Surface Normal Integration for Convex Space-time Multi-view Reconstruction

Martin R. Oswald
martin.oswald@in.tum.de
Daniel Cremers
cremers@tum.de

Computer Vision Group,
Department of Computer Science,
Technische Universität München

Abstract

We show that surface normal information allows to significantly improve the accuracy of a spatio-temporal multi-view reconstruction. On one hand, normal information can improve the quality of photometric matching scores. On the other hand, the same normal information can be employed to drive an adaptive anisotropic surface regularization process which better preserves fine details and elongated structures than its isotropic counterpart. We demonstrate how normal information can be used and estimated and explain crucial steps for an efficient implementation. Experiments on several challenging multi-view video data sets show clear improvements over state-of-the-art methods.

1 Introduction

The extension of multi-view 3D reconstruction approaches to the spatio-temporal domain is far from straightforward: Firstly, with the processing of huge amounts of data computational speed becomes more important. Algorithms which take around an hour for the reconstruction of a single frame are hardly scalable to multi-view videos taken at 30 frames per second. Secondly, integrating temporal regularization gives rise to a substantial increase in memory requirements because the reconstructions for multiple time steps need to be computed jointly. Thirdly, the acquisition of actions over time brings about substantial motion blur of fast moving structures – see the rope in Figure 1. And lastly, one typically uses far fewer cameras

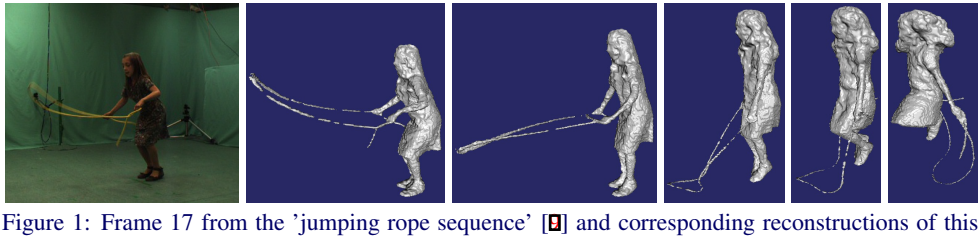


Figure 1: Frame 17 from the 'jumping rope sequence' [9] and corresponding reconstructions of this and the following time frames computed with the proposed method. By minimizing a single convex functional, we obtain a family of reconstructions over time. By integrating normal information into the photoconsistency estimation and into an anisotropic space-time regularization, we are able to preserve fine scale details such as the (substantially motion-blurred) rope.

with lower resolution (the synchronization and joint acquisition being tedious) such that classical photoconsistency approaches often break down.

1.1 Related Work

Multi-view stereo reconstruction for static scenes has been a focus of many works and there exists a vast amount of different approaches. In the following we mainly mention works which relate to normal integration and spatio-temporal reconstruction. We refer to [20] and its related website for an overview of 3D reconstruction methods.

An early work considering surface normals while estimating a 3D surface is by Zabulis and Daniilidis [24]. They estimate voxel occupancy at the surface and corresponding normals in a voxel grid by locally maximizing corresponding surface patch correlation values. The optimization is spatially local and results in rather noisy and disconnected surfaces. Furukawa et al. [4] propose to jointly estimate depth and orientation of surface patches by means of an oriented point cloud which can then be transformed into a mesh e.g. via Poisson surface reconstruction [11]. Goesele et al. [5] built a system for reconstructing 3D scenes from internet photo collections. They show that optimizing surface normal information with respect to the photoconsistency measure significantly improves the reconstruction quality. Both methods [4, 5] are based on an oriented point cloud which is grown and filtered iteratively around existing matches by starting from sparse feature matches. Generally, an extension of such models into a spatio-temporal domain is by no means straightforward because the correspondence of points over time needs to be identified first.

Ladikos et al. [14] used a narrow band graph-cut approach for multi-view reconstruction. They jointly maximized a normalized cross-correlation (NCC) photoconsistency measure and computed the best normal by discretely sampling a dense set of normals in the cone around an initial normal estimate. Kolev et al. [13] improved the results of multi-view 3D reconstruction with an anisotropic regularizer and a given normal field. We use a similar regularizer, but additionally discuss how to compute such a normal field and how it can be used to improve photometric measures.

Goldluecke et al. [6, 7] pioneered spatio-temporal reconstruction in a continuous setting, albeit in a locally optimal level-set formulation. Starck and Hilton [21] extract visual hull volumes and sparse features on the surface and finally merge them via volumetric minimal graph-cuts. Tung et al. [22] fuse feature points and optical flow with a Markov random field. Guillemaut and Hilton [8] jointly solve for a multi-layer segmentation and a depth estimation within a graph-cut framework. Pons et al. [18] jointly estimate scene geometry and scene flow by minimizing the reprojection error in a locally optimal coarse-to-fine approach.

1.2 Contribution

We propose a convex variational approach to space-time reconstruction which estimates surface normal information and integrates it into the photoconsistency estimation as well as into an anisotropic spatio-temporal total variation regularization. As such the proposed method generalizes the works [13], [16]. Although [13] already studied anisotropic regularization they did not estimate normals but used the normals from [4]. The combination of these methods, [13] and [4], is more than 40 times slower than our method as [13] alone needs about 1h to compute a single frame. In contrast, our method only takes about 3 minutes per frame including normal estimation and temporal regularization due to the proposed efficient implementation. Moreover, the method by Kolev et al. [13] does not work well on the 4D

data sets we consider, as shown in [16, Fig. 5]. With the estimated normals at hand, we further propose an improvement of the photoconsistency voting scheme by Hernández and Schmitt [2] resulting in superior accuracy especially for sparse camera setups.

2 Variational Space-Time Reconstruction Model

We aim to find a smooth hypersurface S in the spatio-temporal space $V \times T$ in which $V \subset \mathbb{R}^3$ represents the spatial and $T \subset \mathbb{R}_+$ the temporal domain. A non-static scene is observed from N cameras with known projections $\{\pi_i\}_{i=1}^N$ and approximate silhouettes $\{S_i(t)\}_{i=1}^N$. Similar to [16], we assume the silhouettes to fully enclose the object of interest and restrict the solution space to the visual hull. We do not rely on exact silhouettes as they are difficult to estimate in a general 4D setup. Hence, methods using exact silhoettes such as [1] are not applicable. Note that we will drop temporal indices whenever possible for better readability.

First, we introduce a binary labeling function $u : V \times T \mapsto \{0,1\}$ to represent the hypersurface S by means of an inside-outside labeling in each point. This implicit surface representation easily deals with topology changes and allows to compute minimal surfaces that align with locations of high photometric consistency in a globally optimal manner [12]. We compute a hypersurface as a minimum of the following energy.

$$E(u) = \int_{V \times T} \Big[|\nabla_{\boldsymbol{x}} u|_{D_{\boldsymbol{x}}} + g_t |\nabla_t u| + \lambda\, f u \Big]\; d\boldsymbol{x}dt \tag{1}$$

The parameter $\lambda \geq 0$ steers the smoothness of the solution by balancing the costs of the regularization term and the data term. The function $f : V \times T \mapsto \mathbb{R}$ represents unary potentials which indicate local preferences for either an interior or an exterior label based on the photoconsistency being defined in the next section. The task of the regularization term is to reject outliers, to deal with locations of missing data and to favor a spatially and temporally smooth surface. The regularization term consists of two terms, one for the anisotropic spatial regularization with the norm defined as $|\boldsymbol{y}|_{D_{\boldsymbol{x}}} = \langle \boldsymbol{y}, D_{\boldsymbol{x}} \boldsymbol{y} \rangle^{1/2}$ and the other term takes care of the temporal regularization. Both terms are detailed in the following.

Spatial Regularization. $D_{\boldsymbol{x}}(\boldsymbol{x},t) = \rho(\boldsymbol{x},t)^2 \boldsymbol{n}\boldsymbol{n}^T + \boldsymbol{n}_0 \boldsymbol{n}_0^T + \boldsymbol{n}_1 \boldsymbol{n}_1^T$ is a symmetric positive-definite matrix D_x which accounts for an anisotropic spatial regularization and is defined similarly as in [19]. It lowers smoothing in the surface normal $\boldsymbol{n} \in \mathbb{R}^3$ direction and favors smoothness along the corresponding tangential directions $\boldsymbol{n}_0$ and $\boldsymbol{n}_1 = \boldsymbol{n} \times \boldsymbol{n}_0$. The photoconsistency measure $\rho : V \times T \mapsto [0,1]$ is detailed in the next section. $D_{\boldsymbol{x}}$ performs a change of basis and aligns the local coordinate system along the favored surface normal $\boldsymbol{n}$. As a result, ∇u is more likely to be aligned to $\boldsymbol{n}$. On the one hand the anisotropic regularization better preserves small scale surface details [13], on the other hand it is important when reconstructing fine elongated structures [19] like human arms, or parts of clothes and hair.

Temporal Regularization. In Eq. (1) function $g_t : V \times T \mapsto \mathbb{R}_{\geq 0}$ regulates the temporal smoothness. By setting $g_t(\boldsymbol{x},t) = \exp\big(-a|\nabla f(\boldsymbol{x},t)|\big)$ we make it dependent on the data term in order to reduce temporal smoothing in regions with fast motion. Note that the purpose of this regularization is to reduce surface jittering in scene parts with slow motion. The effect of this term is studied in [16]. We used a between 0.2 and 1 in our setting.

3 Surface Normal Integration

Normal information is used in all stages of our approach, namely during the photoconsistency and data term estimation as well as during the global surface optimization.

3.1 Photoconsistency and Data Term Estimation

For every time step we estimate the photometric consistency of a point on the surface by means of a cost function $C_i : V \times \mathbb{R} \mapsto \mathbb{R}$ based on the NCC score of corresponding small image patches surrounding the projection of that point in each camera.

$$C_i(\boldsymbol{x}, d) = \sum_{j \in \mathcal{C}' \setminus i} w_i^j(\boldsymbol{x}) \cdot \mathrm{NCC}\Big(\pi_i\big(r_i(\boldsymbol{x}, d)\big), \pi_j\big(r_i(\boldsymbol{x}, d)\big)\Big) \tag{2}$$

The value d is the Euclidean distance of $\boldsymbol{x}$ from camera center i along camera ray $r_i(\boldsymbol{x}, \cdot)$ through point $\boldsymbol{x}$. $\mathcal{C}' \subset \mathcal{C}$ is a subset of front-facing cameras of which the angle between the viewing directions is below γ_{max}=85°. The contribution of each camera is weighted by a normalized Gaussian weight $w_i^j(\boldsymbol{x})$ of the angle between the voxel-to-camera directions of cameras i and j. Furthermore, we discard unreliable correlation values by setting $C_i(\cdot)$ to zero if it falls below a threshold $\tau_{\mathrm{ncc}} = 0.3$. To account for image distortion between two cameras during the NCC computation the image coordinates are mapped with the homography $H_{ij} = (\boldsymbol{n}^T \boldsymbol{x}) R_{ij}^T - R_{ij}^T T_{ij} \boldsymbol{n}^T$, with $\boldsymbol{n}$ being the surface normal and R_{ij}, T_{ij} being the relative rotation and translation between local coordinates of cameras i and j [3].

Since the correlation scores $C_i(\cdot)$ are usually noisy and contain many local maxima we denoise them with the voting scheme by Hernández and Schmitt [7] and define the photoconsistency measure $\rho(\boldsymbol{x})$ for the regularizer as

$$\rho(\boldsymbol{x}) = \exp\Big[-\mu \sum_{i \in \mathcal{C}'} \underbrace{\delta\big(d_i^{\max} = \mathrm{depth}_i(\boldsymbol{x})\big) \cdot C_i(\boldsymbol{x}, d_i^{\max})}_{\mathrm{VOTE}_i(\boldsymbol{x})}\Big] \quad . \tag{3}$$

This scheme accumulates only the best score along each camera ray. The point with maximum score is expressed by its distance to the camera center $d_i^{\max} = \arg\max_d C_i(\boldsymbol{x}, d)$. In comparison, for most 3D reconstruction approaches that first estimate depth maps before fusing them into a single 3D model, e.g. [25], the matching scores of single depth estimates are not considered in the depth fusion process. In contrast, the voting scheme accumulates matching scores and we hand them over to the global surface estimation. Another significant difference to such methods is the missing regularization of depth values in the image domain, which often helps to avoid depth ambiguities and to suppress noise. We therefore propose to introduce a dependency between neighboring camera rays by the following modification of the voting scheme:

$$d_i^{\max} = \arg\max_d \int_{V_{\boldsymbol{x}}} C_i(\boldsymbol{x} - \boldsymbol{y}, d)\ \mathcal{G}(\boldsymbol{y}; \Sigma_{\boldsymbol{n}})\ d\boldsymbol{y} \quad , \tag{4}$$

where $V_{\boldsymbol{x}} \subset V$ is a small volume surrounding $\boldsymbol{x}$. Each value of $C_i(\cdot)$ represents the matching score of a small surface patch with location $\boldsymbol{x}$ and orientation $\boldsymbol{n}$ and should also influence neighboring matching scores according to the patch size. We model this dependency by a Gaussian convolution of the matching scores before the maximization. Again, the normal information comes in handy to better represent the shape of the surface patch by an anisotropic

3D Gaussian $\mathcal{G}(\cdot)$ with covariance matrix $\Sigma_{\boldsymbol{n}} = R_{\boldsymbol{n}} \operatorname{diag}(\sigma_n^2, \sigma_t^2, \sigma_t^2) R_{\boldsymbol{n}}^T$. σ_n and σ_t are the standard deviations for normal and tangential directions and rotation matrix $R_{\boldsymbol{n}}$ aligns the x-axis of the coordinate system with the normal $\boldsymbol{n}$. This scheme effectively denoises depth hypotheses and improves the quality of matching scores for piecewise smooth surfaces as it helps to avoid local maxima by integrating information from neighboring viewing rays.

In order to avoid trivial solutions of energy (1) the photoconsistency is further imposed by means of an unary data term f, defined as the log-probability ratio

$$f(\boldsymbol{x},t) = -\ln\left(\frac{1-P(\boldsymbol{x}\in int(\Sigma))}{P(\boldsymbol{x}\in int(\Sigma))}\right) \quad . \tag{5}$$

The probability $P(\boldsymbol{x}\in int(\Sigma))$ that point $\boldsymbol{x}$ belongs to the interior of surface S is defined based on the voting locations and qualities of corresponding camera rays $r_i(\boldsymbol{x},\cdot)$ through point $\boldsymbol{x}$

$$P(\boldsymbol{x}\in int(\Sigma)) = \prod_{i=1}^{N}\prod_{j=1}^{N} \prod_{\text{depth}_i(\boldsymbol{x})<d\leq d_i^{\max}} \frac{1}{Z_j}\exp\left[-\eta\cdot \text{VOTE}_j\big(r_i(\boldsymbol{x},d)\big)\right] \tag{6}$$

Z_j is a normalization constant and parameter η steers how many cameras and which matching scores are necessary to favor an exterior label for all points from $\boldsymbol{x}$ towards the camera. Intuitively, the data term represents a probabilistic space carving and due to the restriction of the solution space to the visual hull, the visual hull is the fall back solution for all areas where photometric information is insufficient (see [16] for more details).

3.2 Normal Estimation

Similar to [14] we experimented with estimating the normal direction by global maximization of the NCC score via discrete sampling around the camera-to-point direction. Generally, pointwise optimization of the surface normal is prone to local minima and we merely got noisy and unsatisfactory results with this approach. Similar results have also been reported by [23]. We also tried estimating normals based on the data term f as done in [19] which also yields defective normals due to the fact that f is very noisy and misses a lot of data for most of our experiments. Kolev et al. [12] estimated normal directions for the photoconsistency computation based on the visual hull. Especially in sparse camera setups we found that the visual hull does not provide good normal estimates for recovering concavities.

Instead we use the camera-to-point direction as a first normal estimate for photoconsistency estimation which is a common (inherent) assumption in most stereo-based methods. We then compute a surface with isotropic spatial regularization and use the surface normals of this solution for a second pass of photoconsistency, data term estimation and surface optimization with anisotropic spatial regularization. For that purpose the surface normals are propagated in space by means of a signed distance function (Sec. 5). In sum, we make use of surface normals at three places within our method: (a) NCC score, (b) voting scheme regularization and (c) anisotropic surface regularization. We run our algorithm in two passes:

Pass 1: camera-to-point direction as normal for (a) and (b), isotropic surface regularization with high λ for (c)

Pass 2: normals from the previous pass for (a),(b) and (c) with lower λ for surface smoothness as desired

This scheme could be further iterated, but in our experience two passes achieve the best trade-off between quality improvements and additional computation time.

4 Global Optimization

The minimization problem in (1) becomes convex by relaxing the image of function u to $[0,1]$. We globally minimize the energy with a preconditioned primal-dual algorithm [7] which solves certain saddle-point problems efficiently. To this end, we introduce a dual variable $p : V \times T \mapsto \mathbb{R}^4$ which tackles the non-differentiability of the total variation norm:

$$u^* = \arg\min_u E(u) = \arg\min_u \max_{p \in P} \int_{V \times T} \left[\langle p_{\boldsymbol{x}}, D_{\boldsymbol{x}}^{1/2} \nabla_{\boldsymbol{x}} u \rangle + \langle p_t, \nabla_t u \rangle + \lambda \, f u \right] \, d\boldsymbol{x} dt \quad , \tag{7}$$

with set P being defined below. The algorithm converges to the globally optimal solution by iterating a projected gradient descent and gradient ascent for the primal and dual variables respectively. The pointwise update equations are as follows.

$$p^{n+1} = \Pi_P \left[p^n + \sigma \left(D_{\boldsymbol{x}}^{1/2} \nabla_{\boldsymbol{x}} \bar{u}^n \, , \, \nabla_t \bar{u}^n \right)^T \right] \tag{8}$$

$$u^{n+1} = \Pi_{[0,1]} \left[u^n + \tau \left(\operatorname{div} \left((D_{\boldsymbol{x}}^{1/2} p_{\boldsymbol{x}}^{n+1} \, , \, p_t^{n+1})^T \right) - \lambda f \right) \right] \tag{9}$$

$$\bar{u}^{n+1} = 2u^{n+1} - u^n \tag{10}$$

$\Pi_{[0,1]}$ projects u onto the unit interval $[0,1]$ via simple thresholding. The projection onto the set $P = \{ p = (p_{\boldsymbol{x}}, p_t)^T : V \times T \mapsto \mathbb{R}^4 \;\; \big| \;\; \|p_{\boldsymbol{x}}\|_2 \leq 1, \, |p_t| \leq g_t \}$ can be done as follows:

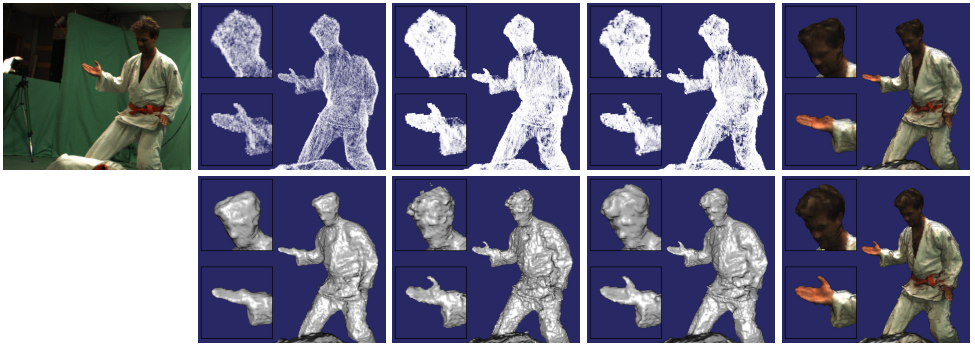
$$\Pi_P(p) = \left(\frac{p_{\boldsymbol{x}}}{\max(1, \|p_{\boldsymbol{x}}\|_2)} , \max\left(-g_t, \min(g_t, p_t) \right) \right)^T \tag{11}$$

Set P can be imagined like a "capsule pill", i.e. a 3D ball shifted along the 4th dimension. The step sizes σ and τ are chosen adaptively by keeping track of the corresponding operator norms as suggested in [7]. Note that the linear operators that transform between primal and dual space contain the discretized differential operators and the diffusion matrix $D_{\boldsymbol{x}}$ which need to be considered for the preconditioning.

For the primal variable u we impose Neumann boundary conditions for both spatial and temporal derivatives and accordingly Dirichlet boundary conditions for the dual variable p, that is, $\nabla_{\boldsymbol{n}} u \Big|_{\partial(V \times T)} = 0$ and $p \Big|_{\partial(V \times T)} = 0$. Finally, we extract an iso-surface of u^* at 0.5 with sub-voxel accuracy for every time step using the Marching Cubes algorithm [15].

5 Implementation

Both the photoconsistency estimation and the surface optimization have been implemented on the GPU using the NVidia CUDA framework. An efficient integration of the anisotropic regularization is challenging because in every point the derivative of the spatially and temporally varying diffusion tensor $D_{\boldsymbol{x}}$ needs to be evaluated based on the normal estimate $\boldsymbol{n}$. A straightforward implementation would easily double the overall memory consumption and render the numerical problem infeasible for reasonable volume resolutions. To save memory we do *not* precompute or save the 3×3 diffusion matrix $D_{\boldsymbol{x}}$, but we recompute $D_{\boldsymbol{x}}$ and its derivative as needed and make use of its symmetry. Further, instead of saving a dense



input | without normal integration [16] | pass 1 | pass 2 (proposed) | comparison column 2 & 4

Figure 2: Effects of the proposed normal integration. Column 2 shows the results without normal integration. The photoconsistency $\rho(x)$ (top) is noisy and less discriminative leading to a reconstruction (bottom) that misses details like the thumb and the hair due to low photometric information. In comparison the photoconsistency for the first pass was denoised with neighborhood information (Eq. (4)). The corresponding reconstruction with isotropic regularization is used to estimate surface normals for the second pass. These normals provide a better estimate than the typically assumed camera-aligned direction used in classical stereo matching. The normals from the first pass further improve photometric scores and fine details (e.g. the thumb) are better preserved due to the anisotropic regularization. The last column compares textured meshes of the results in column 2 and 4. ($|V \times T| = 256^3 \cdot 3$)

normal field for every time step, we store a signed distance function of the previous surface estimate which requires only one additional volume per frame and allows us to densely compute normal estimates as its derivative everywhere in the volume.

As a result, the total amount of required memory per frame is $9 \cdot |V \times T| \cdot 4$ bytes. One volume for the data term, photoconsistency and signed distance function each, two for the primal and four volumes for the dual variable. The second primal variable is needed for the over-relaxation step in Eq. (10). We used $|T| = 3$ and processed longer sequences with a sliding time window approach considering also the frames before and after the current one and took the center frame of the window as the temporally smoothed solution. Further significant memory savings (factor $1/4$ to $1/10$) and speedups (factor 25 to 30) can be achieved by restricting all computations and data structures to the visual hull using indexed lists.

6 Results

We tested our approach on several multi-view sequences with 16 cameras and 1624×1224 image resolution from the INRIA 4D repository [9]. We computed all experiments on a Linux-based PC with a 2.27GHz Xeon CPU, 24GB RAM and an NVidia Titan 6GB graphics card. For quality assessment we compared our method with several state-of-the-art 3D/4D reconstruction methods: PMVS [4] + Poisson surface reconstruction [11], Jancosek and Pajdla [10] and Oswald and Cremers [16]. For all methods we used default parameters, full input image resolution and provided approximate silhouettes if possible (all except [10]).

Fig. 2 shows the influence of normal information on the reconstruction quality in every step of the reconstruction process. For the first pass of our method we used higher standard deviations ($\sigma_n = 0.4, \sigma_t = 0.9$) for the anisotropic Gaussian smoothing kernel in Eq. (4) to achieve a higher denoising of NCC scores from potentially wrong initial normal estimates. The anisotropic smoothing of the NCC scores makes them more discriminative in compari-

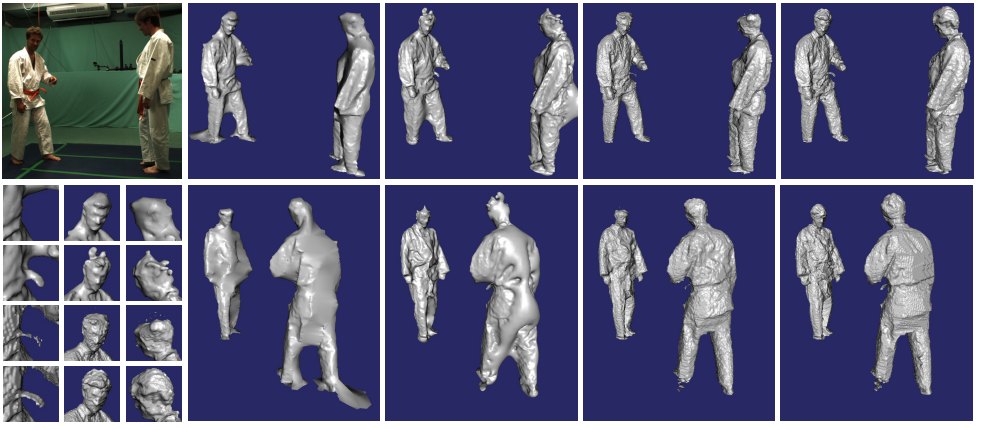


Figure 3: Comparison of our results to other methods. Two views (top/bottom row) of the 'kick one' scene [9] (frame 1) next to an input image and close ups on details. The reconstructions by Jancosek and Pajdla [10] miss many details like the belt and the hand of the left person and parts of both heads (see close ups). Large triangles are generated at locations with low photometric information (bottom row). In contrast, the Poisson reconstruction hallucinates balloonish structures at such locations. The method in [16] yields similar results to the proposed one, but misses fine details like the belt or the hair which is difficult to recover because of noisy photometric information. The proposed normal integration yields superior results ($|V \times T| = 256^3 \cdot 3$).

son to [16] and leads to more distinctive votes (top row). Fine details are only preserved for low smoothness values (bottom row). In the second pass we reduced the Gaussian smoothing ($\sigma_n = 0.3, \sigma_t = 0.7$) to better preserve fine details in the reconstruction. The normal estimates from pass 1 further improve the photoconsistencies (e.g. the hair) and the anisotropic regularization preserves fine details like the thumb also for a higher surface smoothness.

In Fig. 3 we show reconstruction results on a martial art scene in comparison. The method by Jancosek and Pajdla [10] tends to misconnect points which are close but not related to each other. The reconstruction of the left person shows many details on the front, because the method found many inlier points. However, the backside of the left person and most of the right person contains only few triangles which heavily degrade the visual perception of the reconstruction. Generally, this method fails to reconstruct small details and regions with low texture information like the hair or the over-bright cloth section on the shoulders. PMVS [4] performs mostly well in recovering fine details. Since PMVS is a point cloud-based method, point connectivity information is not available for the subsequent Poisson surface reconstruction [11]. This leads to misconnected points and even balloonish surface parts in regions with low photometric information. Moreover, the iterative filtering and expansion approach of [4] in combination with [11] makes the method temporally unstable in sparse camera setups. The method in [16] performs better but cannot fully recover the belt due to bad photometric scores as well as the isotropic regularization scheme which penalizes the surface area and tends to remove thin structures (shrinking bias).

Fig. 4 depicts results of challenging scenes with strong motion blur such as the rope jumping girl or the man with the stick. Our method does not always recover the full geometry, but generally yields better results over full video sequences. Mostly, fine or elongated structures are better preserved such as the fingers of the boy in the cartwheel sequence. Especially, the proposed normal-driven Gaussian smoothing in Eq. (4) yields superior re-

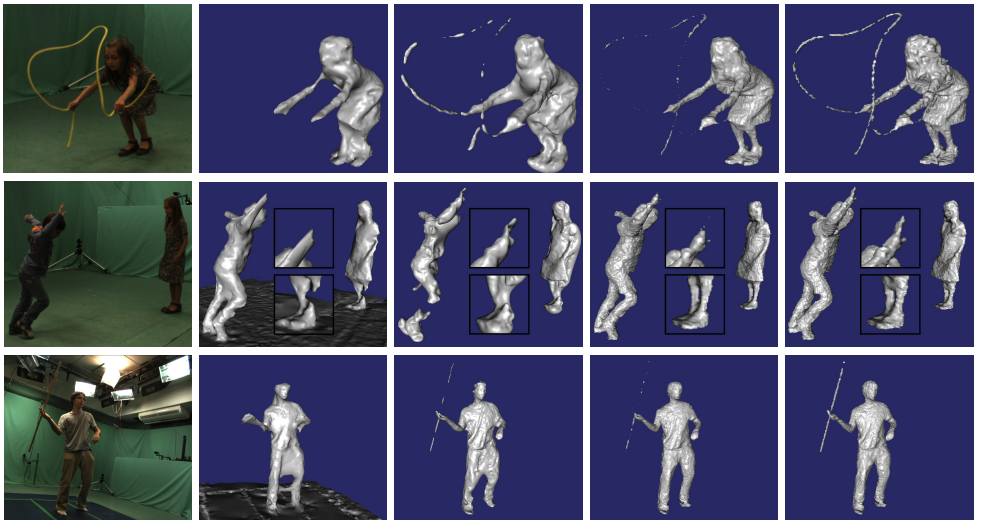


Figure 4: Reconstruction results on different scenes (rope jump, boy cartwheel, stick) from [9]. Although the voxel resolution limits the quality of the rope reconstruction, normal information improves the photometric consistency and helps to better recover fine details in the matching phase and to preserve them during the surface optimization, e.g. the boys thumb or the fast moving stick or rope. Both methods [4, 10] reconstruct frames independently and show severe surface jittering. Enforcing temporal coherence visibly reduces the jittering ($|V \times T| = 384^3 \cdot 3$).

sults in regions with noisy photoconsistency. In particular, the hair is consistently better reconstructed in all sequences we have evaluated. However, in areas where the photometric information is very sparse, the Gaussian smoothing can also degrade the matching score and lead to slightly worse results, e.g. the back of the person in Fig. 3. Essentially, the reconstruction with the isotropic regularization in the first pass only serves as a smoothing of the estimated normal field. Due to the smoothing the recovered normals encode rather low-frequency details of the surface. This is in contrast to the related works mentioned in Sec. 1.1 which estimate normals to better recover the high-frequency details of the surface. However, experiments show that the estimated normals from the first pass can be estimated with moderate effort and improve the photometric matching scores in many surface regions.

Runtime. Depending on the scene the photoconsistency and data term estimation needed about 15-30s per frame. For a volume size $|V| = 384^3$, the isotropic reconstruction in the first pass needed about 1s for $|T| = 1$ and 2-3s for $|T| = 3$. The anisotropic surface estimation in the second pass needed 5s for $|T| = 1$ and 30s for $|T| = 3$. These timings exclude loading and storing from disk and filling data structures. In comparison our method is considerably faster than PMVS+Poisson [4] which needed about 20min/frame for the 'kick one' scene and 6-7min/frame for the 'cartwheel' scene. The method by Jancosek and Pajdla [10] needed about 7-10min/frame. Note that the runtime comparison is only for qualitative information, because all evaluated methods utilize CPU and GPU parallelism in a different manner and have different runtime and memory complexities. Especially the runtime of PMVS [4] is highly data-dependent because of the iterative filtering approach.

7 Conclusion

We showed how surface normal information can be estimated and effectively used within a spatio-temporal multi-view reconstruction setup. Proper estimates of normal information firstly help to improve the accuracy of photometric measures and secondly improve reconstruction results by reducing the shrinking bias of common regularizers. Further, we demonstrated that a modification of the photoconsistency voting scheme [2] improves robustness and quality of the estimated photoconsistencies, making it more similar to methods that determine a regularized fusion of precomputed depths maps. By harnessing the power of consumer graphics cards we showed that an efficient implementation leads to low computation times despite the large amount of data being processed. Numerous experiments showed the improvements of the proposed approach over competitive reconstruction methods.

Acknowledgments. This work was supported by the ERC Starting Grant 'Convex Vision'.

References

[1] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE TPAMI*, 33:1161–1174, 2011.

[2] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *CVIU*, 96(3):367–392, December 2004. ISSN 1077-3142.

[3] Olivier Faugeras and Renaud Keriven. Variational principles, surface evolution, pde's, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7: 336–344, 1999.

[4] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE TPAMI*, 32(8):1362–1376, August 2010. ISSN 0162-8828.

[5] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-view stereo for community photo collections. In *ICCV*, pages 1–8, 2007.

[6] B. Goldluecke and M. Magnor. Space-time isosurface evolution for temporally coherent 3D reconstruction. In *CVPR*, volume I, pages 350–355, July 2004.

[7] B. Goldluecke, I. Ihrke, C. Linz, and M. Magnor. Weighted minimal hypersurface reconstruction. *IEEE TPAMI*, 29(7):1194–1208, July 2007.

[8] J.-Y. Guillemaut and A. Hilton. Space-time joint multi-layer segmentation and depth estimation. In *3DIMPVT*, pages 440–447, 2012.

[9] Institut national de recherche en informatique et en automatique (INRIA) Rhône Alpes. 4d repository, 2014. http://4drepository.inrialpes.fr/.

[10] Michal Jancosek and Tomás Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, pages 3121–3128, 2011.

[11] Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing*, pages 61–70, 2006.

[12] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous global optimization in multiview 3d reconstruction. *IJCV*, 84(1):80–96, August 2009.

[13] K. Kolev, T. Pock, and D. Cremers. Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. In *ECCV*, Heraklion, Greece, September 2010.

[14] Alexander Ladikos, Selim Benhimane, and Nassir Navab. Multi-view reconstruction using narrow-band graph-cuts and surface normal optimization. In *Proc. of the British Machine and Vision Conference*, pages 1–10, 2008.

[15] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21:163–169, August 1987. ISSN 0097-8930.

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

[17] Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *ICCV*, pages 1762–1769, Washington, DC, USA, 2011. ISBN 978-1-4577-1101-5.

[18] Jean-Philippe Pons, Renaud Keriven, and Olivier D. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *IJCV*, 72(2):179–193, 2007.

[19] Christian Reinbacher, Thomas Pock, Christian Bauer, and Horst Bischof. Variational segmentation of elongated volumetric structures. In *CVPR*, 2010.

[20] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0.

[21] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. ISSN 0272-1716.

[22] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *ICCV*, pages 1709–1716, 2009.

[23] H-H. Vu, R. Keriven, P. Labatut, and J.-P Pons. Towards high-resolution large-scale multi-view stereo. In *CVPR*, Miami, Jun 2009.

[24] Xenophon Zabulis and Kostas Daniilidis. Multi-camera reconstruction based on surface normal estimation and best viewpoint selection. In *3DPVT*, pages 733–740. IEEE Computer Society, 2004. ISBN 0-7695-2223-8.

[25] Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *ICCV*, pages 1–8, 2007.