SobolevFusion: 3D Reconstruction of Scenes Undergoing Free Non-rigid Motion

Miroslava Slavcheva [1,2] Maximilian Baust [1*] Slobodan Ilic [1,2]

[1] Technische Universität München [2] Siemens Corporate Technology

Abstract

We present a system that builds 3D models of non-rigidly moving surfaces from scratch in real time using a single RGB-D stream. Our solution is based on the variational level set method, thus it copes with arbitrary geometry, including topological changes. It warps a given truncated signed distance field (TSDF) to a target TSDF via gradient flow. Unlike previous approaches that define the gradient using an L^2 inner product, our method relies on gradient flow in Sobolev space. Its favourable regularity properties allow for a more straightforward energy formulation that is faster to compute and that achieves higher geometric detail, mitigating the over-smoothing effects introduced by other regularization schemes. In addition, the coarse-to-fine evolution behaviour of the flow is able to handle larger motions, making few frames sufficient for a high-fidelity reconstruction. Last but not least, our pipeline determines voxel correspondences between partial shapes by matching signatures in a low-dimensional embedding of their Laplacian eigenfunctions, and is thus able to reliably colour the output model. A variety of quantitative and qualitative evaluations demonstrate the advantages of our technique.

1. Introduction

The abundance of affordable RGB-D cameras in recent years triggered the creation of a variety of excellent real-time methods for 3D mapping and tracking from a single stream [22, 23, 29, 31, 32, 33, 49]. Nowadays depth sensors are being integrated into new generations of mobile phones, whose limited computational resources call for new solutions. One major challenge is the reduced frame rate, which can be as low as 5 frames per second on a Tango tablet [17]. While static reconstruction methods have been successfully ported to mobile devices [21], when it comes to dynamic scenes, algorithms will have to cope with larger frame-to-frame motions, which is one of the goals of this paper.

DynamicFusion [31] is the breakthrough work that first performed real-time 3D reconstruction of a non-rigid scene using a single depth camera. Several follow-ups improved its tracking via additional constraints, such as colour features [18], albedo [15] or human skeleton [52]. While showing ever-improving visual quality, all of these methods only demonstrate examples of relatively contrived movements. Even state-of-the-art multi-view systems [9] utilize extremely high frame-rate cameras of up to 200 fps [12] to ensure that frame-to-frame motion is minimal.

KillingFusion [40] is the only single-stream approach that has shown capture of more free movements to date. It warps an input TSDF towards the current canonical reconstruction via a variational formulation that estimates a flow field. However, the underlying gradient flow is based on an L^2 inner product, which is known to be susceptible to local minima [45]. In order to counteract this issue and stabilize the level set evolution, KillingFusion employs a combination of regularizers, which are difficult to balance and thus result in over-smoothing and loss of high-frequency details. Here we propose to define the gradient flow in the Sobolev space H^1 [30, 45]. It acts as a preconditioner that favours consistent motion and features a desirable coarse-to-fine evolution of the TSDF. This reduces

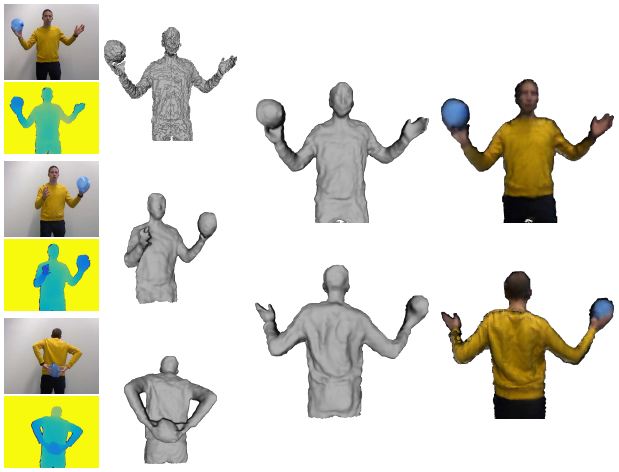


Figure 1. **SobolevFusion example.** Our method reconstructs scenes containing multiple non-rigidly interacting agents, captured with a single RGB-D sensor.

*M. Baust is now with Konica Minolta Laboratory Europe (KMLE).

the risk of getting trapped in local minima without changing the global optimum [44]. Therefore, it lets us define an energy of reduced complexity that is faster to evaluate and yields more detailed reconstructions. Moreover, thanks to improved convergence, SobolevFusion can capture even larger motion, thus only several scans are sufficient to build a realistic 3D model.

As the proposed approach is based on the variational level set method [53], it can handle topological changes, but preserving correspondence information is more challenging than for mesh-based techniques [36]. This usually limits the applicability to tasks such as texture transfer and character animation. Therefore we take inspiration from spectral techniques for matching over voxel representations [28, 38]. Using a low-dimensional embedding of the eigenspace of a shape's Laplacian matrix, the alignment problem is reduced to matching eigenfunction signatures [28]. However, as we are dealing with partial shapes from noisy data, we keep only high-confidence matches. Thus we only obtain a set of sparse correspondences per frame, but can reliably texture the final canonical reconstruction.

To sum up, we propose a variational non-rigid fusion technique, called SobolevFusion, which:

- is based on Sobolev gradient flow, allowing for a more straightforward, faster to compute energy that preserves geometric details without over-smoothing;
- handles topological changes and large motion, thus requiring only a few views to build a model;
- can estimate voxel correspondences and colour the reconstruction.

2. Related Work

Most real-world scenes consist of agents that interact with each other and their surroundings. Reconstructing them is a challenging task due to its high dimensionality. Compelling state-of-the-art capture systems constrain the problem through the use of multiple cameras [4, 6, 8, 10, 20] or template models [3, 16, 55], requiring custom set-ups and recording studios. Here we address the scenario of a single RGB-D camera, which is more convenient for the user.

Dynamic reconstruction Template-free methods for non-rigid fusion using a single depth sensor have been on the rise since 2015 with the development of the offline bundle adjustment scheme of Dou *et al*. [11] and the first real-time solution for simultaneous surface tracking and reconstruction, DynamicFusion [31]. A line of research improved on it, including VolumeDeform [18] which combines its dense depth correspondences with sparse SIFT features, and the integration of surface albedo constraints by Guo *et al*. [15]. However, the examples shown in these publications mostly contain slow motion and no changing topology.

Large motion KillingFusion [40] tackles the problem from another perspective, whereby instead of extracting a mesh from the cumulative model for correspondence estimation, it stays within the TSDF representation and warps it incrementally. As level sets inherently handle topological changes and can recover from large distances, examples on less constrained motion have been shown.

Similar to most approaches derived from the variational level set method, the gradient flow used for warping is defined via an L^2-type inner product [34, 35, 53]. Although widely used, it assumes a metric that may lead to slow convergence and sub-optimal solutions [45]. Techniques to stabilize the evolution include re-initialization or additional regularizers imposing the level set property of unit gradient magnitude [26], as done by KillingFusion. However, it does not hold strictly in the discrete case, and is not valid at the border of voxel truncation, causing over-smoothing effects.

Gradient flow in the Sobolev space H^1 has been shown to have superior performace without changing the global minimum [45]. Re-casting the notion of gradient in this way has a pre-conditioning effect that induces flow with coarse-to-fine behaviour which first evolves lower-frequency components and is thus less susceptible to local minima [5, 44]. The concept was developed in the context of numerical solutions of PDEs. We refer the reader to the book of Neuberger [30] for a mathematical introduction. It has been applied for segmentation [1, 14, 45], registration [44, 46, 54] and sharpening [5] of 2D images or complete 3D volumes in medical imaging. Here we propose to employ it for the profoundly different task of incremental 3D reconstruction from depth images. As its regularity properties will permit us to define an energy functional with fewer terms, we expect faster processing, in addition to the discussed improved convergence and better preservation of geometric details.

Voxel correspondence A major limitation of approaches based on level set evolution is their inability to track correspondences [36, 50]. One possibility to recover them is to convert the resulting shape to a mesh and use spectral matching techniques [2, 19], which utilize the fact that the graph Laplacian of a shape is invariant to isometric deformations [25, 37]. As mesh extraction would entail temporal overhead, our aim is to determine correspondences between an initial TSDF and its warped counterpart. To this end we follow an approach similar to that of Mateus *et al*. [28], who deal with shapes represented as voxel sets. They first find a lower-dimensional embedding of the Laplacian spectrum, then determine an ordering of the eigenvalues by matching eigenfunction signatures, and finally reduce the correspondence estimation problem to rigid alignment in the embedded space. However, their approach is applied only to whole shapes and does not fit into real-time constraints. Therefore we propose a modified strategy over TSDFs of incomplete shapes that only keeps the most likely matches.

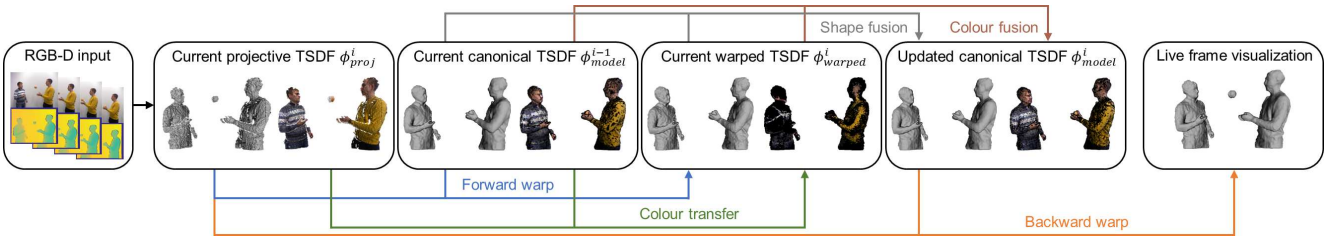


Figure 2. **SobolevFusion pipeline.** Given an input RGB-D pair, we first generate its projective TSDF ϕ^i_{proj} from the current camera pose estimate. Next, we warp it towards the canonical model TSDF ϕ^{i-1}_{model}, obtaining ϕ^i_{warped}. Afterwards we optionally estimate voxel correspondences between ϕ^i_{proj} and ϕ^i_{warped} in order to transfer colour to the warped TSDF. Then we fuse ϕ^i_{warped} into the canonical model, obtaining its updated state ϕ^i_{model}. Finally, we run a backward warp from ϕ^i_{model} to ϕ^i_{proj} to visualize the live frame to the user.

3. Overview

In the following we briefly describe our mathematical notation and outline the proposed SobolevFusion approach.

3.1. Mathematical Preliminaries

Our system takes an RGB-D stream consisting of pairs (I^i_{RGB}, I^i_D), where i is the frame index, I_{RGB} is the 3-channel colour image and I_D is the aligned depth map. We assume a calibrated camera and a projection function $\pi\colon \mathbb{R}^3 \mapsto \mathbb{N}^2$ from 3D coordinates to pixels.

We discretize the pre-defined bounding volume into cubic voxels of a selected side length. They are indexed by integer tuples $(x, y, z) \in \mathbb{N}^3$. Let $(X, Y, Z) \in \mathbb{R}^3$ be the coordinates of the respective voxel's center in 3D space. A single RGB-D frame allows the generation of a projective TSDF $\phi\colon \mathbb{N}^3 \mapsto \mathbb{R}$. We follow the traditional scaling and truncation scheme [39]:

$$d(x, y, z) = I_D(\pi(X, Y, Z)) - Z\,, \tag{1}$$

$$\phi(x, y, z) = \begin{cases} sgn(d(x, y, z)) & \text{if } |d(x, y, z)| \geq \delta\,, \\ d(x, y, z)/\delta & \text{otherwise}\,, \end{cases} \tag{2}$$

$$\omega(x, y, z) = \begin{cases} 1 & \text{if } d(x, y, z) > -\eta\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{3}$$

Here d is the directional signed distance, which is truncated to the interval $[-1, +1]$ to disregard voxels that are far away from the surface. In practice we set the responsible parameter δ to 5-10 times the voxel size. The parameter η determines the expected object thickness and is set to 2-3 voxels. Voxels outside the object and within this thickness receive a confidence weight ω of 1, while non-observed ones get 0.

TSDFs from multiple views are fused together via the weighted averaging scheme of Curless and Levoy [7], resulting in a true (not projective) TSDF.

Finally, we will be estimating a vector flow field $\Psi = (U, V, W)\colon \mathbb{N}^3 \mapsto \mathbb{R}^3$ of the same resolution as the TSDFs. U, V and W denote its x-, y- and z-components respectively, each of which is a scalar grid $\mathbb{N}^3 \mapsto \mathbb{R}$. We denote the vector applied at voxel (x, y, z) by (u, v, w).

3.2. SobolevFusion Pipeline

Our proposed pipeline is displayed in Figure 2. Given an existing state of the cumulative model ϕ^{i-1}_{model} and an incoming RGB-D pair (I^i_{RGB}, I^i_D), we iteratively estimate a deformation field that warps the projective TSDF ϕ^i_{proj} generated from I^i_D towards ϕ^{i-1}_{model}, resulting in ϕ^i_{warped}, using the Sobolev deformation scheme described in Section 4. We then estimate voxel correspondences between the initial and warped TSDFs in order to transfer colour from ϕ^i_{proj} to ϕ^i_{warped}, as explained in Section 5. Then we fuse ϕ^i_{warped} into the global model, obtaining its new state ϕ^i_{model}. Finally, we run a backward deformation from ϕ^i_{model} towards ϕ^i_{proj} in order to provide a live update to the user.

4. Sobolev 3D Reconstruction

Here we describe our variational model for non-rigid reconstruction, as well as how the concept of Sobolev gradient flow is employed for computing a minimizer of this model.

4.1. Deformation Energy

As a new RGB-D frame is acquired and we estimate the approximate camera pose, we generate the corresponding projective TSDF ϕ_{proj}. Next, we warp it towards the canonical TSDF ϕ_{model}. In iteration t, we estimate a deformation field increment $\Psi = (U, V, W)$ and apply it to the current warped TSDF $\phi^{(t)}_{proj}$, obtaining its new state $\phi^{(t+1)}_{proj}$ via trilinear interpolation. We do this following a variational formulation consisting of a data term and a regularizer:

$$E_{def}(\Psi) = E_{data}(\Psi) + w_{reg} E_{reg}(\Psi)\,, \tag{4}$$

where $w_{reg} > 0$ controls the trade-off between data fidelity and regularity. A solution of this model can be found via a gradient descent scheme with step size $\alpha > 0$:

$$\Psi^{(t+1)} = \Psi^{(t)} - \alpha\, \nabla E_{def}\left(\Psi^{(t)}\right)\,, \tag{5}$$

where $\nabla E_{def}\left(\Psi^{(t)}\right)$ denotes the variational derivative of the energy with respect to the deformation field. It is important to note that ∇E_{def} depends on the choice of the underlying inner product as explained in Section 4.2.

Data term Our data term enforces similarity between the TSDF that we are warping and the target canonical model by minimizing their squared voxel-wise difference:

$$E_{data}(\Psi) = \frac{1}{2} \sum_{x,y,z} \big(\phi_{proj}(x+u, y+v, z+w) - \\ -\phi_{model}(x,y,z)\big)^2 . \tag{6}$$

Applying standard calculus of variations we obtain:

$$\nabla E_{data}(\Psi) = \big(\phi_{proj}(\Psi) - \phi_{model}\big)\, \nabla\phi_{proj}(\Psi) . \tag{7}$$

Note that we use the symbol ∇ both for the spatial gradient of ϕ and for the variational derivatives of the energy terms.

Regularizer Our pipeline targets noisy Kinect data, which might cause inconsistencies within voxel neighbourhoods that result in holes in the reconstruction. Therefore we employ a classical Tikhonov-type regularizer that reduces spurious artifacts by imposing uniform motion:

$$E_{reg}(\Psi) = \frac{1}{2} \sum_{x,y,z} \big(|\nabla U(x,y,z)|^2 + \\ +|\nabla V(x,y,z)|^2 + |\nabla W(x,y,z)|^2\big) . \tag{8}$$

Using calculus of variations we obtain:

$$\nabla E_{reg}(\Psi) = -(\Delta U, \Delta V, \Delta W)^\top , \tag{9}$$

where ΔU denotes the Laplace operator applied to the x-component of the flow field, and similarly for V and W.

4.2. Sobolev Gradient Flow

The main idea of Sobolev gradient flows can be summarized as follows: compute the variational derivative of an energy with respect to the inner product of a smooth subspace of L^2, *i.e.* a Sobolev space, to obtain a gradient, which employed in a descent scheme yields a gradient flow that favours globally consistent solutions and is less susceptible to undesired local minima. Sundaramoorthi *et al.* [44] coined the term *coarse-to-fine evolution* for this effect, which accurately summarizes the fact that coarse-scale changes are favoured over fine-scale ones. In the context of incremental 3D reconstruction, this means that the warped TSDF will first adapt to more global deformations before eventually converging also w.r.t. to fine-scale details.

To compute a Sobolev gradient, it is sufficient to project the original gradient ∇E_{def} to the Sobolev space H^1 [5]. Identifing ∇E_{def} from Eq. (5) as the L^2 gradient $\nabla_{L^2} E_{def}$, we obtain:

$$\nabla_{H^1} E_{def} = (Id - \lambda\Delta)^{-1}\, \nabla_{L^2} E_{def} , \tag{10}$$

where Id denotes the identity operator. Eq. (10) involves the solution of an equation system, but it is possible to derive an approximate way of obtaining Sobolev gradients. First we note that Eq. (10) can be realized via

$$\nabla_{H^1} E_{def} = S * \nabla_{L^2} E_{def} , \tag{11}$$

where the filter S is the impulse response of the operator $(Id - \lambda\Delta)^{-1}$. In practice, we approximate S for a chosen value of λ and filter size s by solving the following system:

$$(Id - \lambda\Delta) S = v , \tag{12}$$

where v is a one-hot vector that corresponds to a discretized Dirac impulse of size $s \times s \times s$ voxels, and Δ is the Laplacian matrix discretized via a 7-point finite-difference stencil.

However, 3D convolutions might become prohibitively expensive for large values of s. Thus we further approximate the Sobolev kernel S by three separable 1D convolutions. To this end, we calculate the tensor higher-order SVD decomposition [24] of S and retain only the first singular vector from each resulting U matrix, and after normalization to unit sum obtain the 1D s-element filters S_x, S_y and S_z. As they contain the same entries, the subscript denotes spatial direction of application. Note that this is an approximation of S that has indispensable performance advantages.

At this point it is important to remark the following:

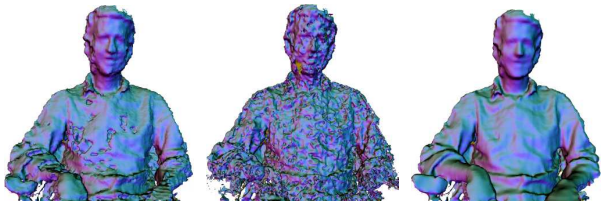
- A Sobolev gradient flow only enforces a more regular evolution to the desired minimum and not a more regular solution itself. Thus it favours globally consistent motions without changing the global optimum [45] and does not hamper the reconstruction of fine details, as we will demonstrate in our experiments.
- Thanks to this more consistent evolution, we do not need to enforce rigidity constraints, such as embedded deformation [43] or as-rigid-as-possible schemes [42] over meshes used by DynamicFusion [31] and its related methods [11, 18], or impose a divergence-free vector field prior like KillingFusion [40].
- Furthermore, our scheme does not require explicit reinitialization [34] or level set regularization [26, 27] to stabilize the evolution of the TSDF. This is in contrast to, for instance, KillingFusion [40] that uses both level set and rigidity priors, which are hard to balance and may cause over-smoothing effects.

4.3. Implementation Details

We use a default setting of neighbourhood size $s = 7$, filter parameter $\lambda = 0.1$, motion smoothness $w_{reg} = 0.2$ and gradient descent step size $\alpha = 0.1$. Our model is robust with regard to the parameter choice and achieves good results with a variety of settings (*c.f.* also an overview in supplementary material). To explain their acceptable ranges, we display reconstructions of the full-loop *Andrew-Chair* sequence from Dou *et al.* [11] in Figure 3.

A Sobolev filter size $s = 3$ is not sufficient to achieve satisfactory results. However, a larger kernel impedes speed, while the differences with $s \geq 7$ become negligible.

The parameter λ has an effect on the convergence rate. We empirically determined that doubling its value reduces



(a) $s = 3$ (b) $w_{reg} = 0$ (c) default

Figure 3. **Extreme versus recommended parameter choices** for Sobolev neigbourhood s, kernel strength λ and motion regularity w_{reg}: (a) a small neigbourhood is not able to fully overcome the effects of noise; (b) no motion regularization results in inconsistent geometry; (c) the default setting yields a detailed reconstruction.

the number of iterations by 3-8%. Moreover, as Figure 3(b) shows, motion regularity is essential to overcome noise. The ranges $\lambda \in [0.05; 0.4]$ and $w_{reg} \in [0.1; 0.5]$ yield high fidelity reconstructions, and we set the default values in the middle of those intervals.

Although our energy consists of only two terms, runtime is dominated by the Sobolev convolutions. Depending on the bounding volume, we use a voxel size in the range 4-12 mm in order to fit our regular voxel grid into GPU memory. Our pipeline achieves 30 fps for 64^3 voxels on a laptop with an Nvidia Quadro K1100M GPU with 2 GB of global memory, and for 128^3 voxels on a desktop PC with an Nvidia Titan Black with 6 GB memory.

5. Voxel Correspondences

Having developed a strategy for reliable non-rigid reconstruction, we now aim to colour the resulting model. However, as level set methods do not preserve correspondences [36, 50], colours would diffuse into each other if we warp an RGB grid in the same way as the TSDF [41].

We therefore turn to techniques based on the spectrum of the Laplacian matrix of a shape, which is invariant to isometric deformations [2, 19]. Its lower-frequency eigenfunctions, corresponding to the smallest eigenvalues, represent the base shape (*e.g.* a human body), while the higher-frequency ones carry details (limbs, wrinkles) [25, 37].

Recently it has been attempted to implicitly transfer correspondences in a level set framework via a term based on the difference of the lowest-frequency eigenfunctions [41]. As the overall scheme involves TSDF evolution, it has been shown to succeed only on constrained motion of complete shapes. We thus develop a scheme for direct voxel matching between TSDFs of incomplete shapes, based on the eigenfunction signature matching proposed by Mateus *et al*. [28].

Spectral embedding Our objective is to find correspondences between ϕ_{proj} and ϕ_{warped}. We first calculate the normalized graph Laplacian matrices of these voxel grids.

Let the number of voxels in the narrow band that is not truncated to ± 1 be l (they do not need to be the same for both shapes). We refer to them as occupied in the current context. This is the main difference between our proposed solution and other spectral methods, which typically consider the entire shape. The adjacency matrix W of size $l \times l$ has an entry 1 when adjacent voxels are occupied, and 0 elsewhere. Note that the diagonal entries are 0, as a voxel is not adjacent to itself. The degree matrix D contains the degree of each voxel, *i.e.* the row-wise sums of elements in W, on its diagonal. Then the normalized Laplacian is:

$$L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} \,. \tag{13}$$

According to Umeyama's theorem, finding correspondences between the two shapes can be done through alignment of their Laplacian eigenspaces [47]. Let $L = U\Lambda U^\top$ be the eigendecomposition. As the number of voxels in our shapes is very large, we resort to a lower-dimensional embedding containing the K smallest non-zero eigenvalues and their eigenvectors [28]. The columns of the respective matrix U^K are the K retained eigenvectors, while its l rows are the K-dimensional coordinates of the embedded shape.

However, there is no guarantee that the eigenvalues are reliably ordered in the embedding, so we need to find a $K \times K$ permutation matrix P that aligns the eigenspaces of our two shapes. In addition, due to sign ambiguity, we have to determine a sign matrix M, resulting in an overall transformation $T = MP$, as described in the next part. It relates the reduced embeddings as follows:

$$(U^K_{warped})^\top = T(U^K_{proj})^\top \,. \tag{14}$$

The correspondences between the embeddings are transferred to the voxels of the original shapes via nearest neighbour search between embedded- and voxel-coordinates.

Eigenfunction signature matching We seek an optimal assignment between the column eigenvectors $\mathbf{u}^i_{proj}$ and $\mathbf{u}^j_{warped}$, $i, j \in \{1, ..., K\}$ of U^K_{proj} and U^K_{warped}. The approach of Mateus *et al*. [28] suggests to construct histograms from these eigenvectors, since they are invariant to the value ordering and the number of entries l, and view them as signatures of the eigenfunctions. We thus build a 200-bin histogram $hist(\cdot)$ from each vector and store the similarity of each eigenvector pair as the ℓ_1 histogram difference in a score matrix A:

$$A_{i,j} = \min(||hist(\mathbf{u}^i_{proj}) - hist(\pm\mathbf{u}^j_{warped})||_1) \,. \tag{15}$$

Additionally, a matrix M' stores the sign of $\pm\mathbf{u}^j_{warped}$ that yielded the lower score.

This is an assignment problem between eigenfunction signatures, which we solve for the lowest cost via the

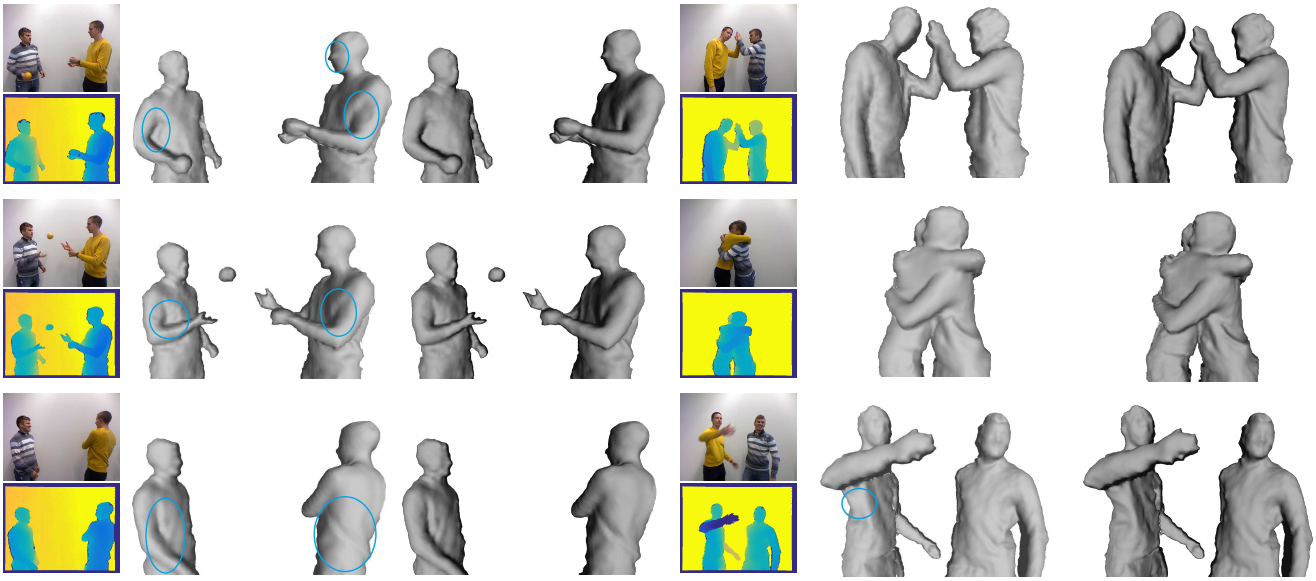


Figure 4. **Comparison of SobolevFusion to KillingFusion [40] on scenes with interacting subjects**, such as two people playing with a ball, high-fiving and hugging. Both methods handle the motion, but SobolevFusion demonstrates better capture of geometric details, while KillingFusion tends to over-smooth and thus, for instance, creates the impression that limbs are fused into the body (see marked regions).

Munkres algorithm [13] over A. We then build the permutation matrix P according to its output, and look up M' for the appropriate sign in M. Thus we obtain the sought transformation matrix $T = MP$ and use it to estimate the correspondence. If a near-surface voxel is assigned to an off-surface voxel, we discard the match.

After obtaining initial matches, we use the Weiszfeld algorithm [48] to determine the geometric median in a $3 \times 3 \times 3$ neighbourhood in order to retain only the most likely correspondence. This step is crucial, since as opposed to Mateus *et al*. [28] and other prior work, we are dealing with partial shapes, so their Laplacian eigenfunctions might carry information about non-overlapping regions.

In our implementation we choose $K \leq 20$, since higher-frequency eigenfunctions might be contaminated by noise or pertain to details of the shape rather than its base structure, which is undesirable for partial TSDFs. As parallelization of the voxel matching procedure is not straightforward, in practice we run it on the CPU while the next frame(s) are being warped on the GPU. It takes 58-500 ms per frame on a 2.80 GHz Intel Core i7 CPU, depending on the volume size. Once done, it continues with the latest warped frame, effectively avoiding temporal overhead.

6. Evaluation

Figure 1 demonstrates that SobolevFusion can reconstruct a complete 3D model of a subject moving in a 360° loop, undergoing large motion and interacting with a balloon, leading to merging and splitting topology.

Figure 5. **Evaluation of geometric error** on objects with ground-truth canonical pose models from KillingFusion [40]. The error is given under the respective output of KillingFusion (KF) and SobolevFusion (SF). In addition to achieving higher geometric accuracy, our method is less susceptible to high-frequency noise on the *Duck* and to over-smoothing on the *Snoopy* sequence.

In this section we carry out various experiments in order to assess the performance of SobolevFusion and compare it to state-of-the art techniques. We test the different aspects of our system separately, namely geometric accuracy, performance under large motion, and ability to transfer colour to the output model.

6.1. Geometric Fidelity

Most related to our method is KillingFusion [40] due to the variational formulation based on signed distance field deformation. In Figure 4 we compare SobolevFusion against our implementation of KillingFusion with default parameters on data that we acquired with a Kinect v1, featuring fast motion, multiple interacting subjects and thus topological changes (more results can be found in our supplementary video). As expected, both methods are able to handle such motion. However, KillingFusion tends to over-

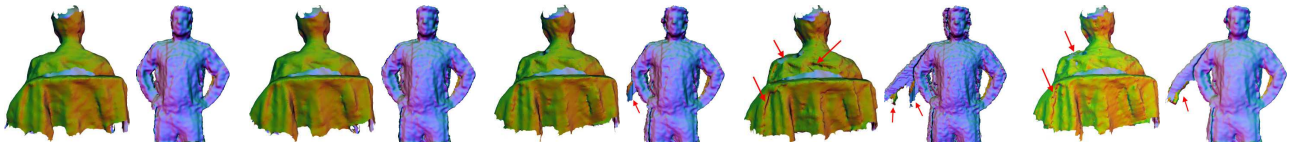


Figure 6. **Lower frame-rate test.** We use only every n^{th} frame, as indicated under the results. SobolevFusion outputs high-fidelity reconstructions using only 20% of the frames. For slow motion, even less frames give good results, while for large motion some of the geometry cannot be recovered, resulting in artifacts. The right-most columns show the KillingFusion [40] result for every 10^{th} frame, exhibiting similar degradation properties as SobolevFusion does for every 15^{th} frame due to its better convergence.

Figure 7. **Canonical model comparison** on the full-loop *Squeeze* sequence from DynamicFusion [31]. SobolevFusion recovers the fine structures on the face better than KillingFusion [40].

smooth facial features and folds on clothes, while these are more clearly visible with our approach. Our reconstructions contain less noise as the underlying Sobolev gradient flow provides higher robustness to it. Moreover, our method captures concavities better and defines sharper edges, both at the shape outline and where surfaces touch. Last but not least, we observed that SobolevFusion requires up to 15 % less iterations to converge.

For quantitative evaluation we test on the fast-motion mechanical toy sequences from KillingFusion [40], where it has already been demonstrated that a TSDF-based approach performs better than a mesh-based technique, such as VolumeDeform [18], under large motion and topological changes. Figure 5 shows that our SobolevFusion further decreases the geometric error and outputs more detailed reconstructions. This is especially noticeable on *Snoopy* for which the regularizers of KillingFusion lead to over-smoothing, while our Sobolev gradient flow keeps fine details while avoiding spurious artifacts caused by noise. Therefore SobolevFusion achieves both an increased level of geometric detail and a lower reconstruction error than KillingFusion.

Similarly, in Figure 7 we demonstrate better preservation of detail than KillingFusion [40] on the 360° *Squeeze* sequence from DynamicFusion [31]. For instance, the facial features are much more conspicuous in our case. Note that due to the used regular voxel grid our result is still less detailed than that of DynamicFusion.

We also compare the level of geometric detail of a TSDF warped via Sobolev gradient flow versus that of a mesh-based technique. For this purpose in Figure 8 we show live frames from the *Umbrella* sequence used in VolumeDeform [18]. Our method recovers similar, or even higher, level of detail as VolumeDeform, without creating spurious elements around the open edge or fusing the strap into the umbrella. Furthermore, KillingFusion over-smooths the tip, while SobolevFusion manages to capture this fine structure using the same voxel size.

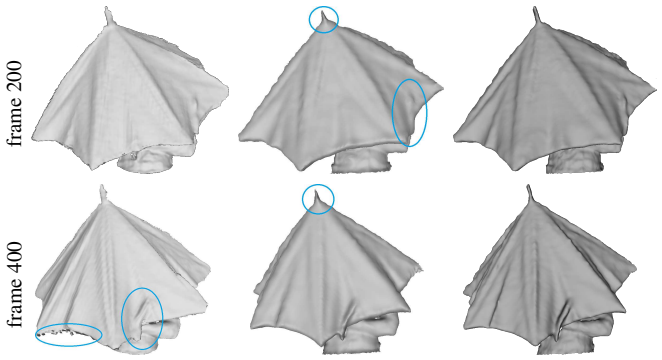


Figure 8. **Warped live frame comparison** on the *Umbrella* from VolumeDeform [18]. SobolevFusion yields similar or higher level of detail as VolumeDeform without artifacts at the edge, while KillingFusion [40] over-smooths thin elements such as the tip.

6.2. Large Motion

Even though datasets from the previous section exhibit large motion, we simulate a lower frame-rate sensor by taking every n^{th} frame from 360° sequences. To this end we use the slow-motion *Andrew-Chair* from Dou *et al*. [11] and the fast *Alex* sequence from KillingFusion [40], as displayed in Figure 6. Naturally, when less frames are fused, the cumulative TSDF is noisier. However, when only every 10^{th} frame is used, the reconstruction is still consistent for the slower *Andrew-Chair* sequence, while the faster *Alex* sequence starts creating artifacts due to misaligned geometry. Moreover, due to improved convergence of the Sobolev scheme, our method manages to recover even larger motion than KillingFusion. This can be concluded from the last two columns of Figure 6, as the KillingFusion result for *Alex* at 10-frame speedup is similar to that of SobolevFusion for 15-frame speedup.

Figure 9. **Colour transfer** from reference frame i to target frame $i+n$. With larger distance the amount of transferred colour decreases, but remains correct due to our carefully designed scheme.

Figure 10. **Coloured canonical-pose models**, obtained with our voxel matching scheme between TSDFs of incomplete shapes.

6.3. Texture Transfer

The reconstruction part of our pipeline is independent of the voxel matching, therefore it can be run separately. Here we assess the quality of colour propagation that we achieve.

In Figure 9 we display the amount of colour that our technique can transfer on the *Minion* sequence from VolumeDeform [18]. We test on consecutive frames, as well as on frames separated by a larger distance. The amount of texture that is being transferred decreases with the increasing pose difference, but our scheme manages to determine stable matches even when views are 15 frames apart. Furthermore, our procedure for match rejection makes sure that only reliable correspondences are returned, and thus there is no transfer of incorrect colours.

Further textured examples are shown in Figure 10. As explained in the implementation part of Section 5, we do not necessarily determine matches for every frame in order not to hamper speed. This is justified, since consecutive frames have a significant overlap. However, a certain amount of voxels might remain un-coloured. In that case, we assign to them the colour that the gradient flow propagates from the initial projective TSDF. Due to the multiple interpolation steps, this colour is typically contaminated by the colours of nearby voxels, but is a plausible estimate.

Figure 1 shows the texture we are able to recover after the subject does a complete 360° loop. Colours on the front are rather crisp, since the difference between the canonical pose and the initial frames is not too large and thus matching is very exact. The back shows more mixed colours, as the poses become more distant and matching becomes more challenging, but the result remains visually pleasing.

Our main goal is to reliably colour the reconstructions we obtain, rather than to estimate a dense set of correspondences. Nevertheless, we quantitatively evaluate on the *yt* sequence with Vicon markers used in BodyFusion [52], which features a human executing various motions. We observed that our matching procedure typically returns a low error for markers on the torso of the subject, which is a region where mesh-based correspondences often exhibit sliding. However, since the lower-frequency Laplacian eigenfunctions do not always capture limbs, often correspondences are not estimated for markers located on the arms. As 12 out of the 18 Vicon markers are placed on the subject's arms, this dataset is not optimally suited for our method, which on average returns matches for half the markers per frame. Yet, our mean ℓ_1 error of 7.7 cm over the entire sequence is comparable to the 4.4 cm of DynamicFusion [31] and 3.7 cm of VolumeDeform [18], considering that we always stay in voxel space and thus accumulate more discretization error, while the other methods explicitly determine correspondences for deformation field calculation (BodyFusion achieves a lower error by combining with a human skeleton prior; *c.f.* Table 1 of their paper [52]). This is a promising result for the incorporation of explicit correspondences into implicit level set frameworks.

7. Limitations and Future Perspectives

Although our framework runs at interactive rates, its speed and memory consumption can be further optimized by replacing the regular voxel grid TSDF representation by an appropriate hashing [33] or hierarchical structure [21].

The voxel matching opens up more avenues for future work. One of our goals is to obtain denser correspondences. A possibility to do this is an expectation-maximization procedure over the spectral matches, which is, however, not feasible in real time [28]. An alternative would be to learn a mapping from sparse to dense fields [51], or even learn correspondences in the spectral embedding. Moreover, segmentation can be helpful in the case of multiple objects, so that for each one we can compute a separate, more representative Laplacian matrix.

8. Conclusion

We have presented a method for non-rigid fusion of scenes undergoing free motion, including fast movements, changing topology and interacting agents. The introduced variational energy formulation is cheaper to compute, converges faster and leads to reconstructions of higher geometric quality than related techniques. It is minimized using a Sobolev gradient flow, for which we have developed an efficient separable 1D convolution implementation. Moreover, we have proposed a correspondence estimation strategy over TSDFs of partial shapes, allowing realistic colouring of the obtained models. Our system uses a single RGB-D stream and can cope with significantly less frames than other approaches, paving the way to applications such as unconstrained performance capture and 3D avatar creation under large motion.

References

[1] M. Baust, D. Zikic, and N. Navab. Variational Level Set Segmentation in Riemannian Sobolev Spaces. In *British Machine Vision Conference (BMVC)*, 2014. 2

[2] M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS)*. 2, 5

[3] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[4] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic Deformable Surface Tracking from Multiple Videos. In *European Conference on Computer Vision (ECCV)*, 2010. 2

[5] J. Calder, A. Mansouri, and A. Yezzi. Image Sharpening via Sobolev Gradient Flows. *SIAM Journal on Imaging Sciences*, 3(4):981–1014, 2010. 2, 4

[6] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-Quality Streamable Free-Viewpoint Video. *ACM Transactions on Graphics (TOG)*, 34(4), 2015. 2

[7] B. Curless and M. Levoy. A Volumetric Method for Building Complex Models from Range Images. In *23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 303–312, 1996. 3

[8] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun. Performance Capture from Sparse Multi-view Video. *ACM Transactions on Graphics (TOG)*, 27(3), 2008. 2

[9] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2Fusion: Real-time Volumetric Performance Capture. In *ACM Transactions on Graphics (TOG)*, 2017. 1

[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 Transactions on Graphics (TOG)*, 35(4), 2016. 2

[11] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3D Scanning Deformable Objects with a Single RGBD Sensor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 4, 7

[12] S. R. Fanello, J. Valentin, C. Rhemann, A. Kowdle, V. Tankovich, P. Davidson, and S. Izadi. UltraStereo: Efficient Learning-Based Matching for Active Stereo Systems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[13] A. Frank. On Kuhns Hungarian Method - A Tribute from Hungary. Technical Report 2004-14, Egérvary Research Group on Combinatorial Optimization, Budapest, Hungary, 2004. 6

[14] E. Göçeri. Fully Automated Liver Segmentation using Sobolev Gradient-based Level Set Evolution. *International Journal for Numerical Methods in Biomedical Engineering*, 32(11), 2016. 2

[15] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time Geometry, Albedo and Motion Reconstruction Using a Single RGBD Camera. *ACM Transactions on Graphics (TOG)*, 2017. 1, 2

[16] C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer. Volumetric 3D Tracking by Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[17] Infineon Technologies AG. REAL3^{TM} Image Sensor Family: 3D Depth Sensing Based on Time-of-Flight. Product Brief, 2015. 1

[18] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time Volumetric Non-rigid Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 4, 7, 8

[19] V. Jain and H. Zhang. Robust 3D Shape Correspondence in the Spectral Domain. In *IEEE International Conference on Shape Modeling and Applications (SMI)*, 2006. 2, 5

[20] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[21] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray. Very High Frame Rate Volumetric Integration of Depth Images on Mobile Devices. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 21(11):1241–1250, 2015. 1, 8

[22] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb. Real-time 3D Reconstruction in Dynamic Scenes using Point-based Fusion. In *International Conference on 3D Vision (3DV)*, 2013. 1

[23] C. Kerl, J. Sturm, and D. Cremers. Dense Visual SLAM for RGB-D Cameras. In *International Conference on Intelligent Robot Systems (IROS)*, 2013. 1

[24] J. B. Kruskal. Multiway Data Analysis. chapter Rank, Decomposition, and Uniqueness for 3-way and N-way Arrays. 1989. 4

[25] B. Levy. Laplace-Beltrami Eigenfunctions Towards an Algorithm That "Understands" Geometry. In *IEEE International Conference on Shape Modeling and Applications (SMI)*, 2006. 2, 5

[26] C. Li, C. Xu, C. Gui, and M. D. Fox. Level Set Evolution Without Re-initialization: A New Variational Formulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2, 4

[27] C. Li, C. Xu, C. Gui, and M. D. Fox. Distance Regularized Level Set Evolution and Its Application to Image Segmentation. *IEEE Transaction on Image Processing (TIP)*, 19(12):3243–3254, 2010. 4

[28] D. Mateus, R. Horaud, D. Knossow, F. Cuzzolin, and E. Boyer. Articulated Shape Matching using Laplacian Eigenfunctions and Unsupervised Point Registration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2, 5, 6, 8

[29] M. Meilland and A. I. Comport. On Unifying Key-frame and Voxel-based Dense Visual SLAM at Large Scales. In

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2013. 1

[30] J. Neuberger. *Sobolev Gradients and Differential Equations*. Springer Science & Business Media, 2009. 1, 2

[31] 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. 1, 2, 4, 7, 8

[32] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *10th International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 1

[33] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. *ACM Transactions on Graphics (TOG)*, 2013. 1, 8

[34] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153 of *Applied Mathematical Science*. Springer, 2003. 2, 4

[35] S. Osher and J. Sethian. Fronts Propagating with Curvature-dependent speed: Algorithms based on Hamilton-Jacobi Formulations. *Journal of Computational Physics*, 79(1):12–49, 1988. 2

[36] J.-P. Pons, G. Hermosillo, R. Keriven, and O. Faugeras. How to Deal with Point Correspondences and Tangential Velocities in the Level Set Framework. In *IEEE International Conference on Computer Vision (ICCV)*, 2003. 2, 5

[37] M. Reuter, F.-E. Wolter, and N. Peinecke. Laplace-Beltrami Spectra As 'Shape-DNA' of Surfaces and Solids. *Computer-Aided Design*, 38(4):342–366, 2006. 2, 5

[38] R. M. Rustamov. Interpolated Eigenfunctions for Volumetric Shape Processing. *The Visual Computer*, 27(11), 2011. 2

[39] C. Schroers, H. Zimmer, L. Valgaerts, A. Bruhn, O. Demetz, and J. Weickert. Anisotropic Range Image Integration. In *Joint German and Austrian Conference on Pattern Recognition (DAGM-OAGM)*, 2012. 3

[40] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. KillingFusion: Non-rigid 3D Reconstruction without Correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4, 6, 7

[41] M. Slavcheva, M. Baust, and S. Ilic. Towards Implicit Correspondence in Signed Distance Field Evolution. In *PeopleCap Workshop, IEEE International Conference on Computer Vision (ICCVW)*, 2017. 5

[42] O. Sorkine and M. Alexa. As-Rigid-As-Possible Surface Modeling. In *Fifth Eurographics Symposium on Geometry Processing (SGP)*, 2007. 4

[43] R. W. Sumner, J. Schmid, and M. Pauly. Embedded Deformation for Shape Manipulation. *ACM Transactions on Graphics (TOG)*, 26(3), 2007. 4

[44] G. Sundaramoorthi, A. Yezzi, and A. Mennucci. Coarse-to-Fine Segmentation and Tracking using Sobolev Active Contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(5):851–864, 2008. 2, 4

[45] G. Sundaramoorthi, A. Yezzi, and A. C. Mennucci. Sobolev Active Contours. *International Journal of Computer Vision (IJCV)*, 73(3):345–366, 2007. 1, 2, 4

[46] A. Trouvé. Diffeomorphisms Groups and Pattern Matching in Image Analysis. *International Journal of Computer Vision (IJCV)*, 28(3):213–221, 1998. 2

[47] S. Umeyama. An Eigendecomposition Approach to Weighted Graph Matching Problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 10(5):695–703, 1988. 5

[48] E. Weiszfeld and F. Plastria. On the Point for Which the Sum of the Distances to n Given Points is Minimum. *Tôhoku Mathematical Journal*, 1937. 6

[49] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. ElasticFusion: Dense SLAM Without A Pose Graph. In *Robotics: Science and Systems (RSS)*, 2015. 1

[50] R. T. Whitaker. A Level-Set Approach to 3D Reconstruction from Range Data. *International Journal of Computer Vision (IJCV)*, 29(3):203–231, 1998. 2, 5

[51] J. Wulff and M. J. Black. Efficient Sparse-to-Dense Optical Flow Estimation using a Learned Basis and Layers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[52] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. BodyFusion: Real-time Capture of Human Motion and Surface Geometry Using a Single Depth Camera. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 8

[53] H.-K. Zhao, T. Chan, B. Merriman, and S. Osher. A Variational Level Set Approach to Multiphase Motion. *Journal of Computational Physics*, 127(1):179–195, 1996. 2

[54] D. Zikic, M. Baust, A. Kamen, and N. Navab. A General Preconditioning Scheme for Difference Measures in Deformable Registration. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[55] M. Zollhöfer, M. Nießner, S. Izadi, C. Rhemann, 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. *ACM Transactions on Graphics (TOG)*, 33(4), 2014. 2