3D Modeling from Wide Baseline Range Scans using Contour Coherence

Ruizhe Wang Jongmoo Choi Gérard Medioni
Computer Vision Lab, Institute for Robotics and Intelligent Systems
University of Southern California
{ruizhewa, jongmooc, medioni}@usc.edu

Abstract

Registering 2 or more range scans is a fundamental problem, with application to 3D modeling. While this problem is well addressed by existing techniques such as ICP when the views overlap significantly at a good initialization, no satisfactory solution exists for wide baseline registration. We propose here a novel approach which leverages contour coherence and allows us to align two wide baseline range scans with limited overlap from a poor initialization. Inspired by ICP, we maximize the contour coherence by building robust corresponding pairs on apparent contours and minimizing their distances in an iterative fashion. We use the contour coherence under a multi-view rigid registration framework, and this enables the reconstruction of accurate and complete 3D models from as few as 4 frames. We further extend it to handle articulations, and this allows us to model articulated objects such as human body. Experimental results on both synthetic and real data demonstrate the effectiveness and robustness of our contour coherence based registration approach to wide baseline range scans, and to 3D modeling.

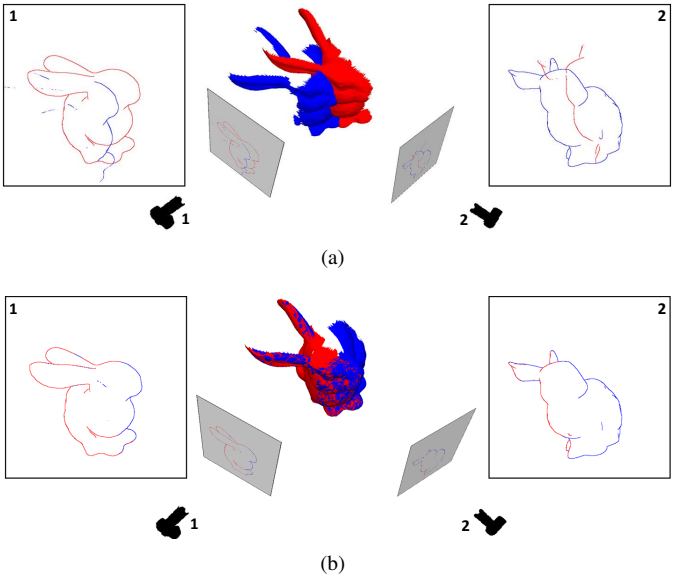


Figure 1. (a) Two roughly aligned wide baseline 2.5D range scans of the Stanford Bunny with the observed and predicted apparent contours extracted. The two meshed points cloud are generated from the two 2.5D range scans respectively (b) Registration result after maximizing the *contour coherence*

1. Introduction

Registering 2 or more range scans is a fundamental problem with application to 3D modeling. It is well addressed in the presence of sufficient overlap and good initialization [3, 4, 5, 14]. However, registering two wide baseline range scans presents a challenging task where two range scans barely overlap and the *shape coherence* no longer prevails. An example of two wide baseline range scans of the Stanford Bunny with approximately 40% overlap is given in Fig. 1(a). The traditional *shape coherence* based methods may fail as most closest-distance correspondences are incorrect.

In computer vision dealing with intensity images, a large body of work has been devoted to study the *apparent contour*, or simply *contour*. An apparent contour is the projection of a contour generator, which is defined as the set of points on the surface where the tangent plane contains the line of sight from the camera. This contour has been shown to be a rich source of geometric information for motion estimation and 3D reconstruction [7, 8, 9, 12].

Inspired by their works, we propose the concept of *contour coherence* for wide baseline range scan registration. *Contour coherence* is defined as the agreement between the observed apparent contour and the predicted apparent contour. As shown in Fig. 1(a), the observed contours extracted from the original 2.5D range scans, *i.e.*, red lines in image 1 and blue lines in image 2, do not match the corresponding contours extracted from the projected 2.5D range scans, *i.e.*, blue lines in image 1 and red lines in image 2. We maximize *contour coherence* by iteratively building robust correspondences among apparent contours and minimizing their distances. The registration result is shown in Fig. 1(b) with the

contour coherence maximized and two wide baseline range scans well aligned. The *contour coherence* is robust in the presence of wide baseline in the sense that only the shape area close to the predicted contour generator is considered when building correspondences on the contour, thus avoiding the search for correspondences over the entire shape.

The recently released low-cost structured light 3D sensors (*e.g*., Kinect, Primesense) enable handy complete scanning of objects in the home environment. While the state-of-the-art scanning methods [17, 23] generate excellent 3D reconstruction results of rigid or articulated objects, they assume small motion between consecutive views and as a result either the user has to move the 3D sensor carefully around the object or the subject must turn slowly in a controlled manner. Even with best effort to reduce the drifting error, sometimes the gap is still visible when closing the loop (Section 7).

In our work, we explicitly employ *contour coherence* under a multi-view framework and develop the Multi-view Iterative Closest Contour (M-ICC) algorithm which rigidly aligns all range scans at the same time. We further extend M-ICC to handle small articulations and propose the Multi-view Articulated Iterative Closest Contour (MA-ICC) algorithm. Using our proposed registration methods, we successfully address the loop-closure problem from as few as 4 wide baseline views with limited overlap, and reconstruct accurate and complete rigid as well as articulated objects, hence greatly reducing the data acquisition time and the registration complexity, while providing accurate results.

To the best of our knowledge, we are first to introduce *contour coherence* for multi-view wide baseline range scan registration. Our main contributions include: 1) The concept of *contour coherence* for robust wide baseline range scan registration; 2) *Contour coherence* based multi-view rigid registration algorithm M-ICC which allows 3D modeling from as few as 4 views; 3) Extension to multi-view articulated registration algorithm MA-ICC and its application to 3D body modeling.

Section 2 presents the relevant literature. Section 3 describes how to extract robust correspondences from the observed and predicted contours. Section 4 employs the *contour coherence* in a multi-view rigid registration framework, which is further extended to consider articulation in Section 5. Section 6 briefly covers implementation. Section 7 presents the experimental evaluation results while section 8 ends with a conclusion and future work.

2. Related Work

We roughly classify all relevant work into 4 categories: rigid registration, articulated non-rigid registration (assuming an articulated structure), general non-rigid registration, and apparent contour or silhouette for pose estimation.

Rigid registration. The classic Iterative Closest Point (ICP) algorithm [5] and its variants [6, 19] prove to be effective in accurate rigid registration for high-quality 3D indoor scene reconstruction [17]. Few efficient attempts have been made for solving rigid range scan registration of partial overlap and poor initialization. Silva *et al*. [22] solve the alignment by an expensive search over the entire pose-space which is speeded up by the genetic algorithm. Gelfand *et al*. [11] propose a shape descriptor based method and perform the registration by matching descriptors under the RANSAC framework. Similar approaches have been proposed [1, 2]. These metric-based registration methods requires extensive computation of similarity values across a set of candidate matches and they are only pairwise. Furthermore, they fail when two range scans overlap on featureless region.

Articulated non-rigid registration. Allen *et al*. [3] extend rigid registration to a template-based articulated registration algorithm for aligning several body scans. Their method utilizes a set of manually selected markers. Pekelney and Gotsman [18] achieve articulated registration by performing ICP on each segment and their method requires a manual segmentation of the first frame. Chang and Zwicker [4] further remove the assumption of known segmentation by solving a discrete labeling problem to detect the set of optimal correspondences and apply graph cuts for optimization.

General non-rigid registration. Li *et al*. [14] develop a registration framework that simultaneously solves for point correspondences, surface deformation, and region of overlap within a single global optimization. More recently, several methods based on the embedded deformation model have been proposed for modeling of non-rigid objects, either by initializing a complete deformation model with the first frame [23] or incrementally updating it [27].

Apparent contour or silhouette for pose estimation. Apparent contour has been used for camera motion estimation [8] relying on the notion of epipolar tangency points. In particular, the work of [26] uses only the two outermost epipolar tangents for the camera pose estimation. More recently, Hernandez *et al*. [12] propose the notion of *silhouette coherence*, which measures the consistency between observed silhouettes and predicted silhouettes, and maximize it to estimate the camera poses. Cheung *et al*. [7] calculate camera poses by locating colored surface points along the bounding edges.

3. Contour Coherence

We perform wide baseline range scan registration by maximizing *contour coherence*, *i.e*., the agreement between the observed and predicted apparent contours. From an implementation point of view, M-ICC and MA-ICC alternate between finding closest contour correspondences and minimizing their distances. However an intuitive closest matching scheme on all contour points fails, mainly due to the

Figure 2. General pipeline of our Robust Closest Contour (RCC) method

presence of self-occlusion, 2D ambiguity, and outliers as described later. Hence we propose the Robust Closest Contour (RCC) algorithm for establishing robust contour correspondences on a pair of range scans (Fig. 2).

Preliminaries. A 2.5D range scan $\mathcal{R}_i$ of frame i provides depth value $\mathcal{R}_i(\mathbf{u})$ at each image pixel $\mathbf{u} = (u, v)^T \in \mathbb{R}^2$. We use a single constant camera calibration matrix K which transforms points from the camera frame to the image plane. We represent $\mathcal{V}_i(\mathbf{u}) = K^{-1}\mathcal{R}_i(\mathbf{u})\tilde{\mathbf{u}}$ as the back-projection operator which maps $\mathbf{u}$ in frame i to its 3D location, where $\tilde{\mathbf{u}}$ denotes the homogeneous vector $\tilde{\mathbf{u}} = [\mathbf{u}^T|1]^T$. Inversely, we denote the projection operator as $\mathcal{P}(\mathcal{V}_i(\mathbf{u})) = g(K\mathcal{V}_i(\mathbf{u}))$ where g represents dehomogenisation.

A meshed points cloud $\mathbf{P}_i$ is generated for each frame i considering the connectivity on the 2.5D range scan $\mathcal{R}_i$. We calculate the normalized 3D normal at each pixel $\mathcal{N}_i(\mathbf{u}) \in \mathbb{R}^3$ following [17]. $\mathcal{N}_i(\mathbf{u})$ is further projected back to the image to obtain normalized 2D normal $\mathbf{n}_i(\mathbf{u})$ of each image pixel. Projecting $\mathbf{P}_j$ to the ith image, given current camera poses, leads us to a projected range scan $\mathcal{R}_{j\to i}$. The inputs to our RCC method are observed and predicted range scans, namely $\mathcal{R}_i$ and $\mathcal{R}_{j\to i}$, and the output is the robust contour correspondences $\mathcal{M}_{i,j\to i}$ (Eq. 5 and Eq. 9).

Extracting contour points. Given pixels belonging to the object in frame i as $\mathcal{U}_i$, we set $\mathcal{R}_i(\mathbf{u}) = \infty$ for $\mathbf{u} \notin \mathcal{U}_i$. The contour points $\mathcal{C}_i$ are extracted considering the depth discontinuity of a pixel and its 8-neighboring pixels $\mathcal{N}^8_{\mathbf{u}}$,

$$\mathcal{C}_i = \{\mathbf{u} \in \mathcal{U}_i | \exists \mathbf{v} \in \mathcal{N}^8_{\mathbf{u}},\ \mathcal{R}_i(\mathbf{v}) - \mathcal{R}_i(\mathbf{u}) > \zeta\}, \quad (1)$$

where ζ is the threshold to detect depth discontinuity. We also extract a set of occlusion points,

$$\mathcal{O}_i = \{\mathbf{u} \in \mathcal{U}_i | \exists \mathbf{v} \in \mathcal{N}^8_{\mathbf{u}},\ \mathcal{R}_i(\mathbf{u}) - \mathcal{R}_i(\mathbf{v}) > \zeta\}, \quad (2)$$

which are boundary points of surface holes created by self-occlusion. An example of $\mathcal{C}_i$ and $\mathcal{C}_{j\to i}$, extracted from $\mathcal{R}_i$ and $\mathcal{R}_{j\to i}$ respectively, is demonstrated in Fig. 3(b).

Pruning contour points. Both $\mathcal{C}_i$ and $\mathcal{C}_{j\to i}$ must be pruned before the matching stage to avoid possible incorrect correspondences. First due to the self-occlusion of frame j, $\mathcal{C}_{j\to i}$ contains false contour points which are actually generated by the meshes in $\mathbf{P}_j$ connected with $\mathcal{C}_j$ and $\mathcal{O}_j$. We mark and remove them to generate the pruned contour points $\mathcal{C}^p_{j\to i}$. Second again due to the self-occlusion of frame j, some contour points in $\mathcal{C}_i$ should not be matched with any contour point in $\mathcal{C}^p_{j\to i}$, *e.g.*, the contour points in frame 2 belonging to the back part of the Armadillo are not visible in view 1 (Fig. 3(b)). Hence we prune $\mathcal{C}_i$ based on the visibility of the corresponding contour generator in view j,

$$\mathcal{C}^p_{i/j} = \{\mathbf{u} \in \mathcal{C}_i | \mathcal{N}_i(\mathbf{u})^T \cdot (\mathbf{o}_{j\to i} - \mathcal{V}_i(\mathbf{u})) > 0\}, \quad (3)$$

where $\mathbf{o}_{j\to i}$ is the camera location of frame j in camera i. An example of pruned contour points is shown in Fig. 3(c).

Bijective closest matching in 3D. After pruning, a one-way closest matching algorithm between $\mathcal{C}^p_{i/j}$ and $\mathcal{C}^p_{j\to i}$ still fails, as contour points are sensitive to minor changes in viewing directions, *e.g.*, camera 1 observes only one leg while the contour points of two legs are extracted from the projected range scan (Fig. 3(c)). Hence we follow a bijective matching scheme [27] when establishing robust correspondences (Eq. 5 and Eq. 9).

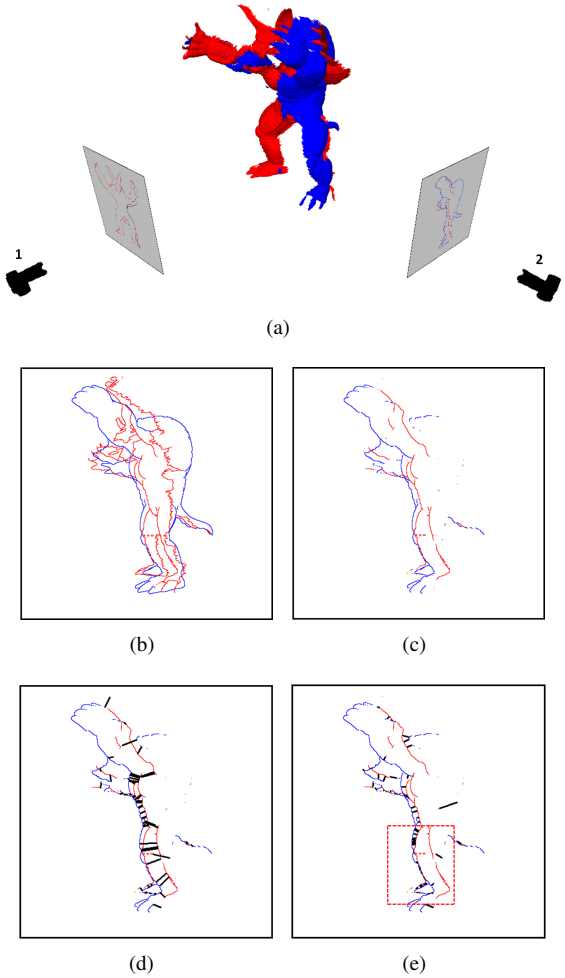


Figure 3. (a) Front range scan (red) and side range scan (blue) of the Stanford Armadillo. (b) $\mathcal{C}_{1\to 2}$ (red) and $\mathcal{C}_2$ (blue). (c) $\mathcal{C}^p_{1\to 2}$ (red) and $\mathcal{C}^p_{2/1}$ (blue). (d) Bijective correspondences (black line) found in 3D. (e) Bijective correspondences (black line) found in 2D with red rectangle indicating mismatches

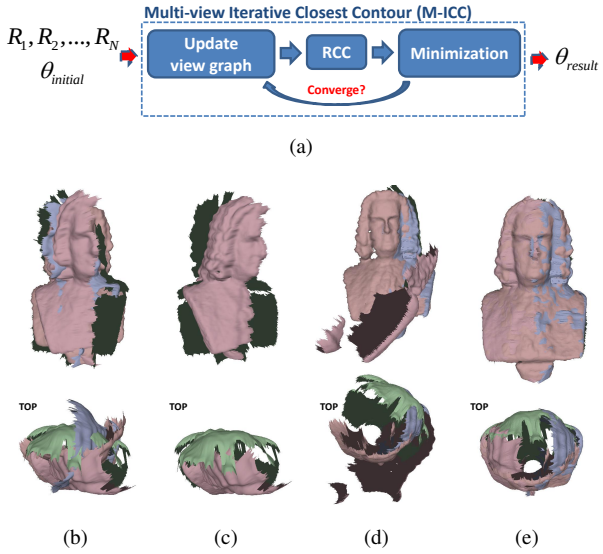


Figure 4. (a) Our M-ICC method. (b) 4 range scans of a Bach head statue taken approximately at 90^o apart, with limited overlap and poor initialization. (c) Initial right range scan and back range scan barely overlap and have a large rotation error. (d) Result of pairwise registration using standard ICP algorithm. (e) Result of our M-ICC method.

Matching directly in the 2D image space leads to many wrong corresponding pairs. An example is shown in Fig. 3(e), where the contour points of the right leg in frame 1 are wrongly matched with the contour points of the left leg in frame 2. The ambiguity imposed by the 2D nature is resolved by relaxing the search to the 3D space (Fig. 3(d)), as we have the 3D point location $\mathcal{V}_i(\mathbf{u})$ for each contour point. It is worth mentioning that while we build correspondences in 3D, we are minimizing the distances between contour correspondences in 2D, as the real data given by most structured-light 3D sensors is extremely noisy along the rays of apparent contour.

4. Multi-View Rigid Registration

The general pipeline of our Multi-view Iterative Closest Contour (M-ICC) method is shown in Fig. 4(a). Given N roughly initialized range scans (Fig. 4(b)), we alternate between updating the view graph, establishing robust contour correspondences from pairs of range scans in the view graph and minimizing distances of all correspondences. While the standard pairwise ICP algorithm fails in the presence of wide baseline (Fig. 4(d)), our M-ICC method jointly recovers accurate camera poses (Fig. 4(e)).

Preliminaries. Frame i is associated with a 6 DOF rigid transformation matrix ${}^wT_i = \begin{bmatrix} R_i & \mathbf{t}_i \\ \mathbf{0}^T & 1 \end{bmatrix}$ where R_i is parameterized by a 3 DOF quaternion, namely $\mathbf{q}_i = [q_i^w, q_i^x, q_i^y, q_i^z]$ with $\|q\|_2 = 1$, and $\mathbf{t}_i$ is the translation vector. Operator ${}^w\Pi_i(\mathbf{u}) = {}^wT_i\tilde{\mathcal{V}}_i(\mathbf{u})$ transforms pixel $\mathbf{u}$ to its corresponding homogeneous back-projected 3D point in the world coordinate system, where $\tilde{\mathcal{V}}_i(\mathbf{u})$ is the homogeneous back-projected 3D point in the camera coordinate system of frame i. Inversely we have operator ${}^i\Pi_w$ such that $\mathbf{u} = {}^i\Pi_w({}^w\Pi_i(\mathbf{u})) = \mathcal{P}(g({}^iT_w\ {}^w\Pi_i(\mathbf{u})))$. Given N frames, we have a total $6 \times N$ parameters stored in a vector $\boldsymbol{\theta}$.

Unlike [23, 24] where pairwise registration is performed before a final global error diffusion step, we do not require pairwise registration and explicitly employ *contour coherence* under a multi-view framework. We achieve that by associating two camera poses with a single contour correspondence. Assuming $\mathbf{u}$ and $\mathbf{v}$ is a corresponding pair belonging to frame i and frame j respectively, then their distance is modeled as $\|\mathbf{v} - {}^j\Pi_w({}^w\Pi_i(\mathbf{u}))\|_2$. Minimizing this distance updates both camera poses at the same time, which allows us to globally align all frames together. It is worth mentioning that pairwise registration is a special case of the multi-view scenario in a way that the pairwise registration ${}^2T_1$ is achieved as ${}^2T_1 = {}^2T_w\ {}^wT_1$.

View graph. View graph $\mathcal{L}$ is a set of pairing relationship among all frames. $(i,j) \in \mathcal{L}$ indicates that frame j is viewable in frame i and hence robust contour correspondences should be established between $\mathcal{R}_i$ and $\mathcal{R}_{j\to i}$. Each frame's viewing direction in the world coordinate is $R_i(0,0,1)^T$ and frame j is viewable in frame i only if their viewing directions are within a certain angle η, *i.e.*,

$$\mathcal{L} = \{(i,j) | \ \text{acos}((0,0,1)R_i{}^T R_j(0,0,1)^T) < \eta\}. \quad (4)$$

It is worth mentioning that $(i,j) \neq (j,i)$ and we establish two pairs of correspondences between frame i and frame j, namely between $\mathcal{C}^p_{j\to i}$ and $\mathcal{C}^p_{i/j}$, and between $\mathcal{C}^p_{i\to j}$ and $\mathcal{C}^p_{j/i}$. Another issue worth raising is that the loop closure is automatically detected and achieved if all N views form a loop. An example is shown in Fig. 4(c), where $\mathcal{L}$ is calculated as $\{(1,2),(2,1),(2,3),(3,2),(1,4),(4,1)\}$ from $\boldsymbol{\theta}_{initial}$, *i.e.*, the gap between frame 3 (back frame) and frame 4 (right frame) is large and the loop is not closed from the beginning. As we iterate and update the camera poses, link $\{(3,4),(4,3)\}$ is added to $\mathcal{L}$ and we automatically close the loop.

Robust Closest Contour and Minimization. For each viewable pair $(i,j) \in \mathcal{L}$, we extract robust contour correspondences $\mathcal{M}_{i,j\to i}$ between $\mathcal{C}^p_{j\to i}$ and $\mathcal{C}^p_{i/j}$ using RCC algorithm as

$$\begin{aligned} \mathcal{M}_{i,j\to i} = \{(\mathbf{u}, {}^j\Pi_w({}^w\Pi_i(\mathbf{v}))) \ | \\ \mathbf{v} = \arg\min_{\mathbf{m}\in\mathcal{C}^p_{j\to i}} d(\mathcal{V}_i(\mathbf{u}), \mathcal{V}_{j\to i}(\mathbf{m})), \\ \mathbf{u} = \arg\min_{\mathbf{n}\in\mathcal{C}^p_{i/j}} d(\mathcal{V}_{j\to i}(\mathbf{v}), \mathcal{V}_i(\mathbf{n}))\} \end{aligned} \quad (5)$$

where $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$ is the Euclidean distance operator. Pixel $\mathbf{v}$ is the closest (*i.e.*, distance in the back-projected 3D space) point on the pruned predicted contour to pixel $\mathbf{u}$ on the pruned observed contour, while at the same time pixel $\mathbf{u}$ is also the closest to pixel $\mathbf{v}$, *i.e.*, the bijectivity in 3D is imposed.

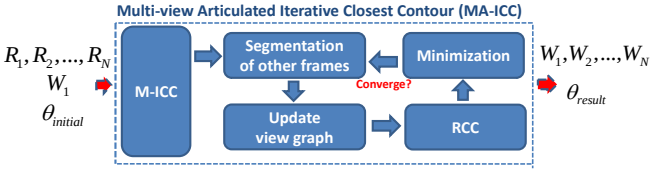
We minimize the sum of point-to-plane [5] distances of all contour correspondences as

$$\mathcal{E}_R = \sum_{(i,j)\in\mathcal{L}} \sum_{(\mathbf{u},\mathbf{v})\in\mathcal{M}_{i,j\to i}} |(\mathbf{u} - {}^{i}\Pi_w({}^{w}\Pi_j(\mathbf{v})))^T \cdot \mathbf{n}_i(\mathbf{u})|. \tag{6}$$

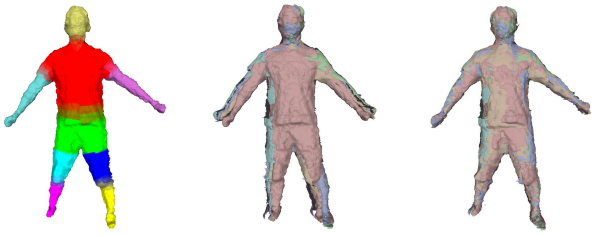
In practice, we find that the point-to-plane error metric allows two contours to slide along each other and reaches better local optimum than the point-to-point error metric.

5. Multi-view Articulated Registration

To handle small articulations, we further extend M-ICC to Multi-view Articulated Iterative Closest Contour (MA-ICC) algorithm (Fig. 5(a)). Given N range scans, articulation structure as well as known segmentation $\mathcal{W}_1$ of all rigid parts in the first frame (Fig. 5(b)), we first regard all range scans as rigid and apply the M-ICC method, after which all range scans are roughly aligned (Fig. 5(c)). We then iteratively segment other frames, update the view graph, establish robust contour correspondences and minimize until convergence (Fig. 5(d)).



(a)



(b) (c) (d)

Figure 5. (a) Our MA-ICC method. (b) Segmentation of human body into rigid parts. (c) Registration result after our M-ICC method. (d) Registration result after our MA-ICC method.

Preliminaries. We employ a standard hierarchical structure, where each rigid segment k of frame i has an attached local coordinate system related to the world coordinate system via transform ${}^{w}T_k^i$. This transformation is defined hierarchically by recurrence ${}^{w}T_k^i = {}^{w}T_{k_p}^i \; {}^{k_p}T_k^i$, where k_p is the parent node of k. For the root node, we have ${}^{w}T_{root}^i = {}^{w}T_i$, where ${}^{w}T_i$ can be regarded as camera pose of frame i. ${}^{k_p}T_k^i$ has a parameterized rotation component and a translation component completely dependent on the rotation component. As such, for a total of N range scans where the complete articulated structure contains M rigid segments, there is a total number of $N \times (M \times 3 + 3)$ parameters stored in the vector $\boldsymbol{\theta}$.

We employ the Linear Blend Skinning (LBS) scheme where each pixel $\mathbf{u}$ in frame i is given a weight vector $\mathcal{W}_i(\mathbf{u}) \in \mathbb{R}^M$ with $\sum_{j=1\ldots M} \mathcal{W}_i(\mathbf{u})_j = 1$, indicating its support from all rigid segments. As such, operator ${}^{w}\Pi_i$ in the rigid case is rewritten as ${}^{w}\Pi_i^{\mathcal{A}}(\mathbf{u}) = \sum_{j=1\ldots M} {}^{w}T_j^i \tilde{\mathcal{V}}_i(\mathbf{u})\mathcal{W}_i(\mathbf{u})_j$ in the articulated case, which is a weighted transformation of all rigid segments attached to $\mathbf{u}$. Similarly we have operator ${}^{i}\Pi_w^{\mathcal{A}}$ as the inverse process such that $\mathbf{u} = {}^{i}\Pi_w^{\mathcal{A}}({}^{w}\Pi_i^{\mathcal{A}}(\mathbf{u}))$.

Segmentation of other frames. Given the segmentation $\mathcal{W}_1$ of the first frame and predicted pose θ, we segment pixel $\mathbf{u} \in \mathcal{U}_i$ of frame i as

$$\mathcal{W}_i(\mathbf{u}) = \mathcal{W}_1(\arg\min_{\mathbf{v}\in\mathcal{U}_1} d(\mathbf{v}, {}^{1}\Pi_w^{\mathcal{A}}({}^{w}\Pi_i^{\mathcal{A}}(\mathbf{u})))), \tag{7}$$

i.e., the same weight as the closest pixel in the first frame.

To simplify the following discussion, we define $\mathcal{F}(\mathcal{S},k) = \{\mathbf{u} \in \mathcal{S} |\, \mathcal{W}_{\mathcal{S}}(\mathbf{u})_k = 1\}$ which indicates the subset of $\mathcal{S}$ with pixels exclusively belonging to the k-th rigid part.

View graph. In the presence of articulation, we only build contour correspondences on the corresponding rigid body parts, as such $(i,j,k) \in \mathcal{L}^{\mathcal{A}}$ indicates that rigid segment k of frame j is viewable in frame i and we should build robust contour correspondences among $\mathcal{F}(\mathcal{C}_{i/j}^p, k)$ and $\mathcal{F}(\mathcal{C}_{j\to i}^p, k)$. Besides considering the viewing direction of cameras, we consider self-occlusion and build contour correspondences only when there are enough contour points (*i.e.*, more than γ) belonging to the rigid segment k in both views,

$$\mathcal{L}^{\mathcal{A}} = \{(i,j,k) |\, \text{acos}((0,0,1)R_i{}^T R_j (0,0,1)^T) < \eta, \\ \#(\mathcal{F}(\mathcal{C}_{i/j}^p, k)) > \gamma,\ \#(\mathcal{F}(\mathcal{C}_{j\to i}^p, k)) > \gamma |\} \tag{8}$$

Robust closest contour and minimization. For each viewable pair $(i,j,k) \in \mathcal{L}^{\mathcal{A}}$, the set of bijective contour correspondences $\mathcal{M}_{i,j\to i,k}^{\mathcal{A}}$ between $\mathcal{F}(\mathcal{C}_{i/j}^p, k)$ and $\mathcal{F}(\mathcal{C}_{j\to i}^p, k)$ are extracted by RCC as

$$\mathcal{M}_{i,j\to i,k}^{\mathcal{A}} = \{(\mathbf{u}, {}^{j}\Pi_w^{\mathcal{A}}({}^{w}\Pi_i^{\mathcal{A}}(\mathbf{v}))) \,| \\ \mathbf{v} = \arg\min_{\mathbf{m}\in\mathcal{F}(\mathcal{C}_{j\to i}^p, k)} d(\mathcal{V}_i(\mathbf{u}), \mathcal{V}_{j\to i}(\mathbf{m})), \\ \mathbf{u} = \arg\min_{\mathbf{n}\in\mathcal{F}(\mathcal{C}_{i/j}^p, k)} d(\mathcal{V}_i(\mathbf{n}), \mathcal{V}_{j\to i}(\mathbf{v}))\}. \tag{9}$$

We minimize the sum of point-to-plane distances between all contour correspondences

$$\mathcal{E}_A = \sum_{(i,j,k)\in\mathcal{L}^{\mathcal{A}}} \sum_{(\mathbf{u},\mathbf{v})\in\mathcal{M}^{\mathcal{A}}_{i,j\to i,k}} |(\mathbf{u} - {}^{i}\Pi^{\mathcal{A}}_{w}({}^{w}\Pi^{\mathcal{A}}_{j}(\mathbf{v})))^T \cdot \mathbf{n}_i(\mathbf{u})| + \alpha\boldsymbol{\theta}^T \cdot \boldsymbol{\theta}, \quad (10)$$

where we use $\alpha\boldsymbol{\theta}^T \cdot \boldsymbol{\theta}$ as the regularization term favoring the small articulation assumption.

6. Implementation

We use a standard stopping condition for our iterative process: (1) the maximum iteration number has been achieved, or (2) the distance per contour correspondence is small, or (3) the decrease in distance per contour correspondence is small. For each iteration Eq. 6 and Eq. 10 are non-linear in parameters, as such we employ the Levenberg-Marquardt algorithm [16] as our solver. The Jacobian matrix for the Levenberg-Marquardt algorithm is calculated by the chain rule.

In all our experiments, we set the depth discontinuity threshold $\zeta = 50mm$ (Eq. 1, 2). The viewable angle threshold is $\eta = 120^o$ (Eq. 4, 8) while the rigid segment minimum number of points threshold $\gamma = 500$ (Eq. 8). The weight for the regularization term is set as $\alpha = 100$ (Eq. 10). For scanning real rigid and articulated objects, we use the Kinect sensor. It is worth mentioning that for a specific range scanning device the parameters work in a large range and do not require specific tuning.

In practice, our M-ICC and MA-ICC methods converge within 10 iterations. Specifically for pair-wise rigid registration, within each iteration, we perform two projections, extract two sets of robust contour correspondences, and minimize the cost function with the Levenberg-Marquardt algorithm. Since the projection is easily parallelized and the closest-point matching is searching over a limited number of contour points in 2D space, our algorithm can easily run in real time on a GPU.

7. Experimental Results

We evaluate our *contour coherence* based multi-view registration algorithms with two sets of experiments. First we evaluate our method on pair-wise registration between synthetic wide baseline range scans. Then we show reconstruction results of rigid and articulated objects using a low-cost Kinect device, and compare with other state-of-the-art scanning algorithms.

7.1. Pair-wise Registration of Synthetic Data

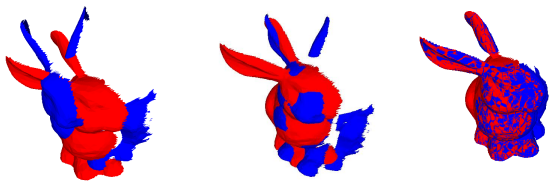
We compare our *contour coherence* method with the Trimmed-ICP (trICP) algorithm [6] which is the most robust variant of ICP in the presence of wide baseline. The basic idea of trICP is that since we don't know the exact percentage of overlap, we can manually select a set of overlap percentages in $[0, 1]$ and run the basic ICP algorithm at each overlap level by forcing the ICP to only use the predefined percentage of closest correpondences. A robust measure of $\psi_{overlap}(\xi) = e(\xi)\xi^{-(1+\lambda)}$ is introduced for the selection of optimal overlap percentage where ξ is the amount of overlap, $e(\xi)$ is the final RMSE at the predefined overlap level and λ is a preset parameter. For our experiment, we use the standard point-to-plane ICP , test a range of overlap percentages ξ from 10% to 100% increased by 10% and set $\lambda = 2$ as suggested in the original paper.

We generate pairs of synthetic wide baseline range scans of different 3D objects by moving a virtual camera around them. When aligning two synthetic range scans, an increasing offset θ_{offset} is added to the main rotation angle while small random perturbations $(0 \sim 10^o)$ are added to its other rotation angles. This setup simulates the 3D reconstruction scenario where we are trying to align two wide baseline range scans yet only have a rough approximation of the main rotation angle between them. We initialize the two range scans by aligning their centers together (Fig. 6(a)). We register the two range scans, estimate the preset offset in the main rotation angle and compare with the ground truth to obtain the error in estimation. Table 1 summarizes the alignment result of two Stanford Bunny range scans with a 45% overlap on the red range scan. Small error indicates successful registration and we denote as error < 1.

The triICP algorithm fails to align these two range scans beyond a 30^o main rotation offset and stops at a local minimum (Fig. 6(b)), while our method successfully recovers up to 54^o as shown in Fig. 6(c). Experiments on wide baseline range scans of other 3D objects, including the Stanford Armadillo and the Stanford Dragon, produce similar results

Table 1. Errors of recovered main rotation angle at different θ_{offset} (in degrees)

θ_b	12	24	30	36	42	48	54	60
trICP	< 1	< 1	< 1	44.1	44.2	44	44.2	44.1
Ours	< 1	< 1	< 1	< 1	< 1	< 1	< 1	34



(a) (b) (c)

Figure 6. (a) Initialization of two synthetic Stanford Bunny range scans of 45% overlap and a 54^o offset in main rotation angle (b) Registration results of triICP algorithm (c) Registration result of our method

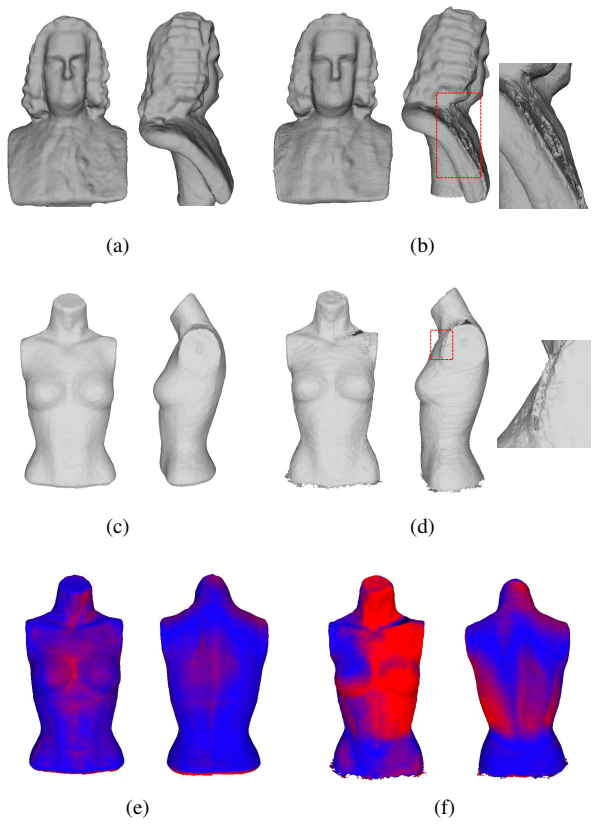


Figure 7. (a) Our modeling result of the Bach head statue from 4 range scans. (b) Modeling result of the KinectFusion algorithm from approximately 1200 range scans. (c) Our modeling result of the mannequin torso from 4 range scans. (d) Modeling result of the KinectFusion algorithm from approximately 1200 range scans. (e) Heatmap of our modeling result compared with a ground truth laser scan with error range from 0mm (blue) to 10mm (red). (f) Heatmap of the KinectFusion modeling result compared with a ground truth laser scan with the same range.

and are not listed here.

7.2. Scanning Real Objects

Rigid Objects. When modeling rigid objects, we capture four depth images of rigid objects at approximately 90^o apart using a single Kinect. The background depth pixels are removed from the range scans by simply thresholding depth, detecting and removing planar pixels using RANSAC [10], assuming the object is the only thing on the ground or table. The four wide baseline range scans are initialized by aligning the centers together and assuming a pairwise 90^o rotation angle (Fig. 4(b)). We register four frames using our proposed M-ICC method and use the Poisson Surface Reconstruction (PSR) algorithm [13] to generate the final water-tight model (Fig. 7(a)). We compare our method with the KinectFusion algorithm [17] (Fig. 7). In practice, KinectFusion easily loses track of the camera even when we carefully hold the camera and move slowly. As a result, we put the object on a turntable, rotate it for approximately 2 minutes and scan it with a fixed Kinect.

As KinectFusion incrementally updates the model, the drifting error accumulates and creates artifacts when closing the loop (Fig. 7(b) 7(d)), while our method successfully reconstruct smooth 3D objects (Fig. 7(a) 7(c)) from only four views. We further compare our reconstructed model with a laser scanned ground truth model and achieve a median error of $2.12mm$ (Fig. 7(e)), while the KinectFusion achieves a median error of $5.17mm$ (Fig. 7(f)). The heatmaps clearly show that while our model accurately captures the global shape, KinectFusion suffers from an accumulated drifting error.

Articulated Objects. Among all articulated objects, human body is of most interest in 3D modeling [15, 23, 25, 27] for its potential applications in 3D printing, animation and apparel design. The subject is scanned by turning in front of a fixed Kinect sensor while showing 4 key poses, *i.e.*, front, left, back and right in order.

When scanning human body, due to Kinect's field of view limit, the subject must stand at approximately 2 meters away which leads to a large degradation in the input data quality. As such, inspired by [15], we ask the subject to come closer and stay rigid for 5 to 10 seconds at each key pose while the Kinect, controlled by a built-in motor, rotates to scan the subject using KinectFusion. The reconstructed partial 3D scene is further projected back to generate a super-resolution range scan.

After acquiring 4 super-resolution range scans of the subject, we align them using our MA-ICC method. Segmentation of the first range scan is performed by heuristically segmenting the bounding contour and then assign the same weight to each pixel as of its closest bounding contour. In practice we segment the whole body into 9 parts (Fig. 5(b)). It is worth mentioning that any other segmentation algorithm can be applied as input, such as [21]. After registration, we again use PSR to generate the final water-tight model as shown in Fig. 8. To the best of our knowledge, we are the first to generate accurate and complete human body models from as few as 4 views with a single sensor, as other *shape coherence* based methods fail in the presence of wide baseline. Our single Kinect 3D body scanning system opens the door for many domestic applications, including our recent work on animating the acquired 3D body model [20].

8. Conclusion and Future Work

We propose the concept of *contour coherence* for solving the problem of wide baseline range scan registration. Our M-ICC and MA-ICC methods allow complete reconstruction of rigid and articulated objects from as few as 4 frames. In the future, we plan to apply *contour coherence*

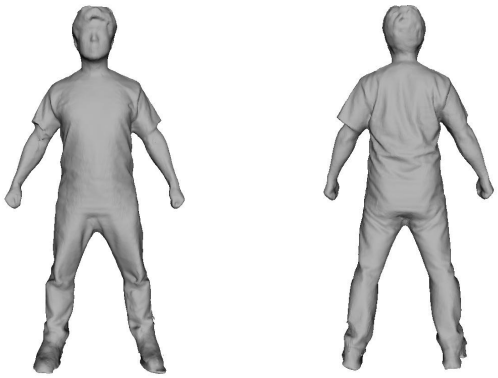


Figure 8. Front and back views of our reconstructed human body model from 4 super-resolution range scans using MA-ICC method

in other interesting fields, *e.g.*, object recognition and pose estimation from range scans.

References

[1] D. Aiger, N. J. Mitra, and D. Cohen-Or. 4-points congruent sets for robust pairwise surface registration. In *ACM Transactions on Graphics (TOG)*, volume 27, page 85. ACM, 2008. 2

[2] A. Albarelli, E. Rodola, and A. Torsello. A game-theoretic approach to fine surface registration without initial motion estimation. In *CVPR*, pages 430–437. IEEE, 2010. 2

[3] B. Allen, B. Curless, and Z. Popović. Articulated body deformation from range scan data. In *TOG*, volume 21, pages 612–619. ACM, 2002. 1, 2

[4] W. Chang and M. Zwicker. Automatic registration for articulated shapes. In *Computer Graphics Forum*, volume 27, pages 1459–1468, 2008. 1, 2

[5] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 1, 2, 5

[6] D. Chetverikov, D. Stepanov, and P. Krsek. Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm. *Image and Vision Computing*, 23(3):299–309, 2005. 2, 6

[7] G. K. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *CVPR*, volume 2, pages II–375, 2003. 1, 2

[8] R. Cipolla, K. E. Astrom, and P. J. Giblin. Motion from the frontier of curved surfaces. In *ICCV*, pages 269–275, 1995. 1, 2

[9] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *IJCV*, 9(2):83–112, 1992. 1

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 7

[11] N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann. Robust global registration. In *Symposium on geometry processing*, volume 2, page 5, 2005. 2

[12] C. Hernandez, F. Schmitt, and R. Cipolla. Silhouette coherence for camera calibration under circular motion. *TPAMI*, 29(2):343–349, 2007. 1, 2

[13] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 7

[14] H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer graphics forum*, volume 27, pages 1421–1430, 2008. 1, 2

[15] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2013)*, 32(6), November 2013. 7

[16] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978. 6

[17] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136. IEEE, 2011. 2, 3, 7

[18] Y. Pekelny and C. Gotsman. Articulated object reconstruction and markerless motion capture from depth video. In *Computer Graphics Forum*, volume 27, pages 399–408, 2008. 2

[19] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001. 2

[20] A. Shapiro, A. Feng, R. Wang, H. Li, B. Mark, G. Medioni, and E. Suma. Rapid avatar capture and simulation from commodity depth sensors. In *Computer Animation and Social Agents (CASA)*. IEEE, 2014. 7

[21] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013. 7

[22] L. Silva, O. R. P. Bellon, and K. L. Boyer. Precision range image registration using a robust surface interpenetration measure and enhanced genetic algorithms. *TPAMI*, 27(5):762–776, 2005. 2

[23] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *Visualization and Computer Graphics, IEEE Transactions on*, 18(4):643–650, 2012. 2, 4, 7

[24] A. Torsello, E. Rodola, and A. Albarelli. Multiview registration via graph diffusion of dual quaternions. In *CVPR*, pages 2441–2448. IEEE, 2011. 4

[25] R. Wang, J. Choi, and G. Medioni. Accurate full body scanning from a single fixed 3d camera. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, pages 432–439. IEEE, 2012. 7

[26] K.-Y. Wong and R. Cipolla. Structure and motion from silhouettes. In *ICCV*, volume 2, pages 217–222, 2001. 2

[27] M. Zeng, J. Zheng, X. Cheng, and X. Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *CVPR*, pages 145–152, 2013. 2, 3, 7