of parts of the body, particularly when scanning multiple people at once, poses a registration challenge. All-pairs correspondences would help these cases as well.

Several potential improvements are possible to our texture recovery pipeline. Currently, during view fusion (Section 3.2), we construct the texture of each view by simple averaging of the raw frames captured by the sensor. Incorporating image registration techniques, and coupling it to our current geometry registration using rigid ICP, would improve the quality of both the texture and the geometry. Extreme changes in lighting between the views (see for instance Figure 14, right) are not always fully corrected by current albedo reconstruction and result in visible artifacts. Correcting for texturing artifacts caused by shadows, or anisotropic materials like velvet, also remains future work. We would like to improve the user experience by making as much of our processing online as possible, so that we can give the user real-time visual feedback of potential problems in the scanning process, allow the user to self-correct large deformations, etc. One interesting avenue for improving the efficiency of our pipeline would be to incorporate priors based on human articulations, in a way that does not compromise our method's robustness to complex poses, loose clothing, and props.

The current visual hull prior for final Poisson reconstruction (Section 3.5) works best for nearly-convex poses. Complex holes and convex regions, which appear solid from the majority of view directions, are sometimes filled in incorrectly by our current voting-based strategy. We would like to explore alternative strategies that are robust to segmentation and scanning errors, while still preserving such voids. Lastly, very thin features, such as canes or the brims of hats, are sometimes lost during scanning and segmentation. Like many of the challenges described above, inevitable improvements in the accuracy and resolution of affordable depth sensors will decrease the prevalence of these artifacts over time.

Acknowledgements. We thank Nadezhda Artemeva, Ksusha Katz, Maddy Kloss, Isabelle Coler, and all the models for their scans. Thank you Benedict J. Brown, Will Chang, and Yan Cui for producing the comparisons. We also thank Oytun Akman and

Mitko Vidanovski (Autodesk Reality Capture) and Bernd Bickel (Disney Research Zurich) for the 3D prints.

References

ANGUELOV, D., SRINIVASAN, P., KOLLER, D., THRUN, S., RODGERS, J., AND DAVIS, J. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, ACM, New York, NY, USA.

BARRON, J. T., AND MALIK, J. 2012. Shape, albedo, and illumination from a single image of an unknown object. *CVPR*.

BARRON, J. T., AND MALIK, J. 2013. Intrinsic scene properties from a single rgb-d image. *CVPR*.

BOJSEN-HANSEN, M., LI, H., AND WOJTAN, C. 2011. Tracking surfaces with evolving topology. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2012) 31*, 4.

BROWN, B., AND RUSINKIEWICZ, S. 2007. Global non-rigid alignment of 3-D scans. *ACM Transactions on Graphics (Proc. SIGGRAPH) 26*, 3 (Aug.).

CHANG, W., AND ZWICKER, M. 2008. Automatic registration for articulated shapes. *Computer Graphics Forum (Proceedings of SGP 2008) 27*, 5.

CHANG, W., AND ZWICKER, M. 2009. Range scan registration using reduced deformable models. *EG*.

CHANG, W., AND ZWICKER, M. 2011. Global registration of dynamic range scans for articulated model reconstruction. *ACM Transactions on Graphics, to appear 30*, 3.

CHEN, Y., DAVIS, T. A., HAGER, W. W., AND RAJAMANICKAM, S. 2008. Algorithm 887: Cholmod, supernodal sparse cholesky factorization and update/downdate. *ACM Trans. Math. Softw. 35*, 3 (Oct.), 22:1–22:14.

CHUANG, M., LUO, L., BROWN, B. J., RUSINKIEWICZ, S., AND KAZHDAN, M. 2009. Estimating the Laplace-Beltrami operator by restricting 3D functions. *SGP 2009* (July).

CUI, Y., CHANG, W., NOLL, T., AND STRICKER, D. 2012. Kinectavatar: Fully automatic body capture using a single kinect. In *Asian Conference on Computer Vision (ACCV)*.

HASLER, N., STOLL, C., SUNKEL, M., ROSENHAHN, B., AND SEIDEL, H.-P. 2009. A statistical model of human pose and body shape. In *Computer Graphics Forum (Proc. Eurographics 2008)*, P. Dutr'e and M. Stamminger, Eds., vol. 2.

HUANG, Q.-X., ADAMS, B., WICKE, M., AND GUIBAS, L. J. 2008. Non-rigid registration under isometric deformations. In *Proceedings of the Symposium on Geometry Processing*, Eurographics Association, SGP '08.

KAZHDAN, M., BOLITHO, M., AND HOPPE, H. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, EG Assoc.

LI, H., SUMNER, R. W., AND PAULY, M. 2008. Global correspondence optimization for non-rigid registration of depth scans. *Computer Graphics Forum (Proc. SGP'08) 27*, 5 (July).

LI, H., ADAMS, B., GUIBAS, L. J., AND PAULY, M. 2009. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (SIGGRAPH Asia 2009) 28*, 5 (December).

LI, H., LUO, L., VLASIC, D., PEERS, P., POPOVIĆ, J., PAULY, M., AND RUSINKIEWICZ, S. 2012. Temporally coherent completion of dynamic shapes. *ACM ToG 31*, 1 (January).

MITRA, N. J., FLÖRY, S., OVSJANIKOV, M., GELFAND, N., GUIBAS, L., AND POTTMANN, H. 2007. Dynamic geometry registration. In *Proceedings of the fifth Eurographics symposium on Geometry processing*, Eurographics Association, SGP '07.

NEWCOMBE, R. A., IZADI, S., HILLIGES, O., MOLYNEAUX, D., KIM, D., DAVISON, A. J., KOHLI, P., SHOTTON, J., HODGES, S., AND FITZGIBBON, A. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, ISMAR '11, 127–136.

PULLI, K. 1999. Multiview registration for large data sets. In *Proceedings of the 2nd international conference on 3-D digital imaging and modeling*.

RUSINKIEWICZ, S., AND LEVOY, M. 2001. Efficient variants of the ICP algorithm. In *Third International Conference on 3D Digital Imaging and Modeling (3DIM)*.

RUSINKIEWICZ, S., HALL-HOLT, O., AND LEVOY, M. 2002. Real-time 3d model acquisition. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA.

SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms.

SHARF, A., ALCANTARA, D. A., LEWINER, T., GREIF, C., SHEFFER, A., AMENTA, N., AND COHEN-OR, D. 2008. Space-time surface reconstruction using incompressible flow. In *ACM SIGGRAPH Asia 2008 papers*, ACM, New York.

SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., AND BLAKE, A. 2011. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Washington, DC, USA.

TEVS, A., BERNER, A., WAND, M., IHRKE, I., BOKELOH, M., KERBER, J., AND SEIDEL, H.-P. 2012. Animation cartography–intrinsic reconstruction of shape and motion. *ACM Trans. Graph. 31*, 2 (Apr.).

TONG, J., ZHOU, J., LIU, L., PAN, Z., AND YAN, H. 2012. Scanning 3d full human bodies using kinects. *IEEE Transactions on Visualization and Computer Graphics 18*, 4.

VLASIC, D., PEERS, P., BARAN, I., DEBEVEC, P., POPOVIĆ, J., RUSINKIEWICZ, S., AND MATUSIK, W. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics 28*, 5, 174.

WAND, M., ADAMS, B., OVSJANIKOV, M., BERNER, A., BOKELOH, M., JENKE, P., GUIBAS, L., SEIDEL, H.-P., AND SCHILLING, A. 2009. Efficient reconstruction of nonrigid shape and motion from real-time 3d scanner data. *ACM Trans. Graph. 28*, 2 (May).

WEI, X., ZHANG, P., AND CHAI, J. 2012. Accurate realtime full-body motion capture using a single depth camera. *ACM Trans. Graph. 31*, 6 (Nov.).

WEISE, T., WISMER, T., LEIBE, B., AND GOOL, L. V. 2011. Online loop closure for real-time interactive 3d scanning. *Comput. Vis. Image Underst. 115*, 5 (May).

WEISS, A., HIRSHBERG, D., AND BLACK, M. 2011. Home 3D body scans from noisy image and range data. In *Int. Conf. on Computer Vision (ICCV)*, IEEE, Barcelona.

Figure 15: *Example 3D self-portraits captured by our system. For each example we show the input raw scans, the initial scans registered by rigid alignment, the scans refined by multi-view non-rigid alignment, the merged surfaces and reconstructed textures. Using only a single static Kinect sensor, our algorithm successfully reconstructs textured geometry of challenging examples (arbitrary poses, subjects with props, loose clothing, multiple subjects, etc.). While being more general and efficient than existing methods, our system can also handle inconsistent illuminations across views and is particularly robust to large deformations.*