Capturing Dynamic Textured Surfaces of Moving Targets

Ruizhe Wang[1(✉)], Lingyu Wei[1], Etienne Vouga[2], Qixing Huang[2,3], Duygu Ceylan[4], Gérard Medioni[1], and Hao Li[1]

[1] University of Southern California, Los Angeles, USA
{ruizhewa,lingyu.wei,medioni}@usc.edu, hao@hao-li.com
[2] University of Texas at Austin, Austin, USA
{evouga,huangqx}@cs.utexas.edu
[3] Toyota Technological Institute at Chicago, Chicago, USA
huangqx@ttic.edu
[4] Adobe Research, San Jose, USA
ceylan@adobe.com

Abstract. We present an end-to-end system for reconstructing complete watertight and textured models of moving subjects such as clothed humans and animals, using only three or four handheld sensors. The heart of our framework is a new pairwise registration algorithm that minimizes, using a particle swarm strategy, an alignment error metric based on mutual visibility and occlusion. We show that this algorithm reliably registers partial scans with as little as 15 % overlap without requiring any initial correspondences, and outperforms alternative global registration algorithms. This registration algorithm allows us to reconstruct moving subjects from free-viewpoint video produced by consumer-grade sensors, without extensive sensor calibration, constrained capture volume, expensive arrays of cameras, or templates of the subject geometry.

Keywords: Range image registration · Particle swarm optimization · Dynamic surface reconstruction · Free-viewpoint video · Moving target · Texture reconstruction

1 Introduction

The rekindling of interest in immersive, 360° virtual environments, spurred on by the Oculus, Hololens, and other breakthroughs in consumer AR and VR hardware, has birthed a need for digitizing objects with full geometry and texture from all views. One of the most important objects to digitize in this way are moving, clothed humans, yet they are also among the most challenging: the human body can undergo large deformations over short time spans, has complex

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46478-7_17) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VII, LNCS 9911, pp. 271–288, 2016.
DOI: 10.1007/978-3-319-46478-7_17

geometry with occluded regions that can only be seen from a small number of angles, and has regions like the face with important high-frequency features that must be faithfully preserved.

Most techniques for capturing high-quality digital humans rely on a large array of sensors mounted around a fixed capture volume. The recent work of Collet et al. [11] uses such a setup to capture live performances and compresses them to enable streaming of free-viewpoint videos. Unfortunately, these techniques are severely restrictive: first, to ensure high-quality reconstruction and sufficient coverage, a large number of expensive sensors must be used, leaving human capture out of reach of consumers without the resources of a professional studio. Second, the subject must remain within the small working volume enclosed by the sensors, ruling out subjects interacting with large, open environments or undergoing large motions.

Using free-viewpoint sensors is an attractive alternative, since it does not constrain the capture volume and allows ordinary consumers, with access to only portable, low-cost devices, to capture human motion. The typical challenge with using hand-held active sensors is that, obviously, multiple sensors must be used simultaneously from different angles to achieve adequate coverage of the subject. In overlapping regions, signal interference causes significant deterioration in the quality of the captured geometry. This problem can be avoided by minimizing the amount of overlap between sensors, but on the other hand, existing registration algorithms for aligning the captured partial scans only work reliably if the partial scans significantly overlap. Template-based methods like the work of Ye et al. [54] circumvent these difficulties by warping a full geometric template to track the moving sparse partial scans, but templates are only readily available for naked humans [4]; for clothed humans a template must be precomputed on a case-by-case basis.

We thus introduce a new shape registration method that can reliably register partial scans even with *almost no overlap*, sidestepping the need for shape templates or sensor arrays. This method is based on a *visibility error metric* which encodes the intuition that if a set of partial scans are properly registered, each partial scan, when viewed from the same angle at which it was captured, should occlude all other partial scans. We solve the global registration problem by minimizing this error metric using a particle swarm strategy, to ensure sufficient coverage of the solution space to avoid local minima. This registration method significantly outperforms state of the art global registration techniques like 4PCS [3] for challenging cases of small overlap.

Contributions. We present the first end-to-end free-viewpoint reconstruction framework that produces watertight, fully-textured surfaces of moving, clothed humans using only three to four handheld depth sensors, without the need of shape templates or extensive calibration. The most significant technical component of this system is a robust pairwise global registration algorithm, based on minimizing a visibility error metric, that can align depth maps even in the presence of very little (15 %) overlap.

2 Related Work

Digitizing realistic, moving characters has traditionally involved an intricate pipeline including modeling, rigging, and animation. This process has been occasionally assisted by 3D motion and geometry capture systems such as marker-based motion capture or markerless capture methods involving large arrays of sensors [12]. Both approaches supply artists with accurate reference geometry and motion, but they require specialized hardware and a controlled studio setting.

Real-time 3D scanning and reconstruction systems requiring only a single sensor, like KinectFusion [18], allow casual users to easily scan everyday objects; however, as with most simultaneous localization and mapping (SLAM) techniques, the major assumption is that the scanned scene is rigid. This assumption is invalid for humans, even for humans attempting to maintain a single pose; several follow-up works have addressed this limitation by allowing near-rigid motion, and using non-rigid partial scan alignment algorithms [24,44]. While the recent DynamicFusion framework [31] and similar systems [13] show impressive results in capturing non-rigidly deforming scenes, our goal of capturing and tracking freely moving targets is fundamentally different: we seek to reconstruct a *complete* model of the moving target at all times, which requires either extensive prior knowledge of the subject's geometry, or the use of multiple sensors to provide better coverage.

Prior work has proposed various simplifying assumptions to make the problem of capturing entire shapes in motion tractable. Examples include assuming availability of a template, high-quality data, smooth motion, and a controlled capture environment.

Template-Based Tracking: The vast majority of related work on capturing dynamic motion focuses on specific human parts, such as faces [25] and hands [32,33], for which specialized shapes and motion templates are available. In the case of tracking the full human body, parameterized body models [5] have been used. However, such models work best on naked subjects or subjects wearing very tight clothing, and are difficult to adapt to moving people wearing more typical garments.

Another category of methods first capture a template in a static pose and then track it across time. Vlasic et al. [45] use a rigged template model, and De Aguiar et al. [1] apply a skeleton-less shape deformation model to the template to track human performances from multi-view video data. Other methods [22,56] use a smoothed template to track motion from a capture sequence. The more recent work of Wu et al. [51] and Liu et al. [26] track both the surface and the skeleton of a template from stereo cameras and sparse set of depth sensors respectively.

All of these template-based approaches handle with ease the problem of tracking moving targets, since the entire geometry of the target is known. However, in addition to requiring constructing or fitting said template, these methods share the common limitation that they cannot handle geometry or topology changes

which are likely to happen during typical human motion (picking up an object; crossing arms; etc.).

Dynamic Shape Capture: Several works have proposed to reconstruct both shape and motion from a dynamic motion sequence. Given a series of time-varying point clouds, Wand et al. [48] use a uniform deformation model to capture both geometry and motion. A follow-up work [47] proposes to separate the deformation models used for geometry and motion capture. Both methods make the strong assumption that the motion is smooth, and thus suffer from popping artifacts in the case of large motions between time steps. Süßmuth et al. [42] fit a 4D space-time surface to the given sequence but they assume that the complete shape is visible in the first frame. Finally, Tevs et al. [43] detect landmark correspondences which are then extended to dense correspondences. While this method can handle a considerable amount of topological change, it is sensitive to large acquisition holes, which are typical for commercial depth sensors.

Another category of related work aims to reconstruct a deforming watertight mesh from a dynamic capture sequence by imposing either visual hull [46] or temporal coherency constraints [23]. Such constraints either limit the capture volume or are not sufficient to handle large holes. Furthermore, neither of these methods focus on propagating texture to invisible areas; in contrast, we use dense correspondences to perform texture inpainting in non-visible regions. Bojsen-Hansen et al. [6] also use dense correspondences to track surfaces with evolving topologies. However, their method requires the input to be a closed manifold surface. Our goal, on the other hand, is to reconstruct such complete meshes from sparse partial scans.

The recent work of Collet et al. [11] uses multimodal input data from a stage setup to capture topologically-varying scenes. While this method produces impressive results, it requires a pre-calibrated complex setup. In contrast, we use a significantly cheaper and more convenient setup composed of three to four commercial depth sensors.

Global Range Image Registration: At the heart of our approach is a robust algorithm that registers noisy data coming from each commercial depth sensor with very little overlap. A typical approach is to first perform global registration to compute an approximate rigid transformation between a pair of range images, which is then used to initialize local registration methods (e.g., Iterative Closest Point (ICP) [8,55]) for further refinement. A popular approach for global registration is to construct feature descriptors for a set of interest points which are then correlated to estimate a rigid transformation. Spin-images [19], integral volume descriptors [15], and point feature histograms (PFH, FPFH) [35,36] are among the popular descriptors proposed by prior work. Makadia et al. [27] represent each range image as a translation-invariant emphextended gaussian Image (EGI) [17] using surface normals. They first compute the optimum rotation by correlating two EGIs and further estimate the corresponding translation using Fourier transform. For noisy data as coming from a commercial depth sensor,

however, it is challenging to compute reliable feature descriptors. Another approach for global registration is to align either main axes extracted by principal component analysis (PCA) [10] or a sparse set of control points in a RANSAC loop [7]. Silva et al. [40] introduce a robust *surface interpenetration measure (SIM)* and search the 6 DoF parameter space with a genetic algorithm. More recently, Yang et al. [53] adopt a branch-and-bound strategy to extend the basic ICP algorithm in a global manner. 4PCS [3] and its latest variant Super-4PCS [29] register a pair of range images by extracting all coplanar 4-points sets. Such approaches, however, are likely to converge to wrong alignments in cases of very little overlap between the range images (see Sect. 5).

Several prior works have adopted silhouette-based constraints for aligning multiple images [2,14,16,28,30,41,49,52]. While the idea is similar to our approach, our registration algorithm also takes advantage of depth information, and employs a particle-swarm optimization strategy that efficiently explores the space of alignments.

3 System Overview

Our pipeline for reconstructing fully-textured, watertight meshes from three to four depth sensors can be decomposed into four major steps. See Fig. 1 for an overview.

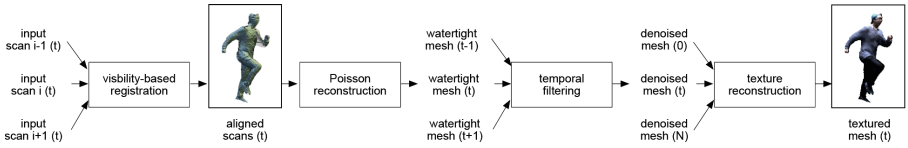


Fig. 1. An overview of our textured dynamic surface capturing system.

1. Data Capture: We capture the subject (who is free to move arbitrarily) using uncalibrated hand-held real-time RGBD sensors. We experimented with both Kinect One time-of-flight cameras mounted on laptops, and Occipital Structure IO sensors mounted on iPad Air 2 tablets (Sect. 6).

2. Global Rigid Registration: The relative positions of the depth sensors constantly change over time, and the captured depth maps often have little overlap (10 %–30 %). For each frame, we globally register sparse depth images from all views (Sect. 4). This step produces registered, but incomplete, textured partial scans of the subject.

3. Surface Reconstruction: To reduce flickering artifacts, we adopt the shape completion pipeline of Li et al. [23] to warp partial scans from temporally-proximate frames to the current frame geometry. A weighted Poisson reconstruction step then extracts a single watertight surface. There is no guarantee, however, that the resulted fused surface has complete texture coverage

(and indeed typically texture will be missing at partial scan seams and in occluded regions.)

4. Dense Correspondences for Texture Reconstruction: We complete regions of missing or unreliable texture on one frame by propagating data from other (perhaps very temporally-distant) frames with reliable texture in that region. We adopt a recently-proposed correspondence computation framework [50] based on a deep neural network to build dense correspondences between any two frames, even if the subject has undergone large relative deformations. Upon building dense correspondences, we transfer texture from reliable regions to less reliable ones.

We next mainly describe the details of the global registration method. Please refer to the supplementary material for more details of the other components.

4 Robust Rigid Registration

The key technical challenge in our pipeline is registering a set of depth images accurately without assuming any initialization, even when the geometry visible in each depth image has very little overlap with any other depth image. We attack this problem by developing a robust pairwise global registration method: let P_1 and P_2 be partial meshes generated from two depth images captured simultaneously. We seek a global Euclidean transformation T_{12} which aligns P_2 to P_1. Traditional pairwise registration based on finding corresponding points on P_1 and P_2, and minimizing the distance between them, has notorious difficulty in this setting. As such we propose a novel *visibility error metric* (VEM) (Sect. 4.1), and we minimize the VEM to find T_{12} (Sect. 4.2). We further extend this pairwise method to handle multi-view global registration (Sect. 4.3).

4.1 Visibility Error Metric

Suppose P_1 and P_2 are correctly aligned, and consider looking at the pair of scans through a camera whose position and orientation matches that of the sensor used to capture P_1. The only parts of P_2 that should be visible from this view are those that overlap with P_1: parts of P_2 that do not overlap should be completely occluded by P_1 (otherwise they would have been detected and included in P_1). Similarly, when looking at the scene through the camera that captured P_2, only parts of P_1 that overlap with P_2 should be visible.

Visibility-Based Alignment Error. We now formalize the above idea. Let

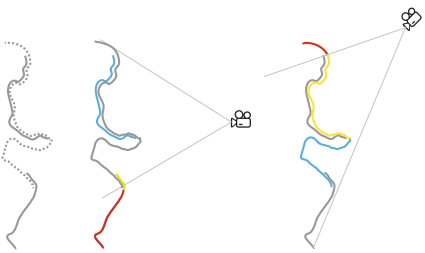


Fig. 2. Left: two partial scans P_1 (dotted) and P_2 (solid) of a 2D human. Middle: when viewed from P_1's camera, points of P_2 are classified into $\mathcal{O}$ (blue), $\mathcal{F}$ (yellow), and $\mathcal{B}$ (red). Right: when viewed from P_2's camera, points of P_1 are classified into $\mathcal{O}$ (blue), $\mathcal{F}$ (yellow), and $\mathcal{B}$ (red). (Color figure online)

P_1, P_2 be two partial scans, with P_1 captured using a sensor at position c_p and view direction c_v. For every point $x \in P_2$, let $I(x)$ be the first intersection point of P_1 and the ray $\overrightarrow{c_p x}$. We can partition P_2 into three regions, and associate to each region an energy density $d(x, P_1)$ measuring the extent to which points x in that region violate the above visibility criteria:

– points $x \in \mathcal{O}$ that are occluded by P_1: $\|x - c_p\| \geq \|I(x) - c_p\|$. To points in this region we associate no energy:

$$d_{\mathcal{O}}(x, P_1) = 0.$$

– points $x \in \mathcal{F}$ that are in front of P_1: $\|x - c_p\| < \|I(x) - c_p\|$. Such points might exist even when P_1 and P_2 are well-aligned, due to surface noise and roughness, etc. However, we penalize large violations using:

$$d_{\mathcal{F}}(x, P_1) = \|x - I(x)\|^2.$$

– points $x \in \mathcal{B}$ for which $I(x)$ does not exist. Such points also violate the visibility criteria. It is tempting to penalize such points proportionally to the distance between x and its closest point on P_1, but a small misalignment could create a point in $\mathcal{B}$ that is very distant from P_1 in Euclidean space, despite being very close to P_1 on the camera image plane. We therefore penalize x using squared distance on the image plane,

$$d_{\mathcal{B}}(x, P_1) = \min_{y \in S_1} \|\mathcal{P}_{c_v} x - \mathcal{P}_{c_v} y\|^2,$$

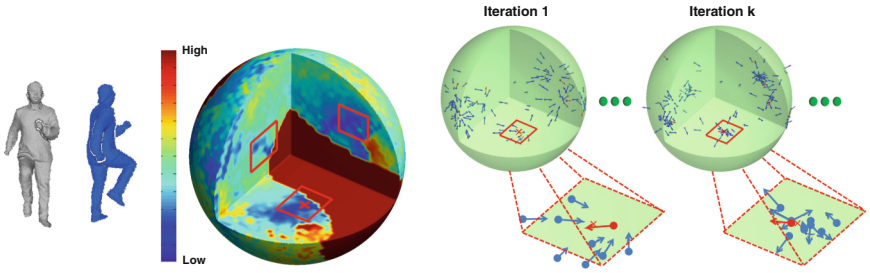
where $\mathcal{P}_{c_v}$ is the projection $I - c_v c_v^T$ onto the plane orthogonal to c_v.

Figure 2 illustrates these regions on a didactic 2D example. Alignment of P_1 and P_2 from the point of view of P_1 is then measured by the aggregate energy $d(P_2, P_1) = \sum_{x \in P_2} d(x, P_1)$. Finally, every Euclidean transformation T_{12} that produces a possible alignment between P_1 and P_2 can be associated with an energy to define our visibility error metric on $SE(3)$,

$$E(T_{12}) = d\left(T_{12}^{-1} P_1, P_2\right) + d\left(T_{12} P_2, P_1\right). \tag{1}$$

4.2 Finding the Transformation

Minimizing the error metric (1) consists of solving a nonlinear least squares problem and so in principle can be optimized using e.g. the Gauss-Newton method. However, it is non-convex, and prone to local minima (Fig. 3(a)). Absent a straightforward heuristic for picking a good initial guess, we instead adopt a Particle Swarm Optimization (PSO) [21] method to efficiently minimize (1), where "particles" are candidate rigid transformations that move towards smaller energy landscapes in $SE(3)$. We could independently minimize E starting from each particle as an initial guess, but this strategy is not computationally tractable. So we iteratively update all particle positions in lockstep: a small set of the



(a) (b)

Fig. 3. (a) Left: a pair of range images to be registered. Right: VEM evaluated on the entire rotation space. Each point within the unit ball represents the vector part of a unit quaternion; for each quaternion, we estimate its corresponding translation component and evaluate the VEM on the composite transformation. The red rectangles indicate areas with local minima, and the red cross is the global minimum. (b) Example particle locations and displacements at iteration 1 and k. Blue vectors indicate displacement of regular (non-guide) particles following a traditional particle swarm scheme. Red vectors are displacements of guide particles. Guide particles draw neighboring regular particles more efficiently towards local minima to search for the global minimum. (Color figure online)

most promising *guide* particles, that are most likely to be close to the global minimum, are updated using an iteration of Levenberg-Marquardt. The rest of the particles receive PSO-style weighted random perturbations. This procedure is summarized in Algorithm 1, and each step is described in more detail below.

Initial Particle Sampling We begin by sampling N particles (we use $N = 1600$), where each particle represents a rigid motion $m_i \in SE(3)$. Since $SE(3)$ is not compact, it is not straightforward to directly sample the initial particles. We instead uniformly sample only the rotational component R_i of each particle

Algorithm 1. Modified Particle Swarm Optimization

Input: A set of initial "particles" (orientations) $\{\mathbf{T}_1^0, ..., \mathbf{T}_N^0\} \in SE(3)^N$
evaluate VEM on initial particles
for each iteration **do**
    select guide particles
    for each guide particle **do**
        update guide particle using Levenberg-Marquardt
    end for
    for each regular particle **do**
        update particle using weighted random displacement
    end for
    recalculate VEM at new locations
end for
Output: The best particle $\mathbf{T}^b$

[39], and solve for the best translation using the following Hough-transform-like procedure. For every $x \in P_1$ and $y \in R_i P_2$, we measure the angle between their respective normals, and if it is less than 20°, the pair (x, y) votes for a translation of $y - x$. These translations are binned (we use 10 mm × 10 mm × 10 mm bins) and the best translation $\mathbf{t}_i^0$ is extracted from the bin with the most votes. The translation estimation procedure is robust even in the presence of limited overlap amount (Fig. 4).

The above procedure yields a set $\mathcal{T}^0 = \{T_i^0\} = \{(R_i^0, \mathbf{t}_i^0)\}$ of N initial particles. We next describe how to step the orientation particles from their values $\mathcal{T}^k$ at iteration k to $\mathcal{T}^{k+1}$ at iteration $k+1$.

Identifying Guide Particles. We want to select as guide particles those particles with lowest visibility error metric; however we don't want many clustered redundant guide particles. Therefore we first promote the particle T_i^k with lowest error metric to guide particle, then remove from consideration all nearby particles, e.g. those that satisfy

$$d_\theta(R_j^k, R_i^k) \leq \theta_r,$$

where $d_\theta(R_i^k, R_j^k) = \theta\left(\log\left[R_j^k\right]^{-1} R_i^k\right)$ is the bi-invariant metric on $SO(3)$, e.g. the least angle of all rotations R with $R_i^k = RR_j^k$. We use $\theta_\tau = 30°$. We then repeat this process (promoting the remaining particle with lowest VEM, removing nearby particles, etc.) until no candidates remain.

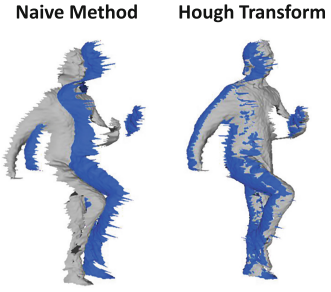


Fig. 4. Translation estimation examples of our Hough Transform method on range scans with limited overlap. The naïve method, which simply aligns the corresponding centroids, fails to estimate the correct translation.

Guide Particle Update. We update each guide particle T_i^k to decrease its VEM. We parameterize the tangent space of $SE(3)$ at T_i^k by two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ with $\exp(\mathbf{u}, \mathbf{v}) = \left(\exp([u]_\times) R_i^k, \mathbf{t}_i^k + \mathbf{v}\right)$, where $[u]_\times$ is the cross-product matrix. We then use the Levenberg-Marquardt method to find an energy-decreasing direction $(\mathbf{u}, \mathbf{v})$, and set $T_i^{k+1} = \exp(\mathbf{u}, \mathbf{v})$. Please see the supplementary material for more details.

Other Particle Update. Performing a Levenberg-Marquardt iteration on all particles is too expensive, so we move the remaining non-guide particles by applying a randomly weighted summation of each particle's displacement during the previous iteration, the displacement towards its best past position, and the displacement towards the local best particle within radius θ_r (measured using d_θ) with lowest energy, as in standard PSO [21]. While the guide particles rapidly descend to local minima, they are also local best particles and drag neighboring regular particles with them for a more efficient search of all local minima, from which the global one is extracted (Fig. 3(b)). Please refer to the supplementary material for more details.

Termination. Since the VEM of each guide particle is guaranteed to decrease during every iteration, the particle with lowest energy is always selected as a

guide particle, and the local minima of E must lie in a bounded subset of $SE(3)$. In the above procedure the particle with lowest energy is guaranteed to converge to a local minimum of E. We terminate the optimization when $\min_i |E(T_i^k) - E(T_i^{k+1})| \leq 10^{-4}$. In practice this occurs within 5–10 iterations.

4.3 Multi-view Extension

We extend our VEM-based pairwise registration method to globally align a total of M partial scans $\{P_1, ..., P_M\}$ by estimating the optimum transformation set $\{T_{12}, ..., T_{1M}\}$. First we perform pairwise registration between all pairs to build a registration graph, where each vertex represents a partial scan and each pair of vertices are linked by an edge of the estimated transformation. We then extract all spanning trees from the graph, and for each spanning tree we calculate its corresponding transformation set $\{T_{12}, ..., T_{1M}\}$ and estimate the overall VEM as,

$$E_M = \sum_{i \neq j} d\left(T_{1j}^{-1} T_{1i} P_i, P_j\right) + d\left(T_{1i}^{-1} T_{1j} P_j, P_i\right). \tag{2}$$

We select the transformation set with the minimum overall VEM. We perform several iterations of Levenberg-Marquardt algorithm to minimize Eq. 2 to further jointly refine the transformation set. We enforce temporal coherence into the global registration framework by adding the final estimated transformation set of the previous frame to the pool of transformation sets of the current frame before selecting the best one.

5 Global Registration Evaluation

Data Sets. We evaluate our registration algorithm on the Stanford 3D Scanning Repository and the Princeton Shape Benchmark [38]. We use 4 models from the Stanford 3D Scanning Repository (the Bunny, the Happy Buddha, the Dragon, and the Amardillo), and use all 1814 models from the Princeton Shape Benchmark. We believe these two data sets, especially the latter, are general enough to cover shape variation of real world objects. For each data set, we generated 1000 pairs of synthetic depth images with uniformly varying degrees of overlap; these range maps were synthesized using randomly-selected 3D models and randomly-selected camera angles. Each pair is then initialized with a random initial relative transformation. As such, for each pair of range images, we have the ground truth transformation as well as their overlap ratio.

Evaluation Metric. The extracted transformation, if not correctly estimated, can be at any distance from the ground truth transformation, depending on the specific shape of the underlying surfaces and the local minima distribution of the solution space. Thus, it is not very informative to directly use the RMSE of rotation and translation estimation. It is rather straightforward to use success percentage as the evaluation metric. We claim the global registration to be successful if the error $d_\theta(R_{est}, R_{gt})$ of the estimated rotation R_{est} is smaller than a

small angle 10°. We do not enforce the translation to be close since it is scale-dependent and the translation component is easily recovered by a robust local registration method if the rotation component is close enough (e.g., by using surface normals to prune incorrect correspondences [34]).

Effectiveness of the PSO Strategy. To demonstrate the advantage of the particle-swarm optimization strategy, we compare our full algorithm to three alternatives on the Stanford 3D Scanning Repository: (1) a baseline method that simply reports the minimum particles from all initially-sampled particles, with no attempt at optimization; (2) using only a traditional PSO formulation, without guide particles; and (3) updating only the guide particles, and applying no displacement to ordinary particles.

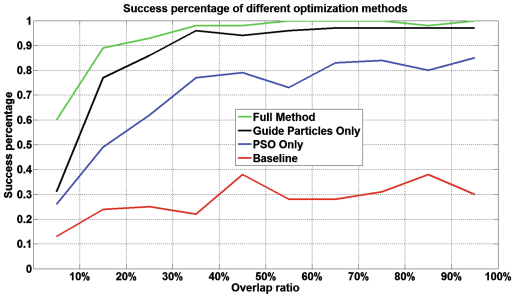


Fig. 5. Success percentage of the global registration method employing different optimization schemes on the Stanford 3D Scanning Repository.

Figure 5 compares the performance of the four alternatives. While updating guide particles alone achieves good registration results, incorporating the swarm intelligence further improves the performance, especially when overlap ratios drop below 30 %.

Comparisons. To demonstrate the effectiveness of the proposed registration method, we compare it against four other alternatives: (1) a baseline method that aligns principal axes extracted with weighted PCA [10], where the weight of each vertex is proportional to its local surface area; (2) Go-ICP [53], which combines local ICP with a branch-and-bound search to find the global minima; (3) FPFH [35,37], which matches FPFH descriptors; (4) 4PCS, a state-of-the-art method that performs global registration by constructing a congruent set of 4 points between range images [3]. We do not compare with its latest variant SUPER-4PCS [29] as only efficiency is improved for the latter. For Go-ICP, FPFH and 4PCS, we use the authors' original implementation and tune parameters to achieve optimum performance.

Figure 6 compares the performance of the five methods on the two data sets respectively. The overall performance on the Princeton Shape Benchmark is

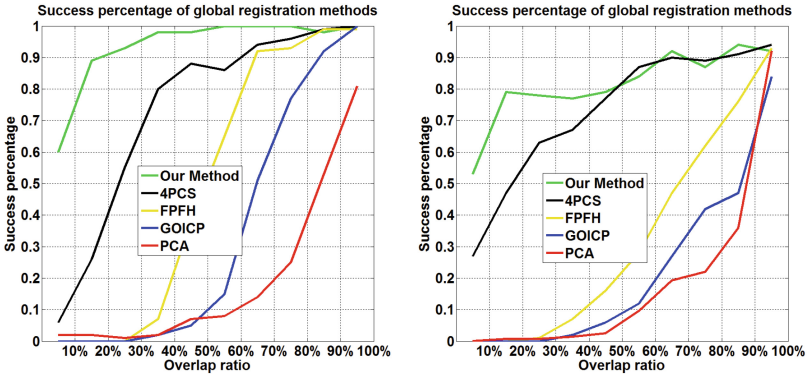


Fig. 6. Success percentage of our global registration method compared with other methods. Left: Comparison on the Stanford 3D Scanning Repository. Right: Comparison on the Princeton Shape Benchmark.

lower as this data set is more challenging with many symmetric objects. As expected the baseline PCA method only works well when there is sufficient overlap. All previous methods experience a dramatic fall in accuracy once the overlap amount drops below 40 %; 4PCS performs the best out of these, but because 4PCS is essentially searching for the most consistent area shared by two shapes, for small overlap ratio, it can converge to false alignments (Fig. 7). Our method outperforms all previous approaches, and doesn't experience degraded performance until overlap falls below 15 %. The average performance of different algorithms is summarized in Table 1.

Table 1. Average success percentage of global registration algorithms on two data sets. Average running time is measured using a single thread on an Intel Core i7-4710MQ CPU clocked at 2.5 GHz.

	PCA	GO-ICP	FPFH	4PCS	Our method
Stanford (%)	19.5	34.1	49.3	73.0	93.6
Princeton (%)	18.5	22.0	33.0	73.2	81.5
Runtime (sec)	0.01	25	3	10	0.5

Performance on Real Data. We further compare the performance of our registration method with 4PCS on pairs of depth maps captured from Kinect One and Structure IO sensors. The hardware setup used to obtain this data is described in detail in the next section. These depth maps share only 10 %–30 % overlap and 4PCS often fails to compute the correct alignment as shown in Fig. 8.

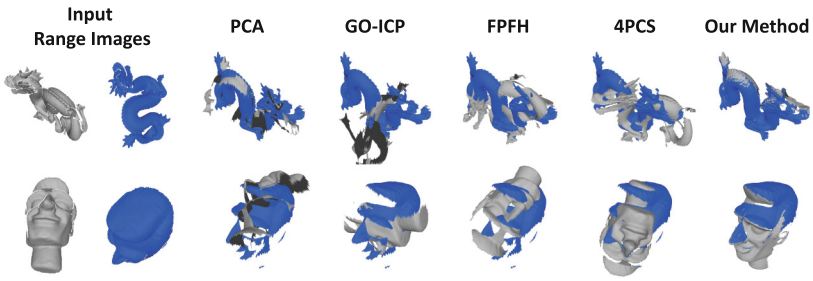


Fig. 7. Example registration results of range images with limited overlap. First and second row show examples from the Stanford 3D Scanning Repository and the Princeton Shape Benchmark respectively. Please see the supplementary material for more examples.

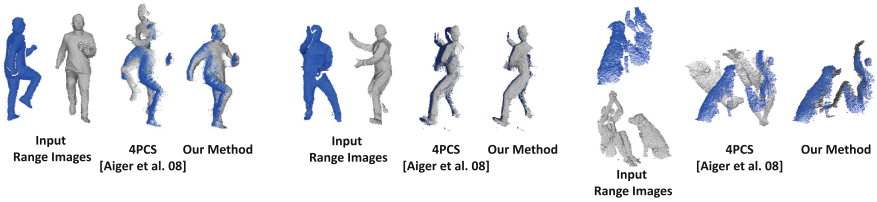


Fig. 8. Our registration method compared with 4PCS on real data. First two examples are captured by Kinect One sensors while the last example is captured by Structure IO sensors.

Limitations. Our global registration method works best when there is sufficient visibility information in the underlying range images, *i.e.*, when the depth sensor's field of view contains the entire object and the background is removed. It tends to fail when the visibility information does not prevail, *e.g.*, range scans of indoor scenes depicting large planar surfaces. We plan to extend our method to handle those challenging cases in future work.

6 Dynamic Capture Results

Hardware. We experiment with two popular depth sensors, namely the Kinect One (V2) sensor and the Structure IO sensor. We mount the former on laptops and extend the capture range with long power extension cables. For the latter, we attach it to iPad Air 2 tablets and stream data to laptops through wireless network. Kinect One sensors stream high-fidelity 512×424 depth images and 1920×1080 color images at 30 fps. We use it to cover the entire human body from 3 or 4 views at approximately 2 m away. Structure IO sensors stream 640×480 for both depth and color (iPad RGB camera after compression) images at 30 fps. Per pixel depth accuracy of the Structure IO sensor is relatively low and unreliable, especially when used outdoor beyond 2 m. Thus, we use it to

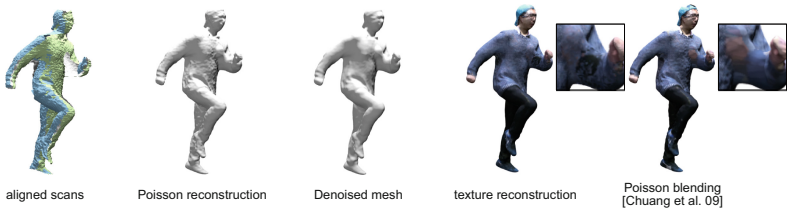


Fig. 9. From left to right: Globally aligned partial scans from multiple depth sensors; The water-tight mesh model after Poisson reconstruction [20]; Denoised mesh after merging neighboring meshes by using [23]; Model after our dense correspondences based texture reconstruction; Model after directly applying texture-stitcher [9].

capture small objects, *e.g.*, dogs and children, at approximately 1 m away. Our mobile capture setting allows the subject to move freely in space in stead of being restricted to a specific capture volume.

Pre-processing. For depth images, first we remove background by thresholding depth value and removing dominant planar segments in a RANSAC fashion. For temporal synchronization across sensors, we use visual cues, *i.e.*, jumping, to manually initialize the starting frame. Then we automatically synchronize all remaining frames by using the system time stamps, which are accurate up to milliseconds.

Performance. We process data using a single thread Intel Core i7-4710MQ CPU clocked at 2.5 GHz. It takes on average 15 s to globally align all the views for each frame, 5 min for surface denoising and reconstruction, and 3 min for building dense correspondences and texture reconstruction.

Results. We capture a variety of motions and objects, including walking, jumping, playing Tai Chi and dog training (see the supplementary material for a complete list). For all captures, the performer(s) are able to move freely in space while 3 or 4 people follow them with depth sensors. As shown in Fig. 9, our geometry reconstruction method reduces flickering artifacts of the original Poisson reconstruction, and our texture reconstruction method recovers reliable texture on occluded areas. Figure 10 provides several examples that demonstrate the effectiveness and flexibility of our capture system. Our global registration method plays a key role as most range images share only 10 % to 30 % overlap. While we demonstrate successful sequences with 3 depth sensors, an additional sensor typically improves the reconstruction quality since it provides higher overlap between neighboring views leading to a more robust registration.

As opposed to most existing free-form surface reconstruction techniques, our method can handle performances of subjects that move through a long trajectory instead of being constrained to a capture volume. Since our method does not require a template, it is not restricted to human performances and can successfully capture animals for which obtaining a static template would be challenging. The global registration method employed for each frame effectively reduces drift for long capture sequences. We can recover plausible textures even in occluded regions.

Fig. 10. Example capturing results. The sequence in the lower right corner is reconstructed from Structure IO sensors, while other sequences are reconstructed from Kinect One Sensors.

7 Conclusion

We have demonstrated that it is possible, using only a small number of synchronized consumer-grade handheld sensors, to reconstruct fully-textured moving humans, and without restricting the subject to the constrained environment required by stage setups with calibrated sensor arrays. Our system does not require a template geometry in advance and thus can generalize well to a variety of subjects including animals and small children. Since our system is based on low-cost devices and works in fully unconstrained environments, we believe our system is an important step toward accessible creation of VR and AR content for consumers. Our results depend critically on our new alignment algorithm based on the visibility error metric, which can reliably align partial scans with much less overlap than is required by current state-of-the-art registration algorithms. Without this alignment algorithm, we would need to use many more sensors, and solve the sensor interference problem that would arise. We believe this algorithm is an important contribution on its own, as a significant step forward in global registration.

Acknowledgments. We thank Jieqi Jiang, Xiang Ao, Jin Xu, Mingfai Wong, Bor-Jeng Chen and Anh Tran for being our capture models. This research is supported in part by Adobe, Oculus & Facebook, Sony, Pelican Imaging, Panasonic, Embodee, Huawei, the Google Faculty Research Award, The Okawa Foundation Research Grant, the Office of Naval Research (ONR)/U.S. Navy, under award number N00014-15-1-2639, the Office of the Director of National Intelligence (ODNI), and Intelligence Advanced Research Projects Activity (IARPA), under contract number 2014-14071600010. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

1. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: ACM SIGGRAPH, pp. 98:1–98:10. ACM, New York (2008)
2. Ahmed, N., Theobalt, C., Dobrev, P., Seidel, H.P., Thrun, S.: Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry. In: IEEE CVPR, pp. 1–8, June 2008
3. Aiger, D., Mitra, N.J., Cohen-Or, D.: 4-points congruent sets for robust pairwise surface registration. In: ACM Transactions on Graphics (TOG), vol. 27, p. 85. ACM (2008)
4. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. ACM Trans. Graph. **24**(3), 408–416 (2005)
5. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular RGB-D sequences, pp. 2300–2308, December 2015
6. Bojsen-Hansen, M., Li, H., Wojtan, C.: Tracking surfaces with evolving topology. ACM Trans. Graph. (SIGGRAPH 2012) **31**(4), 53:1–53:10 (2012)
7. Chen, C.S., Hung, Y.P., Cheng, J.B.: Ransac-based darces: a new approach to fast automatic registration of partially overlapping range images. IEEE Trans. Pattern Anal. Mach. Intell. **21**(11), 1229–1234 (1999)
8. Chen, Y., Medioni, G.: Object modeling by registration of multiple range images. In: ICRA, pp. 2724–2729. IEEE (1991)
9. Chuang, M., Luo, L., Brown, B.J., Rusinkiewicz, S., Kazhdan, M.: Estimating the laplace-beltrami operator by restricting 3D functions. In: Computer Graphics Forum, vol. 28, pp. 1475–1484. Wiley Online Library (2009)
10. Chung, D.H., Yun, I.D., Lee, S.U.: Registration of multiple-range views using the reverse-calibration technique. Pattern Recogn. **31**(4), 457–464 (1998)
11. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. In: ACM SIGGRAPH, vol. 34, pp. 69:1–69:13. ACM, July 2015
12. Debevec, P.: The light stages and their applications to photoreal digital actors. In: SIGGRAPH Asia, Singapore, November 2012
13. Dou, M., Taylor, J., Fuchs, H., Fitzgibbon, A., Izadi, S.: 3D scanning deformable objects with a single rgbd sensor. In: IEEE CVPR, pp. 493–501, June 2015
14. Franco, J., Lapierre, M., Boyer, E.: Visual shapes of silhouette sets. In: Third International Symposium on 3D Data Processing, Visualization, and Transmission, pp. 397–404, June 2006
15. Gelfand, N., Mitra, N.J., Guibas, L.J., Pottmann, H.: Robust global registration. In: Symposium on Geometry Processing, vol. 2, p. 5 (2005)
16. Hernández, C., Schmitt, F., Cipolla, R.: Silhouette coherence for camera calibration under circular motion. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 343–349 (2007)
17. Horn, B.K.: Extended gaussian images. Proc. IEEE **72**(12), 1671–1686 (1984)
18. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In: UIST, pp. 559–568. ACM, New York (2011)
19. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Anal. Mach. Intell. **21**(5), 433–449 (1999)

20. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing, vol. 7 (2006)
21. Kennedy, J.: Particle swarm optimization. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, pp. 760–766. Springer, New York (2010)
22. Li, H., Adams, B., Guibas, L.J., Pauly, M.: Robust single-view geometry and motion reconstruction. In: ACM SIGGRAPH Asia, SIGGRAPH Asia 2009, pp. 175:1–175:10. ACM, New York (2009)
23. Li, H., Luo, L., Vlasic, D., Peers, P., Popović, J., Pauly, M., Rusinkiewicz, S.: Temporally coherent completion of dynamic shapes. ACM TOG **31**(1), 2:1–2:11 (2012)
24. Li, H., Vouga, E., Gudym, A., Luo, L., Barron, J.T., Gusev, G.: 3D self-portraits. In: ACM SIGGRAPH Asia, vol. 32, pp. 187:1–187:9. ACM, November 2013
25. Li, H., Yu, J., Ye, Y., Bregler, C.: Realtime facial animation with on-the-fly correctives. In: ACM SIGGRAPH, vol. 32, pp. 42:1–42:10. ACM, July 2013
26. Liu, Y., Ye, G., Wang, Y., Dai, Q., Theobalt, C.: Human performance capture using multiple handheld kinects. In: Shao, L., Han, J., Kohli, P., Zhang, Z. (eds.) Computer Vision and Machine Learning with RGB-D Sensors, pp. 91–108. Springer International Publishing, Cham (2014)
27. Makadia, A., Patterson, A., Daniilidis, K.: Fully automatic registration of 3D point clouds. In: 2006 IEEE Conference on CVPR, vol. 1, pp. 1297–1304. IEEE (2006)
28. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, pp. 369–374. ACM Press/Addison-Wesley Publishing Co., New York (2000)
29. Mellado, N., Aiger, D., Mitra, N.J.: Super 4pcs fast global pointcloud registration via smart indexing. In: Computer Graphics Forum, vol. 33, pp. 205–215. Wiley Online Library (2014)
30. Moezzi, S., Tai, L.C., Gerard, P.: Virtual view generation for 3D digital video. MultiMedia, IEEE **4**(1), 18–26 (1997)
31. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: reconstruction and tracking of non-rigid scenes in real-time. In: IEEE CVPR, June 2015
32. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: IEEE CVPR, pp. 1862–1869. IEEE (2012)
33. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: IEEE CVPR, pp. 1106–1113. IEEE (2014)
34. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: 3-D Digital Imaging and Modeling, pp. 145–152. IEEE (2001)
35. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (fpfh) for 3D registration. In: 2009 IEEE International Conference on Robotics and Automation, pp. 3212–3217. IEEE (2009)
36. Rusu, R.B., Blodow, N., Marton, Z.C., Beetz, M.: Aligning point cloud views using persistent feature histograms. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3384–3391. IEEE (2008)
37. Rusu, R.B., Cousins, S.: 3D is here: point cloud library (pcl). In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–4. IEEE (2011)
38. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Shape modeling applications, 2004. Proceedings, pp. 167–178. IEEE (2004)
39. Shoemake, K.: Uniform random rotations. In: Graphics Gems III, pp. 124–132. Academic Press Professional, Inc. (1992)

40. Silva, L., Bellon, O.R., Boyer, K.L.: Precision range image registration using a robust surface interpenetration measure and enhanced genetic algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 762–776 (2005)
41. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Comput. Graph. Appl. **27**(3), 21–31 (2007)
42. Süßmuth, J., Winter, M., Greiner, G.: Reconstructing animated meshes from time-varying point clouds. In: SGP, SGP 2008, pp. 1469–1476 (2008)
43. Tevs, A., Berner, A., Wand, M., Ihrke, I., Bokeloh, M., Kerber, J., Seidel, H.P.: Animation cartography—intrinsic reconstruction of shape and motion. ACM TOG **31**(2), 12:1–12:15 (2012)
44. Tong, J., Zhou, J., Liu, L., Pan, Z., Yan, H.: Scanning 3D full human bodies using kinects. IEEE TVCG **18**(4), 643–650 (2012)
45. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM SIGGRAPH, SIGGRAPH 2008, pp. 97:1–97:9. ACM, New York (2008)
46. Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., Matusik, W.: Dynamic shape capture using multi-view photometric stereo. In: ACM SIGGRAPH Asia, SIGGRAPH Asia 2009. pp. 174:1–174:11 (2009)
47. Wand, M., Adams, B., Ovsjanikov, M., Berner, A., Bokeloh, M., Jenke, P., Guibas, L., Seidel, H.P., Schilling, A.: Efficient reconstruction of nonrigid shape and motion from real-time 3D scanner data. ACM TOG **28**(2), 15:1–15:15 (2009)
48. Wand, M., Jenke, P., Huang, Q., Bokeloh, M., Guibas, L., Schilling, A.: Reconstruction of deforming geometry from time-varying point clouds. In: SGP, SGP 2007, pp. 49–58 (2007)
49. Wang, R., Choi, J., Medioni, G.: 3D modeling from wide baseline range scans using contour coherence. In: 2014 IEEE Conference on CVPR, pp. 4018–4025 (2014)
50. Wei, L., Huang, Q., Ceylan, D., Vouga, E., Li, H.: Dense human body correspondences using convolutional networks. In: IEEE CVPR. IEEE (2016)
51. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera. ACM Trans. Graph. **32**(6), 161:1–161:11 (2013)
52. Wu, C., Varanasi, K., Liu, Y., Seidel, H.P., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination, pp. 1108–1115. IEEE, November 2011
53. Yang, J., Li, H., Jia, Y.: Go-icp: Solving 3D registration efficiently and globally optimally. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1457–1464. IEEE (2013)
54. Ye, G., Deng, Y., Hasler, N., Ji, X., Dai, Q., Theobalt, C.: Free-viewpoint video of human actors using multiple handheld kinects. IEEE Trans. Cybern. **43**(5), 1370–1382 (2013)
55. Zhang, Z.: Iterative point matching for registration of free-form curves and surfaces. IJCV **13**(2), 119–152 (1994)
56. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., Stamminger, M.: Real-time non-rigid reconstruction using an rgb-d camera. In: ACM SIGGRAPH, vol. 33, pp. 156:1–156:12. ACM, New York, July 2014