Deformable 3D Fusion: From Partial Dynamic 3D Observations to Complete 4D Models

Weipeng Xu[1,2] Mathieu Salzmann[2,3] Yongtian Wang[1] Yue Liu[1]
[1]Beijing Institute of Technology, China
[2]NICTA, Canberra, Australia
[3]CVLab, EPFL, Switzerland
{xuwp,wyt,liuyue}@bit.edu.cn, mathieu.salzmann@epfl.ch

Abstract

Capturing the 3D motion of dynamic, non-rigid objects has attracted significant attention in computer vision. Existing methods typically require either mostly complete 3D volumetric observations, or a shape template. In this paper, we introduce a template-less 4D reconstruction method that incrementally fuses highly-incomplete 3D observations of a deforming object, and generates a complete, temporally-coherent shape representation of the object. To this end, we design an online algorithm that alternatively registers new observations to the current model estimate and updates the model. We demonstrate the effectiveness of our approach at reconstructing non-rigidly moving objects from highly-incomplete measurements on both sequences of partial 3D point clouds and Kinect videos.

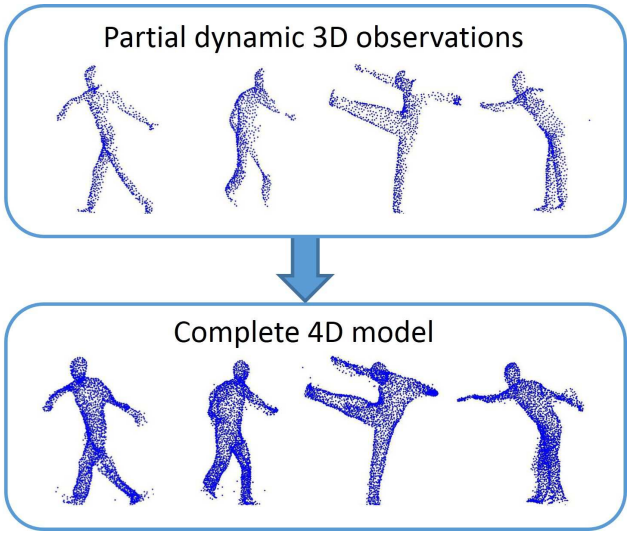

Figure 1. Deformable 3D fusion.

1. Introduction

In this paper, we introduce an approach to estimating a temporally-coherent 3D model of a non-rigid object given a dynamic sequence of highly-incomplete 3D observations of the object undergoing large deformations. Capturing the 3D motion of dynamic objects, or 4D reconstruction, has been a longstanding goal of computer vision. Ultimately, the resulting methods should yield a temporally-coherent shape representation of the observed deformable object.

Multiview reconstruction methods have been well-studied to address 4D reconstruction. While current methods achieve impressive results [12, 9, 6, 27, 32, 36], they typically require well-engineered and expensive setups, where the deforming object is captured from multiple viewpoints essentially covering its entire 3D surface. By contrast, simpler acquisition devices, such as stereo cameras or the increasingly popular low-cost depth sensors, allow us to acquire 3D data in a more affordable manner. Unfortunately, these devices typically only produce partial observations of the object, depicting, at best, half of its 3D surface.

In the case of rigid motion, several fusion techniques have been proposed to combine multiple partial 3D observations [15, 28, 41]. However, when it comes to capturing a dynamically deforming object, the literature remains very sparse. More specifically, most existing methods [17, 7, 18, 35, 39, 42] rely on a pre-processing stage, where the object undergoes (quasi-) rigid motion, to acquire a complete 3D template of the object, which will then be deformed to match new non-rigid data.

By contrast, in this paper, we introduce a template-less 4D reconstruction method that directly estimates a complete temporally-coherent 3D model from partial observations of a deforming objects. In other words, as illustrated in Fig. 1, we incrementally fuse the partial observations into a complete model while accounting for the deformations of these observations. Ultimately, this lets us estimate the complete shape of the deforming object in each frame of a video sequence, thus predicting the hidden parts of the object de-

spite the very small amount of observations in a frame and the fact that the object moves non-rigidly.

At the core of our algorithm, we make use of a subspace representation of the object deformations, which has proven powerful in the context of non-rigid structure-from-motion (NRSfM) [8, 10, 14]. In contrast to typical NRSfM methods, however, we tackle the scenario where the points are only visible in small portions of the sequence and without readily available correspondences, but exploit depth information. To this end, we therefore design an online, two-step algorithm: For each new frame, we first perform a non-rigid registration of the partial observations to the shape estimate of the previous frame. We then update the deformation subspace to integrate the new 3D points and better represent the new frame, as well as still fit to the observations of the previous frames. Both steps of our approach can be performed via iterative algorithms that only involve simple and efficient mathematical operations. Ultimately, we predict the 3D location of each point in each frame of the sequence, whether the point was observed or not.

We demonstrate the effectiveness of our approach on several challenging sequences, including sequences of partial 3D point clouds and kinect videos. Our quantitative and qualitative evaluations evidence that our method can recover accurate 4D models from partial observations of objects undergoing large deformations.

2. Related Work

Many methods have been proposed to address the problem of 4D reconstruction of non-rigid motion. In particular, a vast portion of the literature has focused on the problem of multiview reconstruction [12, 9, 23, 6, 27, 32, 36], where multiple cameras are placed so as to observe most of the surface of the object of interest. With the recent availability of low-cost depth sensors, several methods have proposed to rely on multiview depth maps, thus exploiting more informative RGBD data [33, 29, 31, 38, 11]. Here, however, we focus on the problem of single-view 4D reconstruction, which alleviates the need for the relatively complex setups required by multiview approaches.

When considering a standard RGB camera, single-view reconstruction also has a long history in computer vision. In particular, non-rigid structure-from-motion [4] quickly emerged as a generalization of the rigid factorization approach of [30]. Recently, great progress has been made in this area, such as implicit low-rank shape models [24], prior-free approaches [8], dense reconstruction techniques [13] and methods handling different types of motions [25]. While effective, these methods can still only handle small amounts of missing data (i.e., points visible in only some frames of the sequence), and are thus mostly limited to relatively small deformations, or deformations of open surfaces. By contrast, we propose to exploit the increasingly popular depth sensors, which provide richer information, to perform 4D reconstruction of deforming objects from partial observations that depict, at best, half of the object's surface.

Several methods have tackled the problem of non-rigid 4D reconstruction from partial depth measurements. In particular, some techniques tackle the case of open surfaces, where the surface can be entirely observed in at least some frames of the sequence [37, 16]. While some work has tackled the more challenging case of volumetric surfaces, most existing methods rely on a pre-processing step, where a 3D model of the object of interest is acquired under rigid, or quasi-rigid motion [17, 7, 18, 35, 39, 40, 42]. While methods relying on quasi-rigidity during this model-building step do account for some degree of deformations, they are very far from attempting to directly estimate the 4D motion of, e.g., a human dancing in front of the sensor.

To the best of our knowledge, only two methods have proposed to tackle such a challenging scenario [20, 19]. In [20], a moving human body was reconstructed in a piecewise rigid manner from a kinect. This method, however, exploits the availability of a 3D skeleton model fitted to each frame of the sequence. As a consequence, it does not generalize to other objects, or even to people wearing loose garments. By contrast, [19] addresses the more general scenario of non-rigid shape estimation from partial 3D observations by warping the partial observations of each frame in a sequence to one specific, reference frame. The final shape model in the reference frame is then obtained by fusing the different partial observations using a volumetric signed-distance function (SDF) representation. As a consequence, this method suffers from the computational drawback of having to perform the warping and fusion operations multiple times, by sequentially treating each frame as reference frame. Furthermore, due to its SDF-based shape representation, it only produces unrelated reconstructions in each frame, and thus fails to perform 4D reconstruction.

By contrast, here, we introduce an online algorithm to fuse partial 3D observations of a deforming object. As a result, we produce a complete 4D model for the entire sequence, that accurately infers the missing parts of the object in all the frames.

The work of [22], virtually concurrent to ours, aims at the same goal as us. Our approach fundamentally differs from this work in the representation of the surface deformations, i.e., warp-field for [22] versus subspace for us. The main benefit of our representation is that it allows us to estimate the position of newly observed points even in the previous frames where these points were hidden.

3. Deformable 3D Fusion

In this section, we introduce our approach to estimating a complete 4D model from a sequence of partial 3D ob-

Figure 2. Measurement matrix $\mathbf{X}_f$ with missing entries in black.

servations of a deforming object. Our algorithm, given by Algorithm 1, works in an online manner, and, at each frame f, performs the following two steps: (i) Register the partial observations to the current model while accounting for deformations; (ii) Update the deformation model and the shape in each frame up to f. Below, we first discuss our approach to addressing the second step, and then focus on non-rigid registration.

3.1. Subspace Learning for 4D Reconstruction

Let us first assume that we have established correspondences between the partial observations acquired up to frame f, and that the rigid component of the motion in each frame has been removed. The details of this registration step will be given in Section 3.2. Here, we propose to model the remaining deformations with a linear subspace. Such a subspace model not only provides us with a compact representation of the shape in each frame, but also allows us to predict the locations of the points that are occluded in each frame. We therefore cast 4D reconstruction as the problem of learning a subspace from partial observations.

More specifically, let $\mathbf{x}_j^i = [x_j^i, y_j^i, z_j^i]^T$ denote the 3D location of point i in frame j. At frame f, given the correspondences between the partial observations of each frame, we can build a measurement matrix $\mathbf{X}_f \in \mathbb{R}^{3N_f \times f}$, with N_f the total number of points observed up to frame f, of the form

$$\mathbf{X}_f = \begin{bmatrix} \mathbf{x}_1^1 & \cdots & \mathbf{x}_f^1 \\ \vdots & & \vdots \\ \mathbf{x}_1^{N_f} & \cdots & \mathbf{x}_f^{N_f} \end{bmatrix} ,$$

where the missing (unobserved) entries in each frame are replaced by zeros (or any arbitrary value). Note that, as illustrated by Fig. 2, $\mathbf{X}_f$ contains a large number of missing entries, since, in our scenario, at most half of the object is seen in each frame.

As mentioned before, we assume that the deformations of the object lie on a low-dimensional subspace. This lets us predict a temporally-coherent 3D surface as

$$\hat{\mathbf{X}} = \bar{\mathbf{X}} + \mathbf{S}\mathbf{W} ,$$

where $\mathbf{S} \in \mathbb{R}^{3N_f \times d}$ and $\mathbf{W} \in \mathbb{R}^{d \times f}$ denote the deformation subspace and the corresponding coefficients, respectively. $\bar{\mathbf{X}} = \bar{\mathbf{x}}\mathbf{1}_f^T$, with $\mathbf{1}_f$ an f-dimensional column vector of ones, is the matrix containing f copies of the mean shape $\bar{\mathbf{x}}$. In practice, we compute this mean shape by averaging over the observations of each point.

Reconstructing a 4D surface can then be achieved by finding the subspace $\mathbf{S}$ and coefficients $\mathbf{W}$ that best fit the given partial observations $\mathbf{X}_f$. This can be expressed as the optimization problem

$$\min_{\mathbf{S},\mathbf{W}} \ \|\mathbf{\Omega} \odot (\mathbf{X} - \bar{\mathbf{X}} - \mathbf{S}\mathbf{W})\|_F^2 , \tag{1}$$

where $\mathbf{\Omega}$ is the (known) visibility matrix, $\odot$ denotes the Hadamard (elementwise) product and $\|\cdot\|_F$ is the Frobenius norm.

Due to the large number of occluded points, the reconstruction of the unobserved points obtained with our subspace may still be noisy. To address this issue, we make use of a Laplacian regularizer that constrains the local deformations of the surface [26]. Following [26], we estimate the Laplacian of 3D point i in frame j as

$$\mathbf{l}(\mathbf{x}_j^i) = \mathbf{x}_j^i - \frac{1}{K} \sum_{k \in \mathcal{N}_K(i)} \mathbf{x}_j^k , \tag{2}$$

where K is the number of nearest neighbors of $\mathbf{x}_j^i$ taken into account, and $\mathcal{N}_K(i)$ is the set of indices of these neighbors. An affine-invariant Laplacian regularizer for point i in frame j can then be written as

$$r_j^i = \|\mathbf{T}_j^i \mathbf{l}(\mathbf{x}_{r_i}^i) - \mathbf{l}(\mathbf{x}_j^i)\|_2^2 ,$$

where r_i is the index of the frame in which a reference Laplacian for point i is computed, and $\mathbf{T}_j^i$ is the affine transformation that aligns this reference Laplacian to the one in frame j. Note that, in contrast to [26], since no frame depicts the entire surface, in our case the reference Laplacian needs to be computed in a different frame for each point. We find this reference frame, as well as the best value for $K \in \{4, 5, 6\}$, by finding the Laplacian with smallest L_2-norm (i.e., the frame and value of K for which the K nearest neighbors of point i best approximate $\mathbf{x}_j^i$, according to Eq. 2). As shown in [26], the affine transformation $\mathbf{T}_i^j$ can directly be obtained from the coordinates of the points in frame j. Therefore, the regularizers r_j^i for all points i and all frames j can be encoded with a single matrix $\mathbf{L} \in \mathbb{R}^{3N_f \times 3N_f}$ acting on the coordinates of the points.

By adding such a Laplacian regularizer to our objective function, we can express 4D reconstruction as the solution to the optimization problem

$$\min_{\mathbf{S},\mathbf{W}} \ \|\mathbf{\Omega} \odot (\mathbf{X} - \bar{\mathbf{X}} - \mathbf{S}\mathbf{W})\|_F^2 + \gamma \|\mathbf{L}(\bar{\mathbf{X}} + \mathbf{S}\mathbf{W})\|_F^2 , \tag{3}$$

where γ is the weight of the Laplacian regularizer. Solving (3) in an efficient manner is made difficult by the Hadamard product that accounts for the missing observations. In fact, this is a well-known problem in the matrix

factorization literature [5, 1, 3]. Furthermore, compared to standard matrix factorization, we have an additional term in our objective function. To account for these difficulties, here, we introduce an algorithm based on the Alternating Direction Method of Multipliers [2] , which has proven more effective than simple alternating schemes in practice.

More specifically, let us first re-write (3) as

$$\begin{aligned} \min_{\mathbf{S},\mathbf{W},\mathbf{Z}} \ & \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{Z})\|_F^2 + \gamma \|\mathbf{L}(\bar{\mathbf{X}} + \mathbf{SW})\|_F^2 \\ \text{s.t.} \ & \mathbf{Z} = \bar{\mathbf{X}} + \mathbf{SW} \ , \end{aligned} \tag{4}$$

where we introduced an auxiliary variable $\mathbf{Z}$. The augmented Lagrangian of (4) can be expressed as

$$\begin{aligned} L(\mathbf{S},\mathbf{W},\mathbf{Z},\Lambda) \ = \ & \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{Z})\|_F^2 + \gamma \|\mathbf{L}(\bar{\mathbf{X}} + \mathbf{SW})\|_F^2 \\ + \ & \mathrm{tr}\left(\Lambda^T (\mathbf{Z} - \bar{\mathbf{X}} - \mathbf{SW})\right) + \rho/2 \|\mathbf{Z} - \bar{\mathbf{X}} - \mathbf{SW}\|_F^2 \ , \end{aligned}$$

where Λ is the matrix of Lagrange multipliers corresponding to the constraints, and ρ is the standard parameter of the ADMM. The ADMM then consists of iteratively minimizing the augmented Lagrangian w.r.t. each variable $\mathbf{Z}$, $\mathbf{S}$, $\mathbf{W}$, and updating the Lagrange multipliers. In our case, each of these operations has a simple closed-form solution.

Computing Z:
The auxiliary variable $\mathbf{Z}$ are obtained by solving

$$\min_{\mathbf{Z}} \ \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{Z})\|_F^2 + \frac{\rho}{2} \|\mathbf{Z} - \bar{\mathbf{X}} - \mathbf{SW} + 1/\rho \Lambda\|_F^2 \ , \tag{5}$$

where we grouped the linear and quadratic terms of the augmented Lagrangian in a single quadratic term [2]. The solution to this problem can be obtained independently for each element $\mathbf{Z}_j^i$ of the matrix $\mathbf{Z}$. For a single element, it can be written as

$$\mathbf{Z}_j^i = \left(2\Omega_j^i \mathbf{X}_j^i + \rho(\bar{\mathbf{X}}_j^i + \mathbf{S}^i \mathbf{W}_j) - \Lambda_j^i\right) / \left(2\Omega_j^i + \rho\right) \ , \tag{6}$$

where $\mathbf{S}^i$ is the i^{th} row of $\mathbf{S}$ and $\mathbf{W}_j$ the j^{th} column of $\mathbf{W}$.

Computing S:
The subspace $\mathbf{S}$ is obtained by solving

$$\min_{\mathbf{S}} \ \gamma \|\mathbf{L}(\bar{\mathbf{X}} + \mathbf{SW})\|_F^2 + \frac{\rho}{2} \|\mathbf{Z} - \bar{\mathbf{X}} - \mathbf{SW} + 1/\rho \Lambda\|_F^2 \ , \tag{7}$$

which is simply a least-squares problem. The subspace can thus be obtained as the solution of the linear system

$$\left(2\gamma \mathbf{L}^T \mathbf{L} + \rho \mathbf{I}\right) \mathbf{SWW}^T = \left(\rho(\mathbf{Z} - \bar{\mathbf{X}}) + \Lambda - 2\gamma \mathbf{L}^T \mathbf{L} \bar{\mathbf{X}}\right) \mathbf{W}^T, \tag{8}$$

which can be computed efficiently in matrix form.

Computing W:
The coefficients $\mathbf{W}$ are obtained by solving

$$\min_{\mathbf{W}} \ \gamma \|\mathbf{L}(\bar{\mathbf{X}} + \mathbf{SW})\|_F^2 + \frac{\rho}{2} \|\mathbf{Z} - \bar{\mathbf{X}} - \mathbf{SW} + 1/\rho \Lambda\|_F^2 \ , \tag{9}$$

which, as before, is a least-squares problem whose solution can be efficiently computed in closed-form.

Updating Λ:
At each iteration t, the Lagrange multipliers are updated as

$$\Lambda^t = \Lambda^{t-1} + \rho \left(\mathbf{Z} - \bar{\mathbf{X}} - \mathbf{SW}\right) \ . \tag{10}$$

In practice, we initialize ρ to a small value ρ_0 (typically, we use $\rho_0 = 1e^{-3}$), and increase it at a fixed rate at each iteration. The Lagrange multipliers are initialized to 0. $\mathbf{S}$ and $\mathbf{W}$ are initialized to random matrices. We then run our algorithm until convergence, or until a maximum number of iterations is reached. Note that our objective function remains non-convex. Therefore, convergence is not guaranteed, and the solution is likely to be suboptimal. In practice, however, we observed that the algorithm behaves well.

3.2. Subspace-based Non-rigid Registration

In the previous section, we assumed that the correspondences between the partial observations of the different frames were given. We now discuss our approach to establishing these correspondences. In particular, our approach relies on our subspace-based deformation model.

More specifically, let $\{\mathbf{x}_f^i\}_{i=1}^M$ be the set of M observations of a new frame f. Our goal is to estimate the correspondence between these new observations and the points on the current model, while accounting for the fact that some points may not have been observed at all in the $f-1$ previous frames. To this end, following [21], we introduce a probability matrix $\mathbf{P} \in \mathbb{R}^{M \times N_{f-1}}$, whose element $\mathbf{P}_j^i$ represents the probability of $\mathbf{x}_f^i$ corresponding to the j^{th} model point. We employ this probabilistic correspondence assignment because it is innately more robust than the binary assignment used in ICP. As shown in [21], this representation encodes a Gaussian Mixture Model whose centroids correspond to the model points. The location of these centroids, as well as the variance σ^2 of the Gaussians, can then be searched for so as to minimize the negative log-likelihood of the observed points. This can be expressed as

$$\min_{\mathbf{P},\hat{\mathbf{x}},\sigma^2} \frac{1}{2\sigma^2} \sum_{i=1}^{M} \sum_{j=1}^{N_{f-1}} \mathbf{P}_j^i \|\mathbf{x}_f^i - \hat{\mathbf{x}}^j\|_2^2 + \frac{3N_P}{2} \log \sigma^2, \tag{11}$$

where $\hat{\mathbf{x}}$ encodes the location of the centroids, or in other words the shape of the surface in the current frame , and $N_p = \sum_{m=1}^{M} \sum_{n=1}^{N_{f-1}} \mathbf{P}_m^n$.

Here, to further regularize this problem, we make use of our subspace representation. However, since the subspace obtained from the previous frames may not be rich enough to accurately represent the shape of the surface in the new frame, we do not explicitly encode $\hat{\mathbf{x}} = \mathbf{Sw}$, but rather encourage the shape to remain close to the model. To this end, and to avoid having to compute the coefficients $\mathbf{w}$, we penalize the deformations that lie in the nullspace $\mathbf{N}$ of our subspace $\mathbf{S}$. This lets us express non-rigid registration as

the optimization problem

$$\min_{\mathbf{P},\hat{\mathbf{x}},\sigma^2} \frac{1}{2\sigma^2}\sum_{i=1}^{M}\sum_{j=1}^{N_{f-1}} \mathbf{P}_j^i \|\mathbf{x}_f^i - \hat{\mathbf{x}}^j\|_2^2 + \frac{3N_P}{2}\log\sigma^2 + \lambda\|\mathbf{N}^T(\hat{\mathbf{x}} - \bar{\mathbf{x}})\|_2^2, \tag{12}$$

where λ is the weight of the subspace prior. We solve this optimization problem by alternating over the variables. The value of each variable, while the other ones are fixed, can be obtained as follows.

Computing P:
Following [21], each element in $\mathbf{P}$ can be computed as

$$\mathbf{P}_j^i := \frac{\exp^{-\frac{1}{2\sigma^2}\|\mathbf{x}_f^i - \hat{\mathbf{x}}^j\|_2^2}}{\sum_{l=1}^{N_{f-1}} \exp^{-\frac{1}{2\sigma^2}\|\mathbf{x}_f^i - \hat{\mathbf{x}}^l l\|_2^2} + \frac{w}{1-w}\frac{(2\pi\sigma^2)^{3/2}N_{f-1}}{M}},$$

where w is a parameter that reflects the expected proportion of outliers.

Computing $\hat{\mathbf{x}}$:
To derive the solution for $\hat{\mathbf{x}}$, let us first rewrite our optimization problem directly in terms of the matrix $\mathbf{P}$, which yields

$$\min_{\mathbf{P},\mathbf{x},\sigma^2} \frac{1}{2\sigma^2}\cdot \left(\mathbf{x}_f^T\left(\mathbf{I}\otimes d(\mathbf{P1})\right)\mathbf{x}_f - 2\mathbf{x}_f^T\left(\mathbf{I}\otimes\mathbf{P}\right)\hat{\mathbf{x}} + \hat{\mathbf{x}}^T\left(\mathbf{I}\otimes d(\mathbf{P^T1})\right)\hat{\mathbf{x}}\right) + \lambda\|\mathbf{N}^T(\hat{\mathbf{x}} - \bar{\mathbf{x}})\|_2^2, \tag{13}$$

where $d(\cdot)$ denotes the diagonal matrix obtained from a vector, and where we have ignored the term that does not depend on $\hat{\mathbf{x}}$. Setting the gradient of the objective function with respect to $\hat{\mathbf{x}}$ to zero yields

$$\frac{1}{\sigma^2}\left(-\left(\mathbf{I}\otimes\mathbf{P}^T\right)\mathbf{x}_f + \left(\mathbf{I}\otimes d(\mathbf{P^T1})\right)\hat{\mathbf{x}}\right) + 2\lambda\mathbf{N}\mathbf{N}^T(\hat{\mathbf{x}} - \bar{\mathbf{x}}) = 0, \tag{14}$$

which results in a closed-form solution for $\hat{\mathbf{x}}$.

Computing σ^2:
The variance σ^2 can be obtained by setting the derivative of the objective function w.r.t. σ^2 to zero, which yields

$$\sigma^2 = \frac{1}{3N_{\mathbf{P}}}\sum_{i=1}^{M}\sum_{j=1}^{N_{f-1}} \mathbf{P}_j^i\|\mathbf{x}_f^i - \hat{\mathbf{x}}^j\|_2^2.$$

We run this alternating scheme until convergence, and then extract the correspondences from the resulting $\mathbf{P}$. More precisely, we start from the maximum probability in $\mathbf{P}$ and iteratively find the corresponding point for each observation, while avoiding duplicate correspondences (i.e., two observations corresponding to the same model point). If the maximum probability of an observed point is less than the outlier probability, we treat this observation as a new point in our model, which will then create 3 new rows in the measurement matrix. The outlier probability P_w is computed following [21] as

$$P_w = \frac{w}{1-w}\frac{(2\pi\sigma^2)^{3/2}N_{f-1}}{M}.$$

In practice, we found that the correspondences could be improved by updating the subspace. Therefore, for each frame, we iterate between estimating the correspondences with the current subspace and refining the subspace from the new correspondences, following the technique of Section 3.1. We stop this iterative procedure when the number of points registered to the model is stable (i.e., when no new points are registered compared to the previous iteration), which typically only requires a few iterations.

Rigid Motion and Initialization:
Subspace models are best-suited to only represent non-rigid deformations. Therefore, we seek to remove the rigid motion of the new observations. To this end, we first establish correspondences between the current model and the new observations using CPD [21]. We then estimate the rigid transformation (rotation and translation), which can be achieved by singular value decomposition, and remove this rigid motion from the observations. The subspace can then be re-estimated with this rigidly aligned data and with the CPD correspondences, following the technique of Section 3.1.

Algorithm 1 Deformable 3D Fusion

Initialization:
F := Number of frames
$\hat{\mathbf{x}}_1$:= observed points in frame 1
$\mathbf{X}_1 := \hat{\mathbf{x}}_1$

Iteration:
for $f = 2 : \mathrm{F}$ **do**
  $\mathbf{x}_f$:= observed points in frame f,
  Estimate correspondences between $\mathbf{x}_f$ and $\hat{\mathbf{x}}_{f-1}$ using CPD.
  Estimate and remove rigid motion of $\mathbf{x}_f$.
  while New points registered **do**
    Form $\mathbf{X}_f$ from $\mathbf{X}_{f-1}$ and $\mathbf{x}_f$.
    Estimate mean shape $\bar{\mathbf{x}}$ and Laplacian $\mathbf{L}$.
    Estimate deformation subspace and coefficients $\mathbf{S}, \mathbf{W}$, see Section 3.1.
    Estimate correspondences between $\mathbf{x}_f$ and $\hat{\mathbf{x}}_{f-1}$, see Section 3.2.
  end while
  Form $\mathbf{X}_f$ from $\mathbf{X}_{f-1}$ and $\mathbf{x}_f$.
  Estimate mean shape $\bar{\mathbf{x}}$ and Laplacian $\mathbf{L}$.
  Estimate deformation subspace and coefficient $\mathbf{S}, \mathbf{W}$, see Section 3.1.
end for

Output: 4D reconstruction: $\hat{\mathbf{X}} = \bar{\mathbf{X}} + \mathbf{SW}$

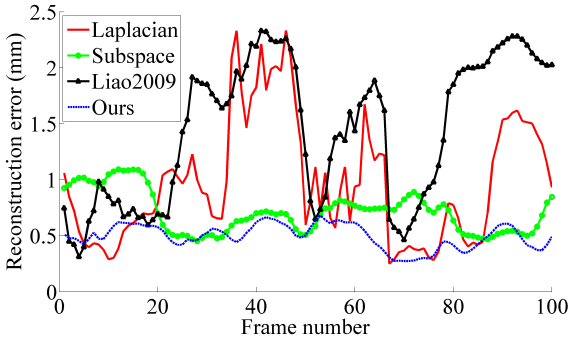

Figure 3. RMS reconstruction error for the pants sequence. Average error over all frames: Laplacian: 0.9767; Subspace: 0.7103; Liao2009: 1.4390; **Ours: 0.4963**.

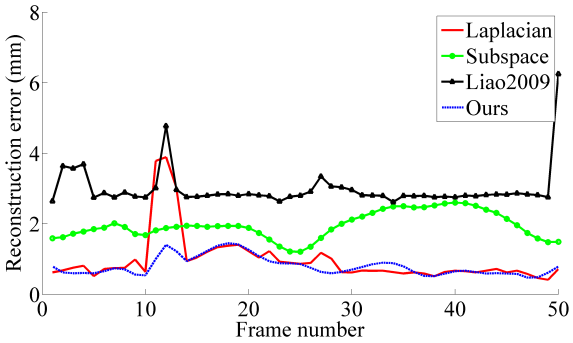

Figure 4. RMS reconstruction error for the woman sequence. Average error over all frames: Laplacian: 0.9619; Subspace: 1.9525; Liao2009: 2.9833; **Ours: 0.7976**.

4. Experimental Results

We now evaluate our 4D reconstruction method. To this end, we made use of three point-cloud sequences (the pants sequence of [34], and the man and woman sequences of [9]), and of two sequences captured with a Kinect. The three point-cloud sequences allow us to perform quantitative evaluations, while the Kinect sequences illustrate the use of our approach on depth measurements. For the point-cloud sequences, we generated realistic partial observations by computing depth maps corresponding to specific viewpoints[1], and removing the occluded points. For computational reasons, we downsampled the man, woman and Kinect sequences. In our experiments, we set the Laplacian weight in (3) to $\gamma = 1$, the subspace weight in (12) to $\lambda = 1$, and the outlier parameter to $w = 10^{-15}$ for the point-cloud sequences and $w = 10^{-8}$ for the Kinect sequences, which include more noise. Below, we first evaluate our subspace learning method of Section 3.1 and then present the results of our full 4D reconstruction algorithm.

4.1. Results of Subspace Learning

To evaluate the quality of our subspace learning reconstruction, we make use of the ground-truth correspondences

[1]For the man and woman sequences, we used the calibration data provided with the sequences.

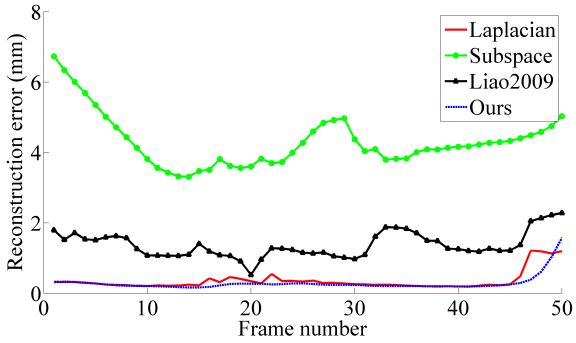

Figure 5. RMS reconstruction error for the man sequence. Average error over all frames: Laplacian: 0.3537; Subspace: 4.3070; Liao2009: 1.3730; **Ours: 0.2874**.

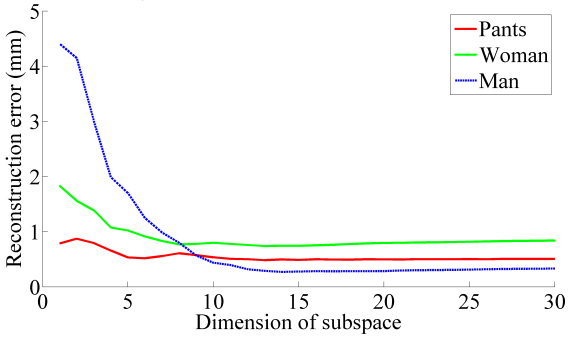

Figure 6. RMS reconstruction error as a function of the subspace dimension for the three point-cloud sequences.

to build the partial measurement matrix $\mathbf{X}_F$, containing the observations of all F frames of a sequence. Note that this can only be achieved with the point-cloud sequences, since no ground-truth is available for the Kinect sequence. To illustrate the importance of the different components of our approach, we compare the results of our subspace learning method against the following baselines:

Laplacian: $\min_{\mathbf{Y}} \|\mathbf{\Omega} \odot (\mathbf{Y} - \mathbf{X})\|_F^2 + \gamma \|\mathbf{L}\mathbf{Y}\|_F^2$,

Subspace: $\min_{\mathbf{S},\mathbf{W}} \quad \|\mathbf{\Omega} \odot (\mathbf{X} - \bar{\mathbf{X}} - \mathbf{S}\mathbf{W})\|_F^2$.

Furthermore, we also report the results of the method of [19], which we refer to as **Liao2009**.

In Figs. 3, 4 and 5, we report the RMS reconstruction error for each frame in the three sequences. While the behaviors of Laplacian and Subspace vary across different sequences, our algorithm performs consistently well. Furthermore, our subspace learning method clearly outperforms Liao2009. In Fig. 6, we report the RMS reconstruction error (averaged over all frames) as a function of the subspace dimension for the three sequences. Note that, with sufficiently large dimensions, our method is very robust to this parameter. In our experiments, we used a 20-dimensional subspace.

4.2. Results of the Full Algorithm

We now evaluate the results of our complete 4D reconstruction algorithm. Here, we make use of a baseline that

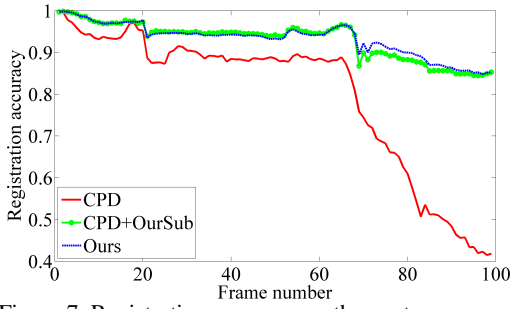

Figure 7. Registration accuracy on the pants sequence.

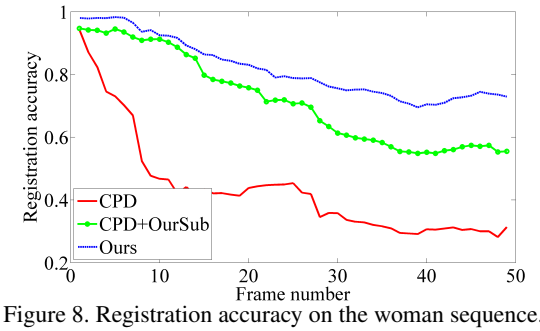

Figure 8. Registration accuracy on the woman sequence.

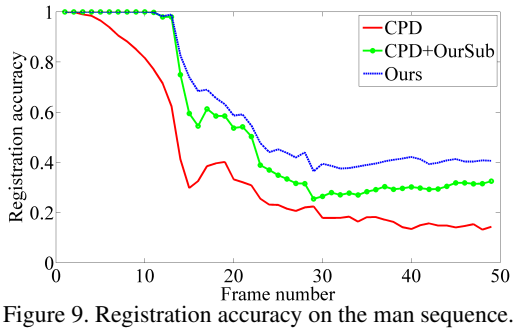

Figure 9. Registration accuracy on the man sequence.

consists of estimating the correspondences using CPD instead of our subspace-based method of Section 3.2, followed by our subspace learning technique of Section 3.1. We refer to this baseline as CPD+OurSub. Furthermore, we also report the results of using CPD followed by the approach of [19], which we refer to as CPD+Liao2009.

In Figs. 7, 8 and 9, we compare the registration accuracy of all methods (computed as the proportion of correctly matched points) for each frame of the three point-cloud sequences. Note that using our subspace-based approach clearly improves over the CPD registration. In Figs. 10, 11 and 12, we report the RMS reconstruction errors of the methods for the three point-cloud sequences. Note again that our approach outperforms the baselines. While CPD+OurSub performs well, recall that it also relies on our subspace learning approach.

To evaluate the robustness of our method to noise, we added zero mean Gaussian noise with different standard deviations to the points in the woman sequence. As shown in Table 1, while the reconstruction error increases, our method remains relatively robust to noise.

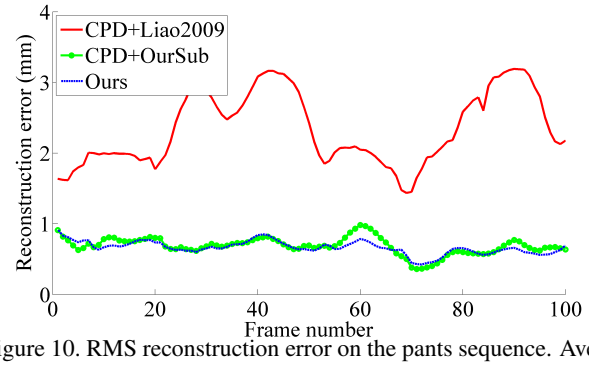

Figure 10. RMS reconstruction error on the pants sequence. Average error over all frames: CPD+Liao2009: 2.3437; CPD+OurSub: 0.6875; **Ours: 0.6640**.

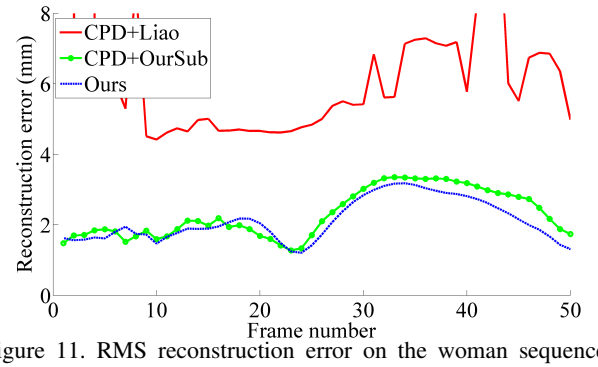

Figure 11. RMS reconstruction error on the woman sequence. Average error over all frames: CPD+Liao2009: 7.6480; CPD+OurSub: 2.3050; **Ours: 2.1173**.

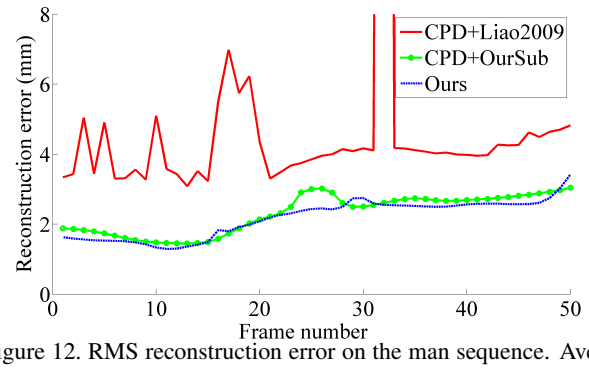

Figure 12. RMS reconstruction error on the man sequence. Average error over all frames: CPD+Liao2009: 5.5454; CPD+OurSub: 2.3023; **Ours: 2.1711**.

Noise std (mm)	0	5	10
RMS errors (mm)	2.12	10.96	12.35

Table 1. Reconstruction errors with different levels of noise.

In Fig. 13, we provide qualitative comparisons of the reconstructions obtained by the different methods on the five sequences. Note that our approach yields more realistic results than the baselines. While the Kinect data is more challenging, due to its lack of exact correspondences, we are still able to obtain accurate reconstructions. Note that our template-free method and subject-independent deformation model allow us to deal with arbitrary types of objects. We therefore evaluated it on a publicly available

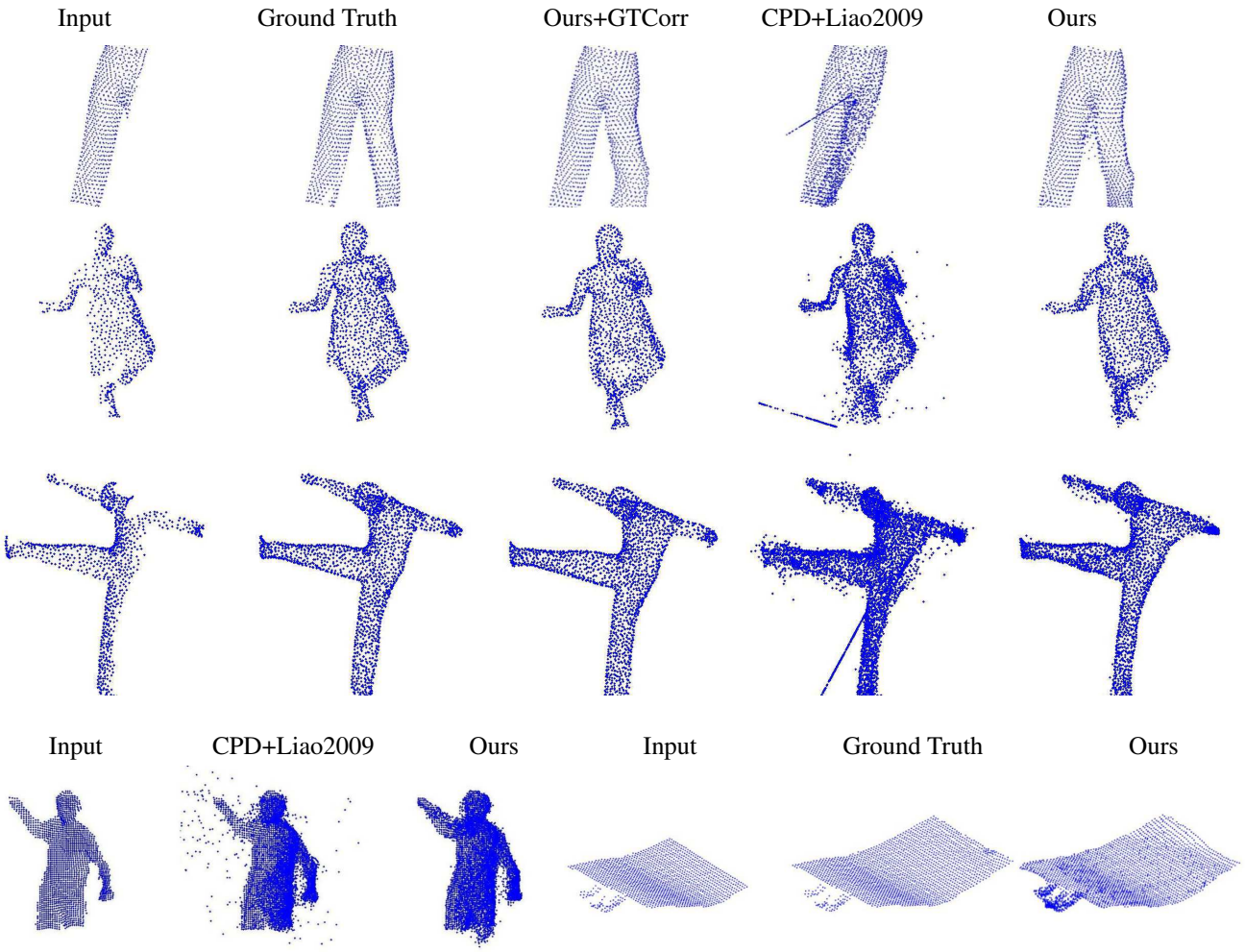

Figure 13. Qualitative comparison of the reconstructions obtained by the different methods on the five sequences (pants, woman, man, Kinect and paper). Ours+GTCorr refers to our subspace reconstruction method with ground-truth correspondences.

Kinect sequence depicting a deforming sheet of paper[2], and, as in [37], augmented the data with a synthetic occluder, which hides roughly half of the surface in the first frame, and is then progressively removed. The results in the bottom row of Fig. 13 illustrate that our approach can accurately reconstruct the entire surface. The complete videos are provided as supplementary material.

5. Conclusion

We have presented a template-less approach to 4D reconstruction of non-rigid objects from highly-incomplete 3D data. Our online algorithm allows us to incrementally fuse the partial observations in a temporally-coherent model. Thanks to the deformation subspace learned from the observations, our approach can predict the hidden parts of the object, and thus reconstruct a complete 4D representation. Our experimental results have demonstrated the effectiveness of our method on several challenging sequences. The main limitations of our method lie in the potential drift and drop in accuracy caused by large inter-frame deformations. In the future, we intend to overcome drift by updating the previous correspondences, and incorporate 3D features to improve the accuracy of registration for large deformations.

Acknowledgements: This work was supported in part by National Program on Key Basic Research Project of China (973 Program) 2013CB328805 and National Natural Science Foundation of China (NSFC) 61370134. The authors would also acknowledge the financial support from China Scholarship Council. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, as well as by the Australian Research Council through the ICT Centre of Excellence program.

References

[1] H. Aanæs, R. Fisker, K. Astrom, and J. M. Carstensen. Robust factorization. *PAMI*, 2002.

[2] Publicly available at http://cvlab.epfl.ch/data/dsr

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011.

[3] S. Brandt. Closed-form solutions for affine reconstruction under missing data. In *ECCV 2002 Workshops*.

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR 2000*.

[5] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *CVPR 2005*.

[6] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic Deformable Surface Tracking From Multiple Videos. In *ECCV 2010*.

[7] Y. Cui, W. Chang, T. Nöll, and D. Stricker. Kinectavatar: fully automatic body capture using a single kinect. In *ACCV 2012 Workshops*.

[8] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *IJCV*, 2014.

[9] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *TOG*, 2008.

[10] A. Del Bue. A factorization approach to structure from motion with shape priors. In *CVPR 2008*.

[11] M. Dou, H. Fuchs, and J.-M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *ISMAR 2013*.

[12] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. In *CVPR 2008*.

[13] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR 2013*.

[14] P. F. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *PAMI*, 2011.

[15] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *UIST 2011*.

[16] A. Jordt and R. Koch. Direct model-based tracking of 3d object deformations in depth and color video. *IJCV*, 2013.

[17] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. *TOG*, 2009.

[18] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *TOG*, 2013.

[19] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong. Modeling deformable objects from a single depth camera. In *ICCV 2009*.

[20] C. Malleson, M. Klaudiny, A. Hilton, and J.-Y. Guillemaut. Single-view rgbd-based reconstruction of dynamic human geometry. In *ICCV 2013 Workshops*.

[21] A. Myronenko and X. Song. Point set registration: Coherent point drift. *PAMI*, 2010.

[22] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR 2015*.

[23] M. R. Oswald, J. Stühmer, and D. Cremers. Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *ECCV 2014*.

[24] M. Paladini, A. Bartoli, and L. Agapito. Sequential non-rigid structure-from-motion with the 3d-implicit low-rank shape model. In *ECCV 2010*.

[25] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *ECCV 2014*.

[26] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *SGP 2004*.

[27] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 2007.

[28] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS 2012*.

[29] A. Tevs, A. Berner, M. Wand, I. Ihrke, M. Bokeloh, J. Kerber, and H.-P. Seidel. Animation cartographyŁntrinsic reconstruction of shape and motion. *TOG*, 2012.

[30] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 1992.

[31] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *TVCG*, 2012.

[32] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *TOG*, 2008.

[33] A. Weiss, D. Hirshberg, and M. J. Black. Home 3d body scans from noisy image and range data. In *ICCV 2011*.

[34] R. White, K. Crane, and D. A. Forsyth. Capturing and animating occluded cloth. In *SIGGRAPH 2007*.

[35] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt. On-set performance capture of multiple actors with a stereo camera. *TOG*, 2013.

[36] F. Xu, Y. Liu, C. Stoll, J. Tompkin, G. Bharaj, Q. Dai, H.-P. Seidel, J. Kautz, and C. Theobalt. Video-based characters: creating new human performances from a multi-view video database. *TOG*, 2011.

[37] W. Xu, M. Salzmann, Y. Wang, and Y. Liu. Nonrigid surface registration and completion from rgbd images. In *ECCV 2014*.

[38] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Performance capture of interacting characters with handheld kinects. In *ECCV 2012*.

[39] M. Zeng, J. Zheng, X. Cheng, and X. Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *CVPR 2013*.

[40] Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *CVPR 2014*.

[41] Q.-Y. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *TOG*, 2013.

[42] M. Zollhöfer, M. Niessner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *TOG*, 2014.