Multi-View Depth Map Estimation With Cross-View Consistency

Jian Wei
jian.wei@graphics.uni-tuebingen.de
Benjamin Resch
benjamin.resch@uni-tuebingen.de
Hendrik P. A. Lensch
hendrik.lensch@uni-tuebingen.de

Computer Graphics
Tübingen University
72076 Tübingen
Germany

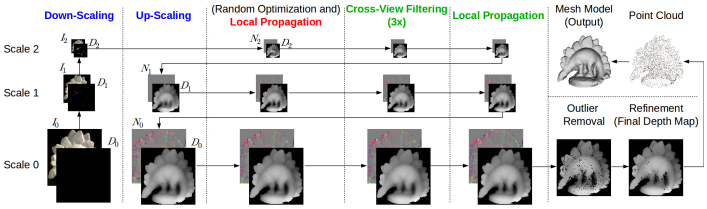

Figure 1: Our processing pipeline for one view of Dino dataset. Our key steps include: hierarchical framework (blue), local propagation (red), and cross-view filtering with an additional propagation pass (green).

Motivation. Multi View Stereo (MVS) aims to establish 3D models from multiple calibrated images. Some works use region growing to estimate depth map per view, and then merge the results. They either only deal with reliable regions, or have difficulty in parallelizing. More crucially, due to the view-independent estimation, inconsistent outliers may exist and grow during propagation, producing unstable estimates across views. This leads to a large amount of estimates removed in the merging stage after consistency checking, and diminishes the reconstruction quality.

To increase robustness of depth-map-based MVS methods, we combine several techniques: Depth estimates are propagated in parallel in the local neighborhood to efficiently spread reliable depth information into regions without prominent structures. A faster coarse-to-fine strategy fills in larger holes. Most importantly, a novel cross-view filtering stage based on free-space constraints and variance filtering, enforces consistency among the depth maps of different views. Our algorithm alternates between correlation and consistency optimization. This way, noisy patches and spikes are excluded so that the subsequent depth map fusion becomes easier.

Workflow. Figure 1 shows our workflow. I_k, D_k, and N_k are the image, depth map, and normal map of a reference view at scale k. I_0 is the input image. Each view selects at most 6 secondary images. Before the first propagation step at each scale, randomly shifted depths and random normals are assigned if smaller matching errors are obtained.

Initialization. For a pixel p, we initialize its depth $D_0(p)$ from bundler if p is feature point; otherwise $D_0(p) = 0$. Its normal $N_k(p)$ including the gradients of the tangent plane in x and y directions, is initialized fronto-parallel at the coarsest scale, *i.e.* $N_2(p) = \{0,0\}$. Before the estimation at each scale, E_k is initialized using the existing depth and normal estimates.

Local Propagation (LP). Good depth and normal estimates are dispersed into the neighborhoods by traversing all pixels if the propagated value improves the correlation measure. The depth hypothesis considers the normal of the tilted patch. Pixels are traversed along parallel scanlines on GPU. We shorten the traversal distance of the work [2] such that more GPU threads can be assigned. In every other iteration vertical and horizontal propagations are applied alternately.

Hierarchical Framework (HF). For textureless regions with few initializations, one propagation alone at the original scale is insufficient due to the locality of short scanlines. We down-scale the depth map and spread the sparse data into neighborhoods. This way, one propagation at the coarsest scale can fill most of the holes. Then the estimates are used for the consecutive finer scale by up-scaling. The overall time is also reduced since the scaling is negligible compared with the speed-up of propagation. We also down-scale the images and up-scale the normal maps.

Cross-View Filtering (CVF). Inspired by the temporally consistent optical flow estimation [3], after local propagation of all views, we perform a cross-view filtering for each reference view to improve the depth consistency. Then a second propagation spreads the optimized estimates.

The projection relationships of pixels between views are considered using the depth information. For each depth value, we find the corresponding pixels in the secondary views, and project them back into the reference view obtaining new depth candidates. These candidates are weighted by the depth difference between the reference and secondary views to get an optimized depth. In some cases, this depth projection from secondary views can even fill holes in the reference, spawning further, more consistent propagation. To avoid slight shifting for some inliers which were accurate before, we additionally check three randomly shifted depth values around the new depth.

Step		Bailer *et al.* [2]	Only LP	LP+HF	LP+CVF	LP+HF+CVF
Downscaling				8.4s		8.4s
1st	Propagation	174.3s	142.2s	13.6s	142.0s	13.6s
	Cross-View Filtering				151.2s	10.8s
	Propagation				226.9s	16.0s
	Upscaling			0.4s		0.4s
2nd	Propagation	1126.6s	880.3s	224.5s	1000.7s	250.0s
	Cross-View Filtering				193.3s	49.2s
	Propagation				951.9s	228.7s
	Upscaling			2.1s		2.1s
3rd	Propagation	418.0s	410.2s	417.4s	450.6s	458.2s
	Cross-View Filtering				204.1s	214.0s
	Propagation				279.9s	280.6s
Outlier removal		42.2s	41.2s	44.2s	48.1s	51.1s
Refinement			121.7s	144.1s	151.3s	189.5s
Overall		1984.4s	1866.8s	1079.1s	4020.3s	1844.1s

Table 1: Timings of each step using Bailer *et al.* [2] and different combinations of our processing steps, when reconstructing all views of Fountain-P11.

Measurement	Bailer *et al.* [2]	Only LP	LP+HF	LP+CVF	LP+HF+CVF	LP+HF+CVF[1]	LP+HF+CVF[2]
Mean Rel. Error ($\times 10^{-3}$) ↓	1.663	1.414	1.236	2.407	1.732	1.505	2.062
Completeness (%) ↑	64.0	63.9	66.9	74.6	79.6	75.9	80.5
Mean Consistency ↑	9.083	9.019	9.124	9.611	9.556	9.253	10.090
Mean Variance ($\times 10^{-6}$) ↓	1.790	1.722	1.626	1.602	1.092	1.179	1.099
Mean Rel. Error of LP+HF+CVF on Pixels of Other Methods ($\times 10^{-3}$) ↓	1.102	1.068	1.142	1.292		1.319	1.368

Table 2: Statistical comparisons for the center view of Fountain-P11 after outlier removal. LP+HF+CVF[1] uses cross-view filtering only for post-processing, and LP+HF+CVF[2] uses propagation-filtering at each scale without the second propagation. The arrows indicate preferred directions.

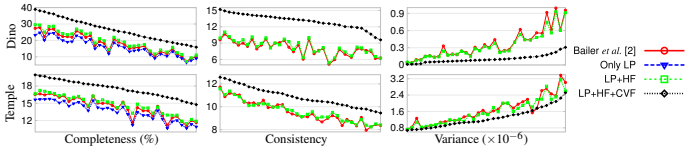

Figure 2: Completeness, mean consistency rating, and mean variance comparisons for some views of Dino and Temple datasets.

Bailer *et al.* [2] (no depth map refinement) | LP+HF (no depth map refinement) | LP+HF+CVF (no depth map refinement) | LP+HF+CVF (with depth map refinement)

Figure 3: Depth maps and 3D models of a region in Sofa dataset after outlier removal and our final results with depth map refinement.

Outlier Removal and Refinement. Inconsistent outliers are filtered out from the resulting depth maps. Results are finally refined by filling the holes and then filtering the noise.

Results. Some results are presented in Tables 1 and 2, as well as Figs. 2 and 3. The relative error evaluates depth accuracy between the estimates and ground truth. The completeness relates the number of recovered pixels to the image size. The consistency [2] and variance (see the paper) measure the multi-view coherence. Combining improved propagation, hierarchical estimation, and iterative multi-view consistency optimization, our method increases the estimation speed, generates dense depth maps with desirable global consistency, and yields convincing 3D reconstruction results. The benchmark results of our full pipeline using the Middlebury evaluation website [1] demonstrate that, our work is competitive with other methods and placed among the most efficient approaches.

[1] Multi-view stereo evaluation. `http://vision.middlebury.edu/mview/`.

[2] C. Bailer, M. Finckh, and H.P.A. Lensch. Scale robust multi view stereo. In *Proc. ECCV*, 2012.

[3] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 31(4), 2012.