

User Directed Multi-View-Stereo

Yotam Doron¹, Neill D.F. Campbell¹, Jonathan Starck², Jan Kautz^{1,3}

¹University College London, ²The Foundry, ³NVIDIA Research

Abstract. Depth reconstruction from video footage and image collections is a fundamental part of many modelling and image-based rendering applications. However real-world scenes often contain limited texture information, repeated elements and other ambiguities which remain challenging for fully automatic algorithms. This paper presents a technique that combines intuitive user constraints with dense multi-view stereo reconstruction. By providing annotations in the form of simple paint strokes, a user can guide a multi-view stereo algorithm and avoid common failure cases. We show how smoothness, discontinuity and depth ordering constraints can be incorporated directly into a variational optimization framework for multi-view stereo. Our method avoids the need for heuristic approaches that edit a depth-map in a sequential process, and avoids requiring the user to accurately segment object boundaries or to directly model geometry. We show how with a small amount of intuitive input, a user may create improved depth maps in challenging cases for multi-view-stereo.

1 Introduction

Multi-view-stereo (MVS) aims to reconstruct the dense geometry of a scene from a set of calibrated images. A large number of MVS methods reconstruct geometry in the form of a *depth map* that assigns a distance value to each pixel in a given input image. Depth maps are a useful representation of geometry with applications in image-based rendering, 3D modeling and augmented reality, and can be merged to create more complete models of a scene [1].

Depth reconstruction is a mature area of research, with methods that perform well on scenes that obey certain assumptions. The first assumption is that the scene geometry has a similar local appearance across different views, and that surfaces are sufficiently textured to distinguish between correct and incorrect geometry under projection. The second main assumption is that depth varies smoothly in regions of low image texture and that discontinuities in depth coincide with strong image edges. These are encoded in the data and smoothness terms common to almost all methods, with data terms based on photo-consistency scores and smoothness achieved by global optimization [2] or local filtering [3].

In reality, these assumptions are only partially valid, as many scenes contain objects with large areas of low-texture, specular reflections and colour discontinuities that are unrelated to changes in geometry.



(a) Source input image (neighboring views not shown)



(b) Depth map from standard method [2]



(c) Depth map after interactive editing with user-in-the-loop

Fig. 1: Overview of the advantage of our user-assisted multi-view stereo. Standard variational approaches to depth map estimation often have shortcomings when presented with real world scenes, for example the specular highlights on the brochure (a) result in holes in the depth map (b). In our approach, we provide the user with interactive tools (in the form of brush strokes) to correct for such short-comings

As a result, fully automatic methods have trouble reconstructing general scenes. Instead, we propose to enable the user to provide intuitive constraints to aid the reconstruction process. In our method we allow the user to apply depth smoothness, depth discontinuity and depth ordering constraints, all of which are specified by the user as simple scribbles. We incorporate the user’s constraints directly into a variational optimization framework that simultaneously optimizes for both a regularized depth map derived from on a photo-consistency cost volume as well as the constraints.

Our work shows how user input can help correct depth reconstruction. In this paper we make the following contributions:

- We introduce a framework to incorporate user constraints, motivating our choice of algorithm by the need for interactive refinement.
- We define a set of simple edit operations to correct common failure cases in multi-view-stereo. Our edit operations are intuitive, do not require the user to draw accurate object boundaries and do not rely on fixed sequential operations such as superpixel segmentation.
- We demonstrate how user constraints may be directly incorporated into a state of the art variational depth reconstruction algorithm and show how this can be used to correct depth reconstruction.
- We provide a clear derivation of our solver from the energy formulation.

2 Related Work

User-guided monocular shape and depth recovery In scenarios such as 2D-3D conversion, a user may want to recover geometry given their knowledge of the image contents. Several techniques exist to interpolate empirical depth or disparity values painted onto monocular image or video data, while others use relative constraints on depth [4]. Guttmann *et al.* [5] train a classifier on user-scribbled disparity values provided on the first and last frames of a shot, using the results to constrain a coarse-to-fine quadratic optimization. Brosch *et al.* [6] also require users to paint disparity values onto the first and last frame of a sequence. Their technique attempts to maintain consistent depth over time by combining the scribbles with the results of a video over-segmentation algorithm and optical flow. In our approach, rather than directly painting depth values, the user indirectly improves the depth computed by multi-view-stereo by providing constraints.

The Depth Director system of Ward *et al.* [7] uses superpixel data, along with sparse structure-from-motion information to assign depth to video frames. The system provides a segmentation tool, allowing users to adjust depth variation and orientation within a region. Additionally, users may choose a template shape, such as a car, to fit to a selected region. Liao *et al.* [8] ask users to provide relative depth input (ordering and equality) at keyframes. The system also uses depth from SFM features where available and applies a temporal perspective depth correction stage based on optical flow. Their user-study suggests that it

is more intuitive for users to specify the relative depth constraints than to paint colour-coded empirical depth values.

High level user interactions have been applied to intrinsic image decomposition and geometric modelling. Two recent methods [9, 10] include in their optimization user-provided local constraints: constant-reflectance, constant-illumination and fixed-illumination. In single-image modelling, Zhang *et al.* [11] optimize for a mesh satisfying user-provided normal, positional and curvature constraints. Toeppe *et al.* [12] propose a variational framework for single-view single-object modelling using a ballooning energy formulation. The recent work of Chen *et al.* [13] demonstrates modelling of complex objects with simple inputs, using symmetry constraints and automatic alignment to image structures.

User-guided stereo The recent system of Zhang *et al.* [14] aims to improve the quality of disparity maps generated from a stereo video sequence by allowing a range of user edits. The user corrects disparity at keyframes indirectly with several interactive tools which alter unary and pair-wise cost terms in an label-based optimization framework. The system relies on a GrabCut-based object selection tool [15] to segment objects. The user may fit a parametric model, using sparse feature matches within a selected object, a disparity alignment tool, a smoothness brush and a discontinuity brush. The paper also presents a method for propagation of edited disparity through the sequence, with the edited disparity map and disparity costs used soft constraints in non-keyframes. This approach is related to ours in that the user edits also influence energy terms in an optimization, however our constraints do not require the user to first accurately segment object boundaries, and we target the multi-view-stereo case rather than (binocular) stereo video.

Photogrammetry There has been extensive work on image-based modelling using photogrammetry [16], with recent methods [17] providing high quality results from inaccurate user input. In the VideoTrace system [18], a user interactively creates a model from video footage. The interactions are assisted by pre-computed sparse stereo and superpixel information and the user may define symmetries to complete geometry in areas not visible in the footage.

3 Our Approach

The input to our system is a sequence of images $\{I_1 \cdots I_N\}$ for which the camera motion has been calculated using, for example, structure-from-motion with an off-the-shelf tool. This provides us with a projection matrix P_i defining the camera pose, and intrinsic parameters such as focal length, at each image (or frame of a video sequence).

Given this calibrated image sequence, the goal of our method is to estimate a dense depth map for one or more of the images. Without loss of generality, we shall discuss calculating the depth for a single image at a time; this image shall be referred to as the *source* or *reference* image and denoted I_s . For the source

frame I_s , a subset of adjacent frames are used to automatically reconstruct the source frame depth D_s .

In this section we provide details on our approach to interactive depth reconstruction. We begin by providing details of how to build a photo-consistency cost volume that encodes appearance constraints from neighboring views. We then describe the basic energy model that uses this cost volume, along with regularisation enforcing smoothness constraints, to define the most likely depth map. Next, we add user constraints to modify the energy function in a principled manner to improve the quality of the depth map in an interactive refinement process with the user in the loop. Finally, we provide details of the optimization scheme used to solve our energy model, including the user constraints, in an efficient manner that allows for interactive editing.

3.1 Photo-Consistency

Our depth estimation process first builds a photo-consistency cost volume. The cost volume acts as a cached data term in our optimization, as in [2] and is computed before any user interactions are introduced. We test a range of candidate depths for each pixel x in the source image I_s , spaced linearly in inverse depth (disparity). Our photo-consistency error for the pixel x taking a candidate depth d is defined as

$$C(x, d) = \frac{1}{K} \sum_{I_j \in \mathcal{N}(I_s)} |I_s(x) - I_j(\pi_{s,j}(x, d))| \quad . \quad (1)$$

Here we have used $\pi_{s,j}(x, d)$ to denote the back-projection from I_s to a depth d along a ray through x , followed by projection into image coordinates in a neighbouring frame I_j , and, $I_j(x')$ to denote the interpolated colour at pixel x' in I_j . In addition, $\mathcal{N}(I_s)$ denotes the the local neighbourhood of frames around I_s , and K normalizes by the number of frames where the re-projected pixel is inside the image bounds.

3.2 Energy

Taking the lowest cost per pixel in the photo-consistency cost volume, *i.e.*

$$d(x) = \arg \min_x C(x, d) \quad , \quad (2)$$

would provide us with a very noisy estimate of the depth map; instead, [19, 2], we apply spatial regularization using a variational energy formulation that combines the photo-consistency cost with a term that encourages gradient sparsity in the depth map. Formally, we define our energy as

$$E[d(x)] = \int_{\Omega} \lambda C(x, d(x)) + g(x) \|\nabla d(x)\|_{\epsilon} \, dx \quad (3)$$

where we use the photo-consistency cost from (1) and a Total Variation (TV) prior on the gradient of the depth map, $\nabla d(x)$, and integrate over the image

domain Ω . The TV term is weighted by the inhomogeneous contrast sensitive term

$$g(x) = \exp(-\gamma \|\nabla I_s\|) \quad (4)$$

that encourages depth discontinuities to coincide with image intensity discontinuities. We use the Huber norm

$$\|\mathbf{s}\|_\epsilon = \begin{cases} \frac{\|\mathbf{s}\|^2}{2\epsilon} & \text{if } \|\mathbf{s}\| \leq \epsilon \\ \|\mathbf{s}\| - \frac{\epsilon}{2} & \text{if } \|\mathbf{s}\| > \epsilon \end{cases} \quad (5)$$

on the depth map gradient. For clarity, we will drop the dependency on x in our notation such that $d(x) \mapsto d$ and $g(x) \mapsto g$.

The energy in (3) is non-convex and difficult to optimize directly. To overcome this issue we use a quadratic relaxation, similar to [20] and [2]. We introduce an auxiliary depth variable, $v(x) \mapsto v$, and approximate (3) with the auxiliary energy $E_{\text{aux}}[d, v]$ as

$$\int_{\Omega} g \|\nabla d\|_\epsilon + \frac{1}{2\theta} (d - v)^2 + \lambda C(x, v) \, dx \quad (6)$$

We observe that as $\theta \rightarrow 0$ we will have $v \rightarrow d$ and thus (6) \rightarrow (3).

By decoupling the regularization from the data term, we obtain two sub-problems. Fixing $v = v'$, we have a problem $\min_d E_{\text{aux}}[d, v']$ that is convex in d as

$$\min_d \int_{\Omega} g \|\nabla d\|_\epsilon + \frac{1}{2\theta} (d - v')^2 \, dx \quad (7)$$

and by fixing $d = d'$ we have a problem $\min_v E_{\text{aux}}[d', v]$ that can be solved point-wise for v as

$$\min_v \int_{\Omega} \frac{1}{2\theta} (d - v)^2 + \lambda C(x, v) \, dx \quad (8)$$

We will show how to solve this alternation optimization in § 3.4 using a primal-dual saddle point technique; before this, we provide details of how to extend this standard energy model to include user constraints.

3.3 Including User Constraints

We now describe how to extend the basic energy model to include the user in the reconstruction process (user-in-the-loop). This takes the form of the user providing brush strokes from a toolbox of three constraints targeted against specific shortcomings of the standard variational approaches. We now describe each of these three tools in further detail.

Smoothness There are regions where the photo-consistency values in the cost volume may be noisy or incorrect, *e.g.* the Lambertian assumption fails in the presence of specular highlights, or image intensity discontinuities may encourage artificial depth discontinuities. In such regions we can smooth the solution from neighboring regions containing the correct depth by downweighting the photo-consistency term and relying on the gradient regularizer to fill in a smooth surface.

We maintain a brush bitmap $B_{\text{sm}}(x) \in [0, 1]$ and use it to modulate the weight on the data term in (3); we allow λ to vary across the image and set it to $\lambda = \alpha(1 - B_{\text{sm}})$, again dropping the explicit dependency on x for clarity. We used feathered brush strokes to ensure a smooth transition. We note that we must modify the sub-problem in (7) since the quadratic relaxation means that the depth is still influenced by v . We therefore include λ in the coupling term as

$$\min_d \int_{\Omega} g \|\nabla d\|_{\epsilon} + \frac{\lambda}{2\theta} (d - v')^2 dx \quad (9)$$

such that when $\lambda \rightarrow 0$ the data term is decoupled and the regularization will take over and smooth the resulting depth map.

Boundary Discontinuities There are the opposite cases where the solution will be too smooth in regions where there should be a discontinuity in the depth; for example, foreground and background objects with similar coloring may result in a low contrast intensity edge disguising a true discontinuity in depth. To tackle this problem we provide a second brush bitmap $B_{\text{dc}}(x) \in [0, 1]$ that increases the contrast sensitive edge term of (4) as

$$g(x) = \exp \left(-\gamma (1 + \mu B_{\text{dc}}) \|\nabla I_s\| \right) . \quad (10)$$

We also need to downweight the data term in these regions since the photo-consistency term can lead to the phenomenon of foreground ening [3]. Whilst previous approaches have addressed this with adaptive support weights or including view selection in the optimization [21], this requires changing or re-computing the cost volume and removes the efficiency advantages of pre-caching the photo-consistency costs. Instead, we make use of the discontinuity brush and again downweight the λ term again such that

$$\lambda = \alpha (1 - B_{\text{sm}}) (1 - B_{\text{dc}}) . \quad (11)$$

Ordering Constraints Errors in the photo-consistency volume can give rise to the an incorrect local minimum where even if smoothness and discontinuity constraints are preserved, a distinct surface may appear in the wrong layer (either too close or too far from the camera). Our third tool makes use of two brush strokes where a user can select two nearby image regions and apply the constraint that one is closer to the camera than the other. Multiple instances of such pairwise constraints can be built up as necessary given the interactive reconstruction feedback available to the user.

We will illustrate with a single pair of brush strokes; we define a foreground brush $B_{\text{fore}}(x)$ and corresponding background brush $B_{\text{back}}(x)$ with the constraint that all the foreground brush pixels $\{x_{f,i}\} \in B_{\text{fore}}$ are closer to the camera than the background brush pixels $\{x_{b,j}\} \in B_{\text{back}}$. We proceed by matching each foreground pixel $x_{f,i}$ to the nearest pixel in the background set $x_{b,m(i)}$ such that

$$m(i) = \arg \min_{j \in B_{\text{back}}} \|x_{f,i} - x_{b,j}\| \quad . \quad (12)$$

This can be performed efficiently using a k -d tree. We then form a set of linear inequality constraints with a minimum threshold distance in depth t_{dist} that must separate the two layers which gives us that

$$\Phi [d] + t_{\text{dist}} \mathbf{1} < \mathbf{0} \quad (13)$$

where $\Phi [d]$ denotes

$$\begin{bmatrix} 0 & \dots & -1 & \dots & 1 & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \cdot \\ d(x_{f,i}) \\ \cdot \\ d(x_{b,m(i)}) \\ \cdot \end{bmatrix} \quad . \quad (14)$$

We can apply this constraint to the energy model using a set of Lagrangian multipliers $\mathbf{r} \in \mathcal{R}^{|\{x_{f,i}\}|}$ to augment sub-problem (9) to

$$\min_d \max_{\mathbf{r}} \int_{\Omega} g \|\nabla d\|_{\epsilon} + \frac{\lambda}{2\theta} (d - v')^2 + \langle \mathbf{r}, (\Phi [d] + t_{\text{dist}} \mathbf{1}) \rangle dx \quad (15)$$

and maximizing with respect to \mathbf{r} such that $\mathbf{r} \geq 0$.

3.4 Optimization

In section § 3.2 we described how to split our energy model into two sub-problems to be solved in alternation. We first consider the sub-problem, with the auxiliary variables v fixed, of (15); this can be solved with a primal-dual approach [22].

Auxiliary Sub-Problem Taking (15), we first dualize the regularisation term $f(d) = g \|d\|_{\epsilon}$, with $g > 0$. The Legendre-Fenchel transform of $f(\cdot)$ is given by

$$f^*(p) = g \max_d \{g^{-1} \langle d, p \rangle - \|d\|_{\epsilon}\} \quad (16)$$

$$= \frac{\epsilon}{2g} \|p\|^2 + \delta \left(\frac{p}{g} \right) \quad (17)$$

where $\delta(\cdot)$ is the indicator function

$$\delta(p) = \begin{cases} 0 & \text{if } \|p\| \leq 1 \\ \infty & \text{otherwise} \end{cases} \quad (18)$$

and we use the scaling property

$$f(x) = a h(x) \implies f^*(p) = a h^*\left(\frac{p}{a}\right) \quad (19)$$

for $a > 0$. We then add a dual variable $\mathbf{q}(x) \mapsto \mathbf{q}$ and obtain the saddle point problem $\int_{\Omega} L dx$ as

$$\int_{\Omega} \langle \mathbf{q}, \nabla d \rangle - \frac{\epsilon}{2g} \|\mathbf{q}\|^2 - \delta \left(\frac{\mathbf{q}}{g} \right) + \frac{\lambda}{2\theta} (d - v)^2 + \langle \mathbf{r}, (\Phi [d] + t_{\text{dist}} \mathbf{1}) \rangle dx \quad (20)$$

that we minimize with respect to d and maximise with respect to \mathbf{q} and \mathbf{r} . Taking partial derivatives we obtain

$$\frac{\partial L}{\partial \mathbf{q}} = \nabla d - \frac{\epsilon}{g} \mathbf{q} \quad (21)$$

$$\frac{\partial L}{\partial \mathbf{r}} = (\Phi [d] + t_{\text{dist}} \mathbf{1}) \quad (22)$$

$$\frac{\partial L}{\partial d} = -\nabla \cdot \mathbf{q} + \frac{\lambda}{\theta} (d - v) . \quad (23)$$

We then discretize for descent in d and ascent in \mathbf{q} and \mathbf{r} and solve for \mathbf{q}^{k+1} , \mathbf{r}^{k+1} and d^{k+1} as

$$\mathbf{q}^{k+1} = g \frac{(\mathbf{q}^k + \sigma_q \nabla d^k)}{(g + \sigma_q \epsilon)} \quad (24)$$

$$\mathbf{r}^{k+1} = \mathbf{r}^k + \sigma_r (\Phi [d] + t_{\text{dist}} \mathbf{1}) \quad (25)$$

$$d^{k+1} = \frac{(\theta d^k + \sigma_d (\theta \nabla \cdot \mathbf{q}^{k+1} + \lambda v))}{(\theta + \lambda \sigma_d)} \quad (26)$$

with appropriate step sizes σ_q , σ_r and σ_d , as in [23]; the updates are shown in Algorithm 1. We note that two projection steps are required. The \mathbf{q} updates require projection into the norm ball, represented by the operation $\pi[\cdot]$. The \mathbf{r} updates require projection into the positive half-plane, represented by $I^+[\cdot]$.

Cost Volume Sub-Problem The second sub-problem, with the primal variables d fixed, of (8) may be solved using a simple point-wise search in the cost volume. As noted in [2], this search may be accelerated by keeping track of a search depth range for each pixel over subsequent iterations. We also perform a single Newton step in v in each iteration, as in [2], to obtain sub-sample accuracy within the cost volume and reduce depth quantization artefacts.

Efficiency The overall optimization process is given in Algorithm 1. We note that each update step can be performed efficiently and in parallel for each pixel. As demonstrated by the results of [2], this allows for real-time performance when implemented on the GPU. In addition, the cost volume calculation (Equation 1) may be performed in parallel on the GPU using the texture units to perform efficient image interpolation.

Algorithm 1 Depth optimization

$$d^0(x) \leftarrow \min C(x, d(x))$$

$$v^0(x) \leftarrow d^0(x)$$

$$n \leftarrow 0$$

while not converged **do**

Projected gradient ascent for dual variable

$$\mathbf{q}^{k+1} \leftarrow g \pi \left[(g + \sigma_q \epsilon)^{-1} g \{ \mathbf{q}^k + \sigma_q \nabla d^k \} \right]$$

Projected gradient ascent for constraint multipliers

$$\mathbf{r}^{k+1} \leftarrow I^+ \left[\mathbf{r}^k + \sigma_r (\Phi [d] + t_{\text{dist}} \mathbf{1}) \right]$$

Gradient descent step for primal variable

$$d^{k+1} \leftarrow \left(1 + \frac{\lambda \sigma_d}{\theta} \right)^{-1} \left(d^k + \sigma_d (\nabla \cdot \mathbf{q}^{k+1} + \frac{\lambda v}{\theta}) \right)$$

Exhaustive point-wise search for auxiliary variable

$$v^{k+1} \leftarrow \min_v \left(\lambda C(x, v^k) + \frac{\lambda}{2\theta} (d^{k+1} - v^k)^2 \right)$$

Update coupling parameter

$$\theta \leftarrow \theta(1 - \beta n)$$

$$n \leftarrow n + 1$$

end while

return d

4 Results

We demonstrate the results of our system, showing the effect of the different user interactions on the reconstructed depth. In this paper we show a reconstruction of the flower and lawn dataset from Zhang *et al.* [24] (in Figures 3 and 4) and a desk scene (in Figure 2) that features a number of violations of standard MVS assumptions. In our experiments, we select the 8 closest views to the reference frame and sample 100 depth values to construct the cost volume. For the flower and lawn scenes, we use the camera pose provided by [24] and for the desk scene we use the NUKEX camera tracker [25].

Figure 2 shows the impact of the different interactions compared to the result of the baseline variational method [2]. The desk scene features a glossy non-Lambertian surface with strong image edges that do not coincide with depth discontinuities. The resulting hole in the depth map is corrected with smoothness constraints in our method. The discontinuity edits reduce the depth smearing on the top edge of the cloth in the desk scene due to low image contrast.

Figures 3 and 4 compare our results for the flower and lawn datasets to the depth maps computed by [24] and to the results of baseline variational method [2] without any additional constraints. The user annotations are shown in the last row. In the flower and lawn datasets, the method of [24] produces smooth-looking

depth maps but loses a significant amount of detail, while the baseline method [2] is able to recover more detail but suffers from edge fattening artefacts and in large areas of low texture. Our method allows the user to selectively maintain detail, for example in the leaves of the flower, while imposing smoothness in other areas of the image, such as the gravel and the bottom edge of the image. In the lawn dataset, the severe artefacts in the sky and the boundaries of the figure and bench are improved.

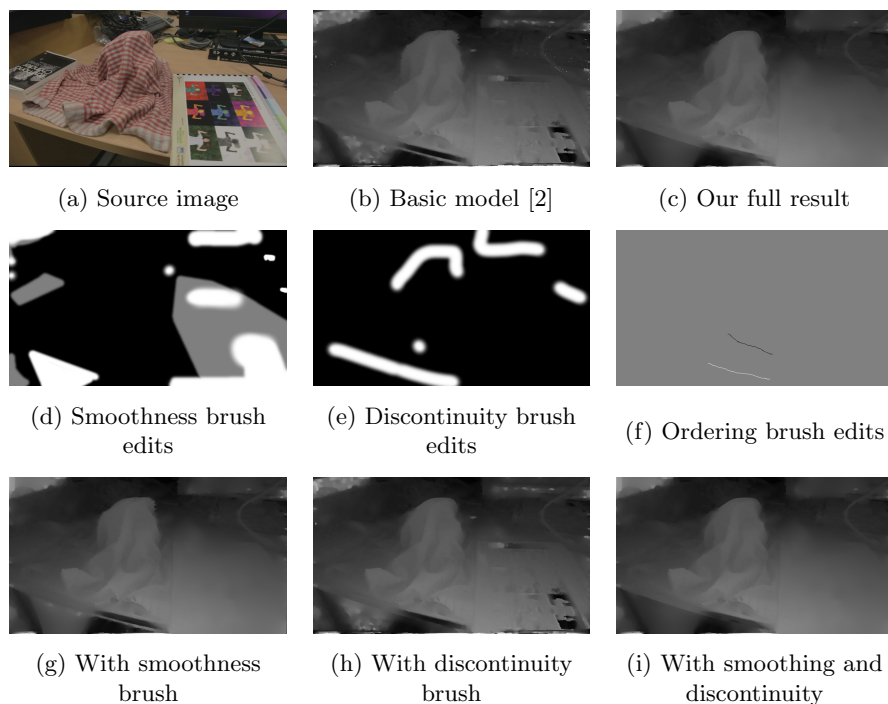


Fig. 2: Our results for the desk dataset. We demonstrate that considerable improvement can be made over the basic model by adding the user to the reconstruction loop. We show our full result in (c) for comparison to the basic model result in (b). We have filled in the holes in the reconstruction and improved the quality and ordering of the discontinuities. We show the individual brush strokes in (d)-(f) and results for subsets of the brush strokes in (g)-(i).

5 Conclusion

In this paper we have shown how simple user annotations may significantly improve the quality of depth maps reconstructed from multiple images. Our method builds on state-of-the-art continuous depth recovery algorithms with

the annotations modifying terms in the energy formulation, and does not require users to draw precise object boundaries or to paint absolute depth values. We believe this is particularly well-suited to mobile devices, where user input is limited to coarse strokes and image-based-rendering applications such as depth defocus are gaining popularity.

6 Acknowledgements

We would like to thank Anastasios Roussos and Fabio Viola for helpful discussions. This work has been supported by the EngD VEIV Centre at UCL, The Foundry and EPSRC grant EP/I031170/1.

References

1. Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: 10th European Conference on Computer Vision. Volume 5302 of LNCS. (2008) 766–779
2. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense Tracking and Mapping in Real-Time. In: ICCV. (2013)
3. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '11, Washington, DC, USA, IEEE Computer Society (2011) 3017–3024
4. Sýkora, D., Sedlacek, D., Jinchao, S., Dingliana, J., Collins, S.: Adding Depth to Cartoons Using Sparse Depth (In)equalities. *Computer Graphics Forum* **29** (2010) 615–623
5. Guttman, M., Wolf, L., Cohen-Or, D.: Semi-automatic stereo extraction from video footage. In: ICCV. (2009)
6. Brosch, N., Rhemann, C., Gelautz, M.: Segmentation-based depth propagation in videos. *ÖAGM/AAPR* (2011)
7. Ward, B., Kang, S.B., Bennett, E.P.: Depth director: A system for adding depth to movies. *IEEE Computer Graphics and Applications* **31** (2011) 36–48
8. Liao, M., Gao, J., Yang, R., Gong, M.: Video stereolization: Combining motion analysis with user interaction. *Visualization and Computer Graphics, IEEE Transactions on* **18** (2012) 1079–1088
9. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. In: *ACM Transactions on Graphics (TOG)*. Volume 28., ACM (2009) 130
10. Shen, J., Yang, X., Jia, Y., Li, X.: Intrinsic images using optimization. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE (2011) 3481–3487
11. Zhang, L.Z.L., Dugas-Phocion, G., Samson, J.S., Seitz, S.: Single view modeling of free-form scenes. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* **1** (2001)
12. Toeppe, E., Oswald, M., Cremers, D., Rother, C.: Image-based 3d modeling via cheeger sets. In: *Asian Conference on Computer Vision (ACCV). HONORABLE MENTION*. (2010)

13. Chen, T., Zhu, Z., Shamir, A., Hu, S.M., Cohen-Or, D.: 3-Sweep: Extracting Editable Objects from a Single Photo. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2013)* **32** (2013) Article 195
14. Zhang, C., Price, B.: High-Quality Stereo Video Matching via User Interaction and Space-Time Propagation. *3DTV-Conference, 2013 ...* (2013)
15. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": Interactive foreground extraction using iterated graph cuts. In: *ACM SIGGRAPH 2004 Papers. SIGGRAPH '04*, New York, NY, USA, ACM (2004) 309–314
16. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-based Approach. In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '96*, New York, NY, USA, ACM (1996) 11–20
17. Arikian, M., Schwärzler, M., Flöry, S., Wimmer, M., Maierhofer, S.: O-Snap: Optimization-Based Snapping for Modeling Architecture. *ACM Transactions on Graphics* **32** (2013) 6:1—6:15
18. van den Hengel, A., Dick, A., Thormählen, T., Ward, B., Torr, P.H.S.: VideoTrace: rapid interactive scene modelling from video. In: *ACM Transactions on Graphics (TOG)*. Volume 26., ACM (2007) 86
19. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **40** (2011) 120–145
20. Steinbrücker, F., Pock, T., Cremers, D.: Large displacement optical flow computation without warping. In: *ICCV*. (2009)
21. Zheng, E., Dunn, E., Jovic, V., Frahm, J.M.: Patchmatch based joint view selection and depthmap estimation. (2014)
22. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision* **40** (2011) 120–145
23. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: *ICCV*. (2011)
24. Zhang, G., Jia, J., Wong, T.T., Bao, H.: Consistent depth maps recovery from a video sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31** (2009) 974–988
25. The Foundry: NUKEX. <http://www.thefoundry.co.uk/products/nuke-product-family/nukex/> (2014)

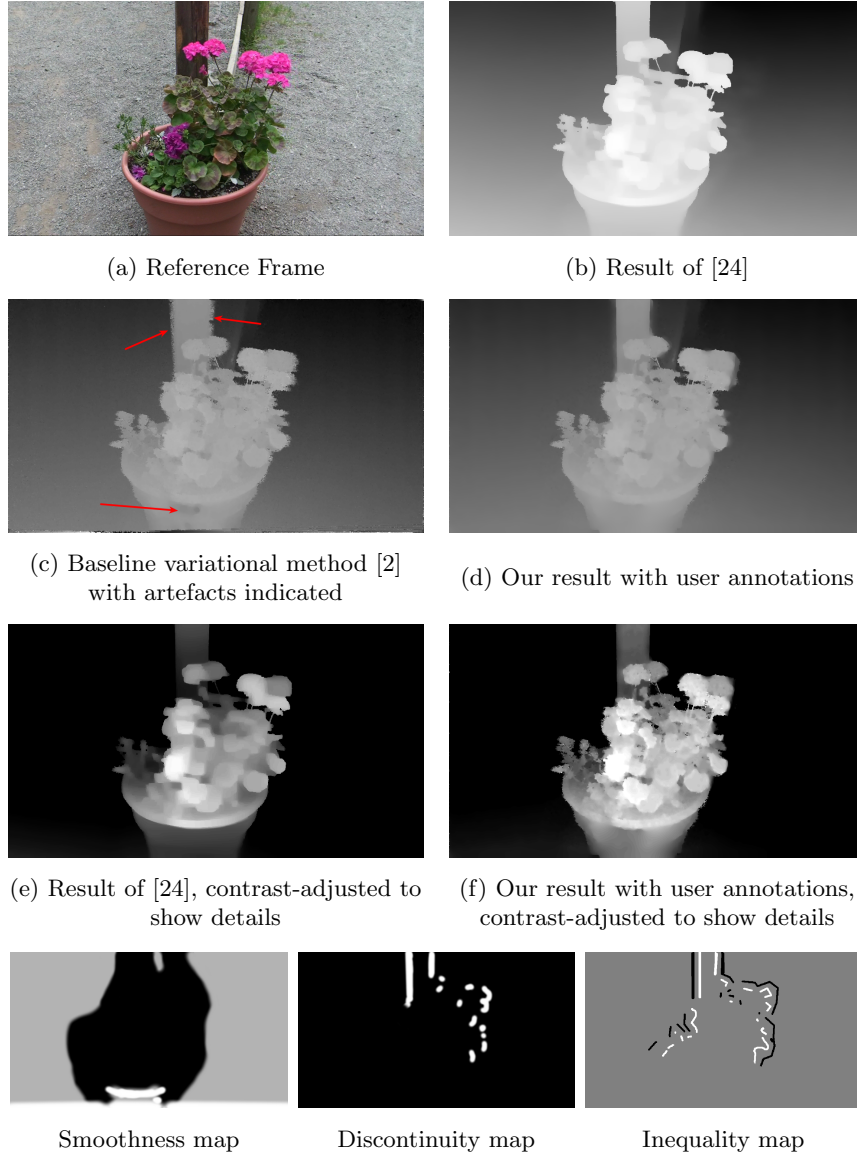


Fig. 3: Results on the flower dataset of [24]. By controlling the smoothness and discontinuity terms locally, a user can obtain a smooth result while selectively maintaining fine detail, illustrated here by the detail in the leaves. The depth maps in (e) and (f) were adjusted globally with a colour-curve tool for visualisation.

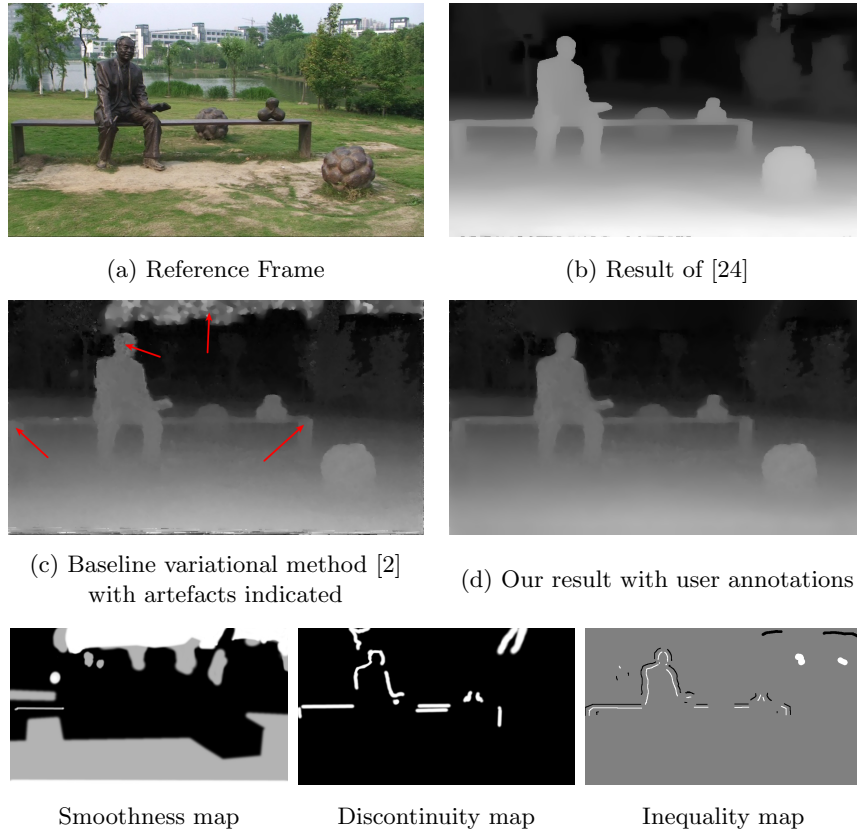


Fig. 4: Results on the lawn dataset of [24]. Both [24] and [2] fail to recover depth discontinuities in areas of low contrast between the bench and the lawn. The baseline method cannot determine depth for the textureless sky region, while [24] loses detail in the background trees and enlarges foreground objects such as the legs and sides of the bench. These problem areas are improved with user annotations.