Multi-View Dynamic Shape Refinement Using Local Temporal Integration

Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer

Multi-View Dynamic Shape Refinement Using Local Temporal Integration

Vincent Leroy Jean-Sebastien Franco Edmond Boyer
INRIA Grenoble Rhône-Alpes, LJK - Grenoble Universities, France
{vincent.leroy,jean-sebastien.franco,edmond.boyer}@inria.fr

Abstract

We consider 4D shape reconstructions in multi-view environments and investigate how to exploit temporal redundancy for precision refinement. In addition to being beneficial to many dynamic multi-view scenarios this also enables larger scenes where such increased precision can compensate for the reduced spatial resolution per image frame. With precision and scalability in mind, we propose a symmetric (non-causal) local time-window geometric integration scheme over temporal sequences, where shape reconstructions are refined framewise by warping local and reliable geometric regions of neighboring frames to them. This is in contrast to recent comparable approaches targeting a different context with more compact scenes and real-time applications. These usually use a single dense volumetric update space or geometric template, which they causally track and update globally frame by frame, with limitations in scalability for larger scenes and in topology and precision with a template based strategy. Our templateless and local approach is a first step towards temporal shape super-resolution. We show that it improves reconstruction accuracy by considering multiple frames. To this purpose, and in addition to real data examples, we introduce a multi-camera synthetic dataset that provides ground-truth data for mid-scale dynamic scenes.

Figure 1. A challenging dynamic scene with fast motions and a mid-scale acquisition space, hence low image resolution on shapes in addition to motion blur. Temporal integration helps recovering highly detailed models.

1. Introduction

We address multi-view 4D modeling of dynamic scenes observed with a set of color cameras. We are particularly interested in challenging scenes, of mid-scale size (dozen square meters or more), with possibly fast motions and multiple people. These are a prominent feature of numerous moving surface capture scenarios, for instance sport moves with running, combat, or dancing over a large area. Addressing this use case enhances the creative possibilities for many applications typically associated to 3D content creation and such as sports analysis, cultural heritage preservation or virtual reality applications.

Increasing the acquisition space of multi-camera set-ups raises challenges since it generally requires larger camera field of views and more distant cameras, leading to lower pixel coverage of the scene for fixed sensor resolutions. For dynamic scenes, we expect anyway scene details to be accessible by considering observations not only over space, with different cameras, but also over time with moving objects. This requires going beyond static per-frame reconstruction methods [40, 18, 12] and turn to temporal redundancy for detail refinement.

A number of global 4D strategies have been devised for such task, with the general strategy of globally optimizing a spatio-temporal scene representation, e.g. implicit variational [14], volumetric with convex relaxation [26] or graph cut based [23]. These robust schemes optimize over all in-

put data and hence are likely to filter out local shape details in space and time and, furthermore, they do not easily scale to long actions observed from many viewpoints.

Recently, online causal accumulation strategies based on dense TSDF representations have stood out, in particular for real-time interactive applications, where a single update volume [10] or geometric template[25, 16] is updated by globally aligning the current shape estimate to data of the incoming frame. These approaches focus on compact scenes and interactive applications, with therefore limitations on scale, topology evolution and local precision. We pursue a different and complementary objective with offline modeling of mid-scale dynamic scenes.

Remarkably, few approaches address these mid-scale scenarios and no mid-scale datasets are yet publicly available. We propose a local, non-causal and detail preserving filtering approach to this 4D reconstruction problem, with the focus on offline temporal refinement for higher accuracy. The approach fuses reliable shape information over a sliding time window by using local warps between neighboring frames. To this purpose, it relies on an implicit TSDF representation and a space discretization which adapts to the input image resolution rather than considering an implicit form over a fixed voxel grid where most cells will be empty in a dynamic mid-scale scenario.

We validate the approach on several real datasets with multiple subjects or people, where qualitative improvements are shown in terms of noise reduction and better completeness in occluded regions. We also set up a quantitative evaluation protocol using two synthetic mid-scale datasets, which we will make available to the community. A significant quality improvement is measured for our temporal integration algorithm on these datasets versus static and causal tracking strategies.

2. Related Work

Multi-view reconstruction with temporal continuity. While initially addressed on a frame-by-frame basis based on silhouette and stereo cues, *e.g.* [31], multi-view reconstruction has been variously shown to benefit from low level temporal continuity assumptions, *e.g.* by carving pairs of photoconsistent voxels across two frames [36], with global 4D hypersurface filtering [14], or by carving 4D Delaunay Triangulation-based representation of the sequence [2]. These smoothness constraints may be guided by optical flow [19] or scene flow [28, 24]. In some cases optical flow has been used to propagate stereo information across pairs of views [34], but no full window integration based on 3D warps was demonstrated as proposed. Topology constraints can be additionally enforced over a sequence, ensuring consistent extraction of thin objects (rope) [27] or ensuring a particular silhouette topology [23]. Rather than focusing on the propagation or use of such purely geometric priors, our approach leverages the propagation of observed stereo/depth data across a temporal window.

Template-based capture. The 4D capture problem is very often formulated as a template-based shape tracking and alignment problem. The template may be a laser-scanned [38, 9] or reconstructed surface, and use underlying kinematic [38], body-space [4], volumetric [9] or surface-cohesion [5] constraints to model the non-rigid deformability of the scene. While most methods track a single template for the whole sequence, thus not closely adjusting to the topological and geometric reality of the observed data at each frame, [7] builds and tracks keyframed templates which are discarded every few frames but are locally more faithful to the data. No method of this family refines the reconstructed representation as proposed.

Real-time, causal approaches. Several relevant approaches exist that tackle the problem [17, 25, 16, 10] showing how a TSDF representation can be used to accumulate passed geometry information over a static or non-rigid object, but these methods rely on a global non-rigid tracking step aligning passed data to the current frame, which is prone to accuracy and topological drift, especially in the presence of topological splits or merge and fast motion which are common in many dataset. Scalability is also an issue with large scenes due to dense volumetric reference shape representation. Our approach targets a different, offline context where scalability and precision are the main goal, achieved through implicit TSDF reprensentation, robust local propagation and geometry refinement.

Large scene reconstructions. All previous approaches address 4D reconstruction scenarios where the acquisition area is limited to a few square meters. only a handful of approaches address larger scenes, *e.g.* [6] applies TSDF depth-map fusion on large static scenes, and [15] reconstruct players in stadium events with frame-by-frame reconstruction. They do not however address temporal filtering enhancements as proposed.

3. Method Overview

Our objective is to exploit visual cues on dynamic scenes over both space and time in order to recover high precision shape models. We particularly consider mid-scale dynamic scenes which favors multi color camera apparatus as they provide flexibility in the acquisition space and time resolution. Our approach exploits temporal redundancy over a sliding time window in a sequence of multi-view frames. Within such a time window, we propagate depth cues between frames over a single shape instance. To this aim, we do not track a global shape template but use instead a local strategy that can benefit from shape regions with locally reliable shape information. Our integration framework, fig 2

considers therefore as input the multi-view color images within a time window and outputs a single high precision 3D shape mesh for that window. To this aim, the information over several frames, typically 3 to 7, is fused by alternating shape and local warp estimation as detailed in the following. In order to address the specific issues that results from mid-scale scenarios, *e.g*. heterogeneous scene coverage and wide baselines, we devise a novel method that combines stereo based dense depth map estimation with robust fusion over space and time through implicit forms.

4. Depth Map Estimation

The first step of our framework consists in building depth maps for the input images. This step is performed independently per frame. The objective here is to provide a dense coverage of the scene using a local strategy that can yield precise, though noisy, depth estimations and to leave the integration operation to a further global step based on truncated sign distance function (TSDF). The principle is to sample depths along the viewing ray of any image pixel and to keep the best potential candidate with respect to a photoconsistency measure that relies on image features. In order to increase precision and to reduce the false positives along viewing lines, we limit the sampling space using a confidence volume based on the silhouette information. Dropping out the time dimension temporarily to simplify notations, we assume we are given a set of N images $\{I_i\}_{i=1}^N$ observed with a set C of calibrated cameras with known projections $\{\pi_i\}_{i=1}^N$ and centres $\{c_i\}_{i=1}^N$. We assume we are also given a set of silhouettes $\{\Omega_i\}_{i=1}^N$, possibly imprecise.

4.1. Confidence Volume

The silhouettes $\{\Omega_i\}_{i=1}^N$ define, by extrusion, a 3D visual hull that is assumed to contain the observed object. In practice, silhouettes are prone to various errors such as holes or missing parts and do seldom guarantee this containment property with the visual hulls. In addition, our objective is primarily to reduce the search space along viewing rays to segments that are likely to intersect the object surface more than exactly locate the visual hull. Consequently we define the confidence volume V as:

$$V = \{x \in \mathbb{R}^3 : \exists^{>\alpha} i \, (\pi_i(x) \in I_i) \\ \wedge \exists^{>\beta} i \, (\pi_i(x) \in \Omega_i)\}, \tag{1}$$

that is the locus of points in $\mathbb{R}^3$ for which there exist $i > \alpha$ images where they project and $i > \beta$ silhouettes to which they belong. α, β are two user defined constants that restrict weakly supported depth predictions with α and enable predictions away from the exact visual hull when $\beta < \alpha$. Intuitively, V is a dilated version of the visual hull in the space region seen by at least α images, as shown in fig 3.

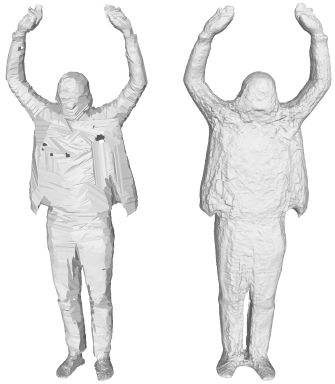

Figure 3. Left: the Confidence Volume with $\alpha = \beta = 54$, equivalent to the Visual hull with the 54 cameras that see the subject; Right: the Confidence Volume with $\alpha = \beta = 10$.

4.2. Photoconsistency measure

In order to predict depth along pixel viewing rays we make use of a photoconsistency measure evaluated along the ray and based on pairwise photometric discrepancy. While Normalized Cross Correlation has been extensively used over the past [12, 27, 11, 39], recent advances in image descriptors have demonstrated the benefit of gradient based descriptors, such as SIFT, GLOH, DAISY [21, 22, 33], especially with noisy photometric information. We chose DAISY as it experimentally gives the best results in our context.

For a point $x \in \mathbb{R}^3$ and given two images I_i and I_j, the pairwise photometric discrepancy $g_{i,j}(x)$ at x is given by the Euclidean distance between the two descriptors D_i and D_j of the point's projection in the images:

$$g_{i,j}(x) = (D_i(\pi_i(x)) - D_j(\pi_j(x)))^2. \tag{2}$$

The photoconsistency measure $\rho_i(x)$ at x, given all the images, is then computed as a normalized robust vote of the image descriptors $D_j(\pi_j(x))$ at x that are similar to $D_i(\pi_i(x))$. In contrast to [39], who consider only local minima of the pairwise discrepancy $g_{i,j}(x)$ and interpolate them, we consider all the discrepancy values. This is based on our observations that, in the mid-scale context, surface points are less likely to define local minima of $g_{i,j}(x)$ than in the small-scale case that presents short baselines. Hence our photoconsistency measure $\rho_i(x)$ is:

$$\rho_i(x) = \sum_{j \in C_i} \bar{\omega}_j W(g_{i,j}(x)), \tag{3}$$

where: the normalized values $\bar{\omega}_j$ of $\omega_j = \cos(\theta_{ij})$ weights camera contributions around camera i using the angle θ_{ij} between the optical axes of camera i and j; C_i is the subset of cameras j such that $\omega_j > 0.7$; and $W()$ is a robust voting function, a Gaussian Parzen-Window in the descriptor space in our experiments. Note that 1 is therefore the best score

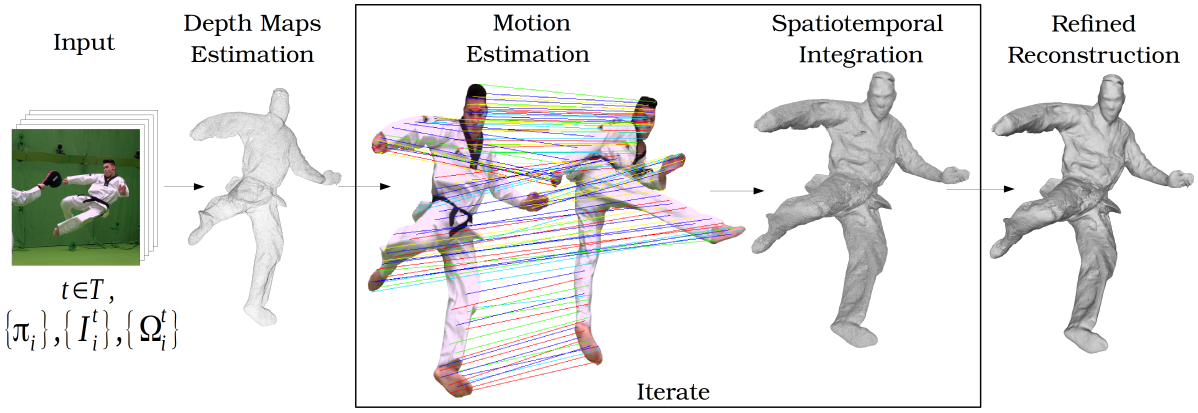

Figure 2. Spatiotemporal refinement framework.

$\rho_i(x)$ when all cameras in C_i present the exact same image descriptors at x and 0 the worst.

The above photoconsistency measure implicitly assumes Lambertian surfaces and while robust to a certain extent to specularities it can still fail when strong highlights occur. Also regarding occlusions, we expect ρ to present local maxima where rays intersect the surface even in the presence of occlusions. In order to reinforce this assumption, we restrict the search along viewing rays within a range close to the surface using the confidence volume (1) as explained below.

4.3. Depth Prediction

For each pixel in every silhouette, depth is predicted along the viewing line using maxima of the photoconsistency measure ρ introduced before. As mentioned before, the photometric information can often be unreliable in mid-scale scenarios. In order to prevent false detections of maxima far from the surface, we adopt a conservative scheme where search for maxima along the viewing rays start from the confidence volume and stop when the accumulated photoconsistency reaches a threshold, hence limiting surface penetration along rays. In spirit, this is similar to [26] who define and integrate interior probabilities along rays using however a photoconsistency measure taken from [39] (see the discussion on photoconsistency measures in the previous paragraph).

More precisely, the best depth candidate d_i^p along ray $r_i(p,d)$ leaving camera i through pixel p is determined as:

$$d_i^p = \begin{cases} d_V(p) \text{ if } \max\limits_{d\in[d_V(p),d_{max}]} \rho_i(r_i(p,d)) < \tau_{photo}, \\ \operatorname{argmax}\limits_{d\in[d_V(p),d_{max}]} (\rho_i(r_i(p,d))) \text{ otherwise.} \end{cases} \tag{4}$$

Where $d_V(p)$ is the first depth value along $r_i(p,d)$ inside the confidence volume V, τ_{photo} a minimum photoconsistency value below which we fall back to silhouette information and the confidence volume, and d_{max} the search limit such that:

$$\int_{x=d_V(p)}^{d_{max}} \rho_i(r_i(p,x))\mathrm{d}x \leq \rho_{max} \tag{5}$$

To speed up depth map computation and add some spatial consistency, we first perform super pixel clustering on images using SLIC [1] and select a few random samples per super pixel. An exhaustive search is performed for these sample pixels in order to provide an approximation for depths within the super pixel. Other pixel depths in the super pixel are then computed around this first approximation $\bar{d}$.

As a post-processing step, bilateral filtering accounting for spatial, color and photoconsistency proximity is performed over depth maps. It efficiently filters out outliers with little impact on the computation burden, which motivates our choice in a 4D dynamic context.

5. Shape Estimation

Given the depth maps $\{d_i^t\}$ estimated for all cameras i and all frames t, we can now fuse depth information over space and time to recover the shape surface mesh S^k at any time instant k. While we consider all cameras in the fusion, we limit the frames taken into account to a temporal window around k, typically 3 to 7 frames in our experiments, within which required shape motion information can be obtained with precision. In order to propagate reliable depth cues between frames, our approach seeks for local regions with consistent displacements and high photoconsistencies. This local strategy better prevents the propagation of wrong depth cues which occurs when a global strategy, such as template tracking, is used. Given a temporal window, we assume that each frame t, within the temporal window, corresponds to an instance of the reference shape S^k deformed

with respect to a 3D motion field W_k^t, with no topology assumption. The approach consists then in iterating the following steps:

1. For all frame k:
   (a) Given inter frame volumetric motions $\{W_k^t\}$ merge all the time window depth maps, warped using $\{W_k^t\}$, into a 3D implicit form.
   (b) From the implicit form estimate the 3D mesh S_k.
2. Given the $\{S^k\}$ estimate the motion fields $\{W_k^t\}$.

To initialize the process, we perform spatial integration only in the above step 1 at the first iteration. The two steps are then repeated a few times, typically 3 in our experiments.

5.1. Spatial Integration

To introduce our integration scheme, we first consider a single frame and the spatial integration of the depth maps d_i for all cameras at that frame. Following several works[8, 17, 25] with a similar objective but in different contexts, *e.g.* small-scale, we fuse all the depth maps into a 3D implicit form and take benefit of the Truncaded Signed Distance Function (TSDF) strategy for that purpose. Our motivation for the TSDF comes from its ability to naturally handle arbitrary depth maps arising from different cameras in addition to different time steps, as shall be dealt with in further sections.

For a point $x \in \mathbb{R}^3$, the truncated signed distance $TD(x) \in \mathbb{R}$ to the surface is defined as the weighted average of all camera predictions $F_i(x), i \in C$:

$$F_i(x) = \begin{cases} \min(\mu, \eta(x)) & iif\ \eta(x) \geq -\mu, \\ \emptyset & \text{otherwise}, \end{cases} \qquad (6)$$
$$\eta(x) = d_i(\pi_i(x)) - \|c_i - x\|,$$

and:

$$TD(x) = \frac{\sum\limits_{i \in C_x} \rho_i'(x) F_i(x)}{\sum\limits_{i \in C_x} \rho_i'(x)}, \qquad (7)$$

where $C_x = \{i \in C : F_i(x) \neq \emptyset\}$ and ρ_i' the photoconsistency measure (3) of the estimated depth along the ray passing through x. If d_i is undefined at x, *e.g.* x is outside the camera visibility domain, then camera i does not contribute to the TSDF. When no camera contributes at x but x is inside the confidence volume V then it is considered as inside, *i.e.* $TD(x) < 0$. Note that contributions are weighted by the normalized photoconsistency measure which means that when cameras disagree about the photoconsistency at x, cameras with higher measures have an increased impact whereas cameras with low photoconsistency measures only marginally impact the reconstruction.

5.2. Spatiotemporal Integration

In order to extend the previous spatial integration to the time domain, we now consider several frames over a temporal window $T = [k - n/2, k + n/2]$ of size n around frame k. In essence, the temporal integration consists then in adding to the TSDF (7) depth contributions from the neighboring frames; using to this aim the estimated motion fields $W_k^t : \mathbb{R}^3 \to \mathbb{R}^3$ that map frame k to frame l (as detailed in Sec. 5.4). As mentioned earlier, these contributions should be weighted by the confidence λ we have in the estimated local motion in addition to their photoconsistencies ρ. We define therefore the integrated implicit form $\overline{TD}_k : \mathbb{R}^3 \to \mathbb{R}$ of the observed shape at frame k as:

$$\overline{TD}_k(x) = \frac{\sum\limits_{t \in T} \lambda_k^t(x_k^t) \sum\limits_{i \in C_x^t} \rho_i'^t(x_k^t) F_i^t(x_k^t)}{\sum\limits_{t \in T} \lambda_k^t(x_k^t) \sum\limits_{i \in C_x^t} \rho_i'^t(x_k^t)}, \qquad (8)$$

$$x_k^t = x + W_k^t(x). \qquad (9)$$

where $C_x^t = \{i \in C : F_i^t(x) \neq \emptyset\}$ and λ_k^t, ρ_i^t, d_i^t and F_i^t are respectively the motion confidence (Sec. 5.4), the photoconsistency measure (Sec. 4.2), the depth prediction (Sec. 4.3) and the truncation function (Sec. 5.1) at frame t.

5.3. Shape Mesh Generation

From the implicit form of the shape detailed in the previous section, we can extract the 3D shape mesh at frame k as the zero level set of the associated implicit function $\overline{TD}_k(x)$. A vast majority of methods consider the Marching Cube [20] (MC) approach for that purpose [12, 17, 26]. Although MC would also work in our case we consider instead a different strategy that addresses some of the limitations of MC: MC is based on a regular discretization of the space and hence dilutes precision inside the shape, unless a specific strategy such as subdivision is applied at the surface; MC is not guaranteed to provide manifold meshes, again unless specific and costly additional steps are performed. In contrast we built on recent works on Voronoi Tesselation [41] showing that better precision can be obtained with discretizations of shapes instead of space. We devise a simple yet efficient version of Voronoi Tesselation that specifically accomodates multi-view capture scenarios. The main steps of the algorithm are as follows:

1. Sample points inside the implicit form defined by the TSDF. This is achieved by randomly selecting pixels in all images and computing the point, along each pixel rays, inside but close to the surface according to the TSDF. The process is iterated until a user defined number of 3D points is reached.

2. Determine the Voronoi diagram: given the points inside the shape surface, a Voronoi diagram of this set of points is computed.

3. Clip the Voronoi diagram with the zero level set of the TSDF. This operation extracts the intersection of the Voronoi cells with the surface.

In the above strategy, sampling points close to the surface, and originating from image viewpoints, ensures that the 3D discretization is denser on the surface than inside the volume and also denser on surface regions observed by images. The latter enables more precision to be given to surface regions for which more image observations are available.

5.4. Motion Estimation

Considering two meshes S^k and S^l at frames k and l, we want to estimate the volumetric motion field $W_k^l : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ that maps S^k into S^l. Recall that our objective is to improve shape estimations, hence we do not necessarily need the complete shape motion, as when tracking or estimating scene flow. Instead, we seek for reliable sparse motion information in surface regions where temporal integration will therefore benefit to the shape reconstruction. Thus, the estimated 3D motion fields needs not fully reproduce the true motion, yet be equipped with confidence measures that identify valid motion and allow to neglect the surface cues associated with invalid motions when propagating information between frames.

Various methods have been proposed to recover motion information on moving shapes. Depending on the prior assumption on the motion model they range from weakly constrained models with scene flow [35] to locally rigid models with ARAP[3] strategies, as for instance with Kinect and Dynamic fusion[17, 25] or [7] and, at the other end of the spectrum, to stronger priors with articulated models and skinning animations as in [37].

In our context, as we do not seek for a complete and flexible motion model we will favor local constrained strategies. In addition, since we consider mid-scale and dynamic scenes, large displacements can occur between frames which advocates for sparse but robust matching. We therefore opt for 3D features to provide robust 3D matches that will be progressively densified over the alternate iterations of shape and motion estimations. We use MeshHog [42] to detect and match 3D features as it demonstrates a good tradeoff between robustness, completeness and accuracy among other efficient methods such as heat kernel [32] or Harris 3D [30].

Let $\{M^k\}$ be the set of corresponding pairs of 3D features between S^k and S^{k+1} obtained with MeshHog and $m \in \{M^k\}$ such a pair. We attach to m a confidence measure λ_m that favors regions with dense and coherent matches. To this aim, the k-nearest neighbors m_j of m in $\{M^k\}$ are first computed. Let δ_m^j be the discrepancy between the displacements vectors associated to m and m_j. λ_m is then the median of the j values $\mathcal{G}(\delta_m^j)$, where $\mathcal{G}$ is a Gaussian kernel. This conservative strategy favors small regions on S^k where m and its neighbors present similar displacements vectors. As more matches will be added over iterations, this can be seen as a growing strategy that progressively extends the motion field around regions where consistent displacements are found over iterations.

Given the corresponding pairs of MeshHog feature $m \in \{M^k\}$, their displacement vectors $\{T_m\}$ from S^k to S^{k+1} and their confidences λ_m, we define the forward motion field $W_k^+ : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ and its confidence $\lambda_k^+ : \mathbb{R}^3 \rightarrow \mathbb{R}$ as:

$$\begin{aligned} W_k^+(x) &= \sum_{m \in \{M^k\}} \lambda_m \, \mathcal{G}_m(x) \, T_m, \\ \lambda_k^+(x) &= \frac{1}{|M^k|} \sum_{m \in \{M^k\}} \lambda_m \, \mathcal{G}_m(x) \end{aligned} \tag{10}$$

where $\mathcal{G}_m()$ is a Gaussian kernel that weights the contribution of m with respect to the spatial distance between x and the feature of m on S^k. The backward motion fields $W_k^-(x)$ that maps S^k onto S^{k-1} is defined in a similar way using MeshHog features between S^k and S^{k-1}. The motion field W_k^l and its confidence λ_k^l are then defined as:

$$W_k^l(x) = \begin{cases} \sum\limits_{t \in [k, l-1]} W_t^+(x) & if \; k < l, \\ \sum\limits_{t \in [k, l+1]} W_t^-(x) & if \; k > l, \\ 0 & if \; k = l, \end{cases} \tag{11}$$

$$\lambda_k^l(x) = \begin{cases} \prod\limits_{t \in [k, l-1]} \lambda_t^+(x) & if \; k < l, \\ \prod\limits_{t \in [k, l+1]} \lambda_t^-(x) & if \; k > l, \\ 1 & if \; k = l, \end{cases} \tag{12}$$

6. Results

In order to demonstrate the benefit of time integration to recover dynamic scene models we conducted different experiments. First, quantitative results were obtained to evaluate how temporal integration improves shape reconstruction. To this purpose, and since dynamic multi-view benchmarks are not available yet, we created a dynamic dataset equipped with ground truth data on geometry and appearance. Then, qualitative results on real data were also obtained to illustrate that temporal integration enhances reconstructed shapes quality. The code and all the data used in the following experiments will be made available to the community.

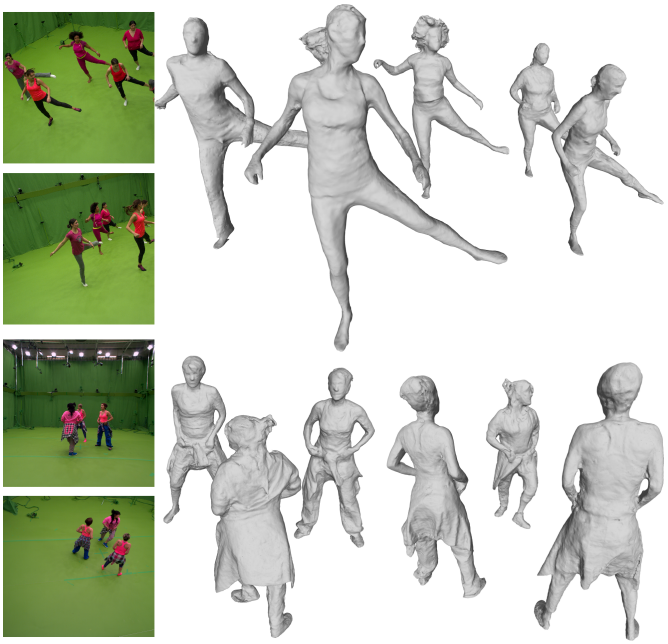

Figure 4. Examples of challenging dynamic mid-scale datasets, and our reconstructions.

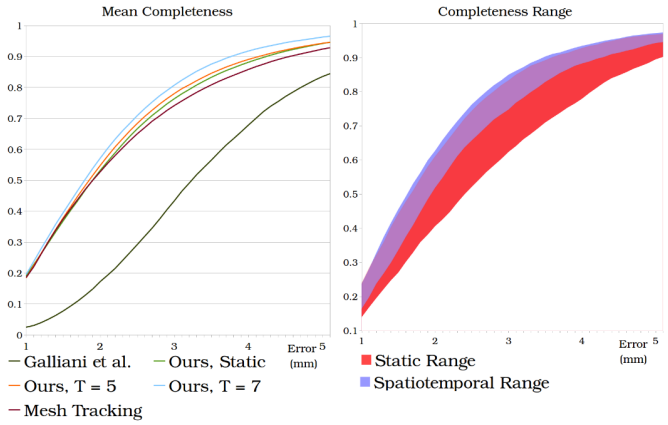

Figure 5. (*left*) Mean completeness comparison between [13] and our reconstructions on 10 frames of the synthetic sequence, (*right*) Min and max values of completeness on 20 frames of the synthetic sequence, time window T = 7, iterations = 3.

6.1. Synthetic Data

Dataset Multiple benchmarks addressing the Static Multi-View Stereo problem, *e.g*. Middleburry [29] or DTU Robot Image Dataset [18], were already made available online. However, to the best of our knowledge, none exists for the dynamic case with surfaces evolving over time. Hence, we built an evaluation dataset with the objective to be as close as possible to real situations with real data while having ground truth information. It should be noticed that such ground truth data is of interest in a context larger than shape recovery and can contribute to tracking or appearance modeling evaluations. The data consists of procedurally generated surfaces, typically clothes, added to real captured data, typically body shapes, for which tracking over time sequences are available. Its main features are:

- The synthetic image generation set-up is similar to real multi camera platforms.
- Underlying shapes and their motions are real captured data and replicate therefore real dynamic situations.
- Local shape deformations are generated and can simulate clothes or any other type of deformation.
- Appearances are generated as well and can yield various effects with low to high contrast textures, specular surfaces, color diffusion, motion blur among others.

Evaluation Given the ground truth data mentioned above we evaluated quantitatively shape reconstructions using standard measures in the field [29, 18], *i.e*. accuracy and completeness. Static and refined reconstructions were performed on a 20 frames synthetic sequence with local cloth-ing deformations, observed by 60 cameras, with a capture volume of approximately 8mx4mx6m.

Figure 5 demonstrates how the mean completeness (ratio of ground truth points closer to the reconstruction than a given error) over 10 frames increases with temporal window of sizes 1, 5 and 7. In order to evaluate the benefit of our local propagation strategy, we also performed comparisons with a strategy based on global surface tracking between adjacent frames [5] very similar in spirit to the tracking method employed in [10]. The global motion was then fed in our temporal integration pipeline similarly to our local strategy. All experiments were conducted using the same set of parameters. Figure 5 shows that such global strategy (mesh tracking in the figure) performs worse than our local strategy or even than static strategies (*i.e*. single frame). This is confirmed on real data in Figure 7 where the mesh tracking based strategy is prone to erroneous and imprecise estimations, leading to an oversmoothed results.

For the sake of completeness, we also compare to [13], top ranked static Multi-View Stereo Reconstruction method on the DTU dataset [18]. While the accuracy comparison would be unfair since [13] does not take silhouettes into account and hence produces points outside the visual hull, we believe that the completeness that measures how close the ground truth is to the reconstructed surface is on the other hand informative.

This figure also shows min and max completeness values over 20 frames of the synthetic sequence. It shows that the temporal integration impact significantly more the min completeness. It is worth noticing that at approximately the pixel resolution, roughly 3mm here, the min completeness is increased by around 15% with the temporal integration.

6.2. Real Data

We also tested our method on different dynamic multi-camera sequences, containing multiple subjects. Every sequence was captured with 68 calibrated RGB cameras (2048x2048) with focal lengths between 8 and 28 mm. Some examples of dynamic mid scale scenes and spatiotemporally refined surfaces are shown in Figure 4.

Figures 6 and 1 depict input images, our reconstructions and the temporal improvement for the former. In addition, Figure 6 shows that the temporal refinement preserved details that are filtered out by a spatial smoothing technique (Laplacian Smoothing).

Figure 7 shows an example of temporal integration with a global mesh tracking strategy, as explained previously. Even though the standing subject is quite well reconstructed, such global approach fails in the case of fast motion and strong topology noise. The temporal integration with a global template motion makes the moving subject's surface noisier and fast moving parts are missing. The thin surfaces such as the belt and the outfit also tend to suffer from the tracking inaccuracies propagated through time and are not correctly recovered with the global mesh tracking strategy.

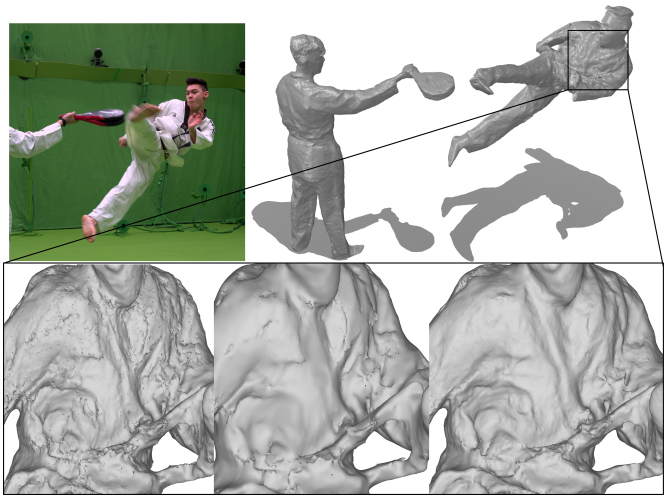

Figure 6. (*top*) An input image and our refined reconstruction. (*bottom*) A close-up view on the model, showing the static reconstruction (*left*), spatially smoothed (*middle*) and our temporal details refinement (*right*). Best viewed magnified

Our C++ multithreaded implementation runs as follows on a 16-core Xeon 3.00GHz PC, 32 Gb RAM and with 68 4Mpixels cameras: 5-20 min/frame to build the implicit TSDF, depending on total number of silhouette pixels; 5 min/frame for motion estimation; 5 min/frame for the surface extraction, for a final mesh of 3M faces. A GPU implementation could be considered as extension for significant speedup.

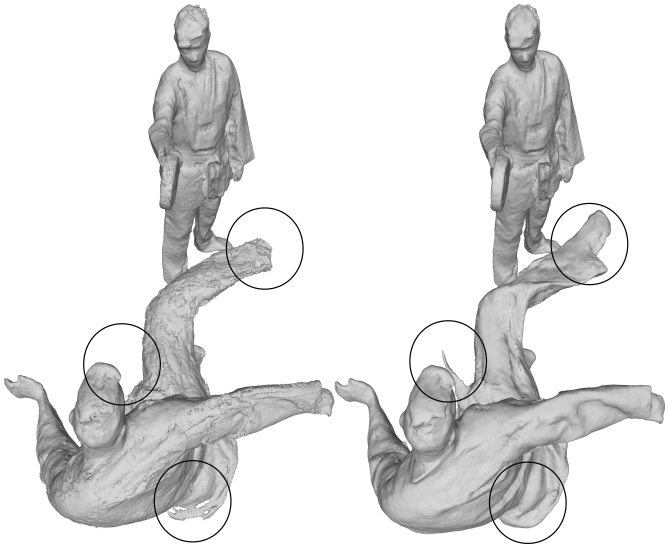

Figure 7. Spatiotemporal integration using motion estimation based on global surface tracking (*left*) and using the proposed local detection approach (*right*).

7. Conclusion

We presented a framework for spatiotemporal integration for surface reconstruction refinement, especially efficient on challenging mid-scale dynamic scenes captured with multi-camera systems. Our approach improves over classic per frame reconstruction, giving smoother and more accurate reconstructions, especially in strongly occluded areas, by propagating photometric cues through time, accumulating implicit forms, and extracting the surfaces using a space discretization attached to the observed shape. A seed growing strategy method is introduced to gradually estimate the motion of the dynamic scene, alternating between a safe temporal accumulation of observations and motion re-estimation. Comparisons against a state of the art MVS methods demonstrate the effectiveness of our method to recover surfaces in standard static cases, but also for mid-scale dynamic data, as validated with a proposed data-set, containing synthetic and real scenes.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. SLIC Superpixels. Technical report, EPFL, 2010.

[2] E. Aganj, J. Pons, F. Ségonne, and R. Keriven. Spatio-temporal shape from silhouette using four-dimensional delaunay meshing. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[3] M. Alexa, D. Cohen-Or, and D. Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pages 157–164, 2000.

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016.

[5] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic Deformable Surface Tracking From Multiple Videos. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV 2010 - 11th European Conference on Computer Vision*, volume 6314, pages 326–339, Heraklion, Greece, Sept. 2010. Springer.

[6] J. Chen, D. Bautembach, and S. Izadi. Scalable real-time volumetric surface reconstruction. *ACM Trans. Graph.*, 32(4):113:1–113:16, 2013.

[7] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. G. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4):69, 2015.

[8] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312, 1996.

[9] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 98:1–98:10, New York, NY, USA, 2008. ACM.

[10] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi. Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4):114:1–114:13, July 2016.

[11] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[13] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 873–881, 2015.

[14] B. Goldlücke and M. A. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *CVPR (1)*, pages 350–355, 2004.

[15] J.-Y. Guillemaut, J. Kilner, and A. Hilton. Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. In *ICCV*, pages 809–816. IEEE Computer Society, 2009.

[16] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 362–379, 2016.

[17] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. W. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 559–568, 2011.

[18] R. R. Jensen, A. L. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 406–413, 2014.

[19] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[20] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*, pages 163–169, 1987.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[22] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*, pages 257–263, 2003.

[23] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. *CoRR*, abs/1603.03381, 2016.

[24] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *International Journal of Computer Vision*, 47(1-3):181–193, 2002.

[25] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2015, Boston, MA, USA, June 7-12, 2015*, pages 343–352, 2015.

[26] M. R. Oswald and D. Cremers. A convex relaxation approach to space time multi-view 3d reconstruction. In *ICCV Workshop on Dynamic Shape Capture and Analysis (4DMOD)*, 2013.

[27] M. R. Oswald, J. Stühmer, and D. Cremers. Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 32–46, 2014.

[28] J. Pons, R. Keriven, and O. D. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

[29] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 519–528, 2006.

[30] I. Sipiran and B. Bustos. A robust 3d interest points detector based on harris operator. In *Eurographics Workshop on 3D Object Retrieval, Norrköping, Sweden, May 2, 2010, Proceedings*, pages 7–14, 2010.

[31] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007.

[32] J. Sun, M. Ovsjanikov, and L. J. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *Comput. Graph. Forum*, 28(5):1383–1392, 2009.

[33] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[34] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1709–1716, 2009.

[35] S. Vedula, S. Baker, P. Rander, R. T. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV*, pages 722–729, 1999.

[36] S. Vedula, S. Baker, S. M. Seitz, and T. Kanade. Shape and motion carving in 6d. In *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000), 13-15 June 2000, Hilton Head, SC, USA*, pages 2592–2598, 2000.

[37] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3), 2008.

[38] D. Vlasic, P. Peers, I. Baran, P. E. Debevec, J. Popovic, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.*, 28(5):174:1–174:11, 2009.

[39] G. Vogiatzis, C. H. Esteban, P. H. S. Torr, and R. Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2241–2246, 2007.

[40] H. Vu, R. Keriven, P. Labatut, and J. Pons. Towards high-resolution large-scale multi-view stereo. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1430–1437, 2009.

[41] L. Wang, F. Hétroy-Wheeler, and E. Boyer. On volumetric shape reconstruction from implicit forms. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 173–188, 2016.

[42] A. Zaharescu, E. Boyer, and R. Horaud. Keypoints and local descriptors of scalar functions on 2d manifolds. *International Journal of Computer Vision*, 100(1):78–98, 2012.