Integrating Approximate Depth Data into Dense Image Correspondence Estimation

Kai Ruhl, Felix Klose, Christian Lipski, and Marcus Magnor
Computer Graphics Lab, TU Braunschweig
Muehlenpfordtstrasse 23, 38106 Braunschweig, Germany
{ruhl,klose,lipski,magnor}@cg.tu-bs.de [*]

ABSTRACT

High-quality dense image correspondence estimation between two images is an essential prerequisite for many tasks in visual media production, one prominent example being view interpolation. Due to the ill-posed nature of the correspondence estimation problem, errors occur frequently for a number of problematic conditions, among them occlusions, large displacements and low-textured regions. In this paper, we propose to use approximate depth data from low-resolution depth sensors or coarse geometric proxies to guide the high-resolution image correspondence estimation. We counteract the effect of uncertainty in the prior by exploiting the coarse-to-fine image pyramid used in our estimation algorithm. Our results show that even with only approximate priors, visual quality improves considerably compared to an unguided algorithm or a pure depth-based interpolation.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Depth cues, Motion*

Keywords

dense image correspondence estimation, optical flow, depth sensor, geometric proxy, view interpolation

1. INTRODUCTION

In visual media production, dense image correspondences are used for a multitude of purposes, among them view interpolation of spatio-temporal in-between frames [5]. Interpolation can occur either using a geometric proxy or image-based, as in our case. We rely on high-quality dense image correspondences to achieve visually plausible results. Erroneous correspondences lead to artifacts in the rendered frames which have to be fixed manually; therefore, it is preferable to reduce errors in correspondence space.

Contemporary dense image correspondence estimation algorithms have come a long way in the last decade and continue to improve [1].

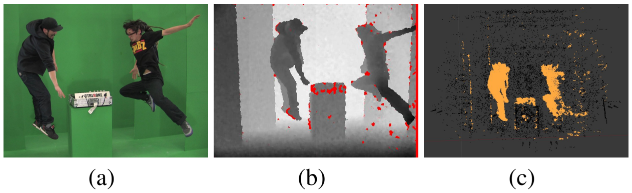


Figure 1: Input data example: approximate depth data from multiple unsynchronized Kinects, to be used as uncertain prior to image correspondence estimation. (a) HD camera image (b) VGA depth map (c) depth points projected into world space.

However, due to the ill-posed nature of the problem, errors cannot be eliminated completely. Frequently problematic areas include occlusions, which a 2D flow simply cannot resolve; large displacements of small regions, which require a large search radius and vanish in the coarse-to-fine pyramid, if one is used; and low-textured regions, which rely solely on the smoothness term to be resolved.

We propose a two-stage process combining geometric information and dense image correspondence estimation. In the first stage, we obtain approximate depth data from arbitrary sources, and then use the depth information as an uncertain prior to guide the subpixel precise correspondence estimation in the second stage.

Depth data can be obtained using stereo algorithms. In the case of static scenes or synchronized cameras, this is very similar to the original correspondence estimation problem, and unfortunately also subject to the same problem areas.

Active light sensors such as time-of-flight cameras [14] or structured light sensors like the Microsoft Kinect [20] are not affected by these issues. Unfortunately, their limitations include comparatively low resolution (typically VGA and less), which in conjunction with high-resolution camera frames requires upsampling or other means of adaptation [28]. Limited range and vulnerability to non-lambertian surfaces must also be considered.

Building geometric proxies by hand is another method to generate depth information usable for view interpolation. While geometric proxies are often a byproduct of visual media production [9], increasing the geometric detail is a time-consuming task.

Correspondences obtained from those depth sensor or proxy data are able to resolve occlusions, large displacements and low-textured regions well. However, many detail errors remain (Section 3).

In the second stage, we therefore employ a guided dense image correspondence estimation algorithm to improve the results. Rather than starting with the correspondences from the first stage, we instead treat it only as a highly uncertain prior. Our algorithm is based

[*] Video and further paper details available under:
`http://graphics.tu-bs.de/publications/ruhl2012cvmp/`

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CVMP '12, December 5 – 6, 2012, London, United Kingdom
Copyright 2012 ACM 978-1-4503-1311-7/12/11 ...$15.00.

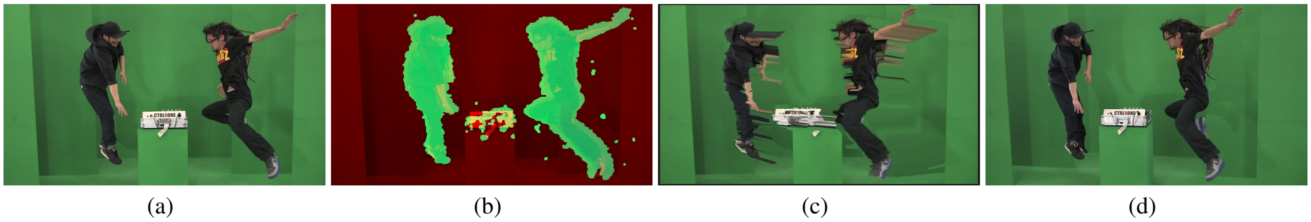


Figure 2: Stage 1: Dense image correspondence prior calculated from approximate depth data gained by unsynchronized devices. (a) source image (b) source image in red channel, approximate correspondences from (background subtracted) depth data in green channel (c) source image forward-warped by approximate correspondences. while large displacements, occlusion and low-textured regions are solved well, detail errors are common (d) "ground truth" target image.

on a variational formulation and uses a coarse-to-fine image pyramid [34]. This allows us to integrate the priors on coarse pyramid levels where the effects of an erroneous recommendation are less prominent and easier to compensate (Section 4).

In our evaluation, we use wide-baseline examples which are hard to solve using 2D image correspondence estimation algorithms. We show that our approach is able to compensate prior mismatches stemming from calibration and temporal misalignments, sensor resolution or coarse proxy definition, and thus reduces the need for frame-by-frame image domain corrections (Section 5).

Our main contribution is the successful *integration of imprecise priors* marred by (a) low resolution of depth sensors (b) coarse definition of geometric proxies (c) subframe temporal offsets of unsynchronized cameras or (d) calibration inaccuracies. We show that *uncertainty compensation* of depth guides is possible to some extent in the context of dense image correspondence estimation.

2. RELATED WORK

A number of research areas are pertinent to our work.

Optical flow. Dense image correspondence estimation and optical flow are areas with large overlap. Being less constrained than stereo reconstruction based on epipolar lines, it can also handle unsynchronized cameras. A survey on recent algorithms was performed by Baker et al. [1]. Contemporary algorithms typically achieve subpixel precision by Taylor linearization, e.g. Zach et al. [34] or dense sampling, e.g. Steinbruecker et al. [32]. Long-range correspondence estimation, particularly of small objects, is also an active research area. Brox et al. [4] address large displacements using pre-segmentation and Lipski et al. [18] use large scale belief propagation. Liu et al. [19] use manual segmentation to improve the flow. Our stage 2 is based on optical flow.

Depth sensors. Real-time depth sensor research has until recently been focused mostly on time-of-flight cameras. A good survey was undertaken by Kolb et al. [14]. Due to low resolution and system-immanent errors, upsampling and correction of depth maps is required, e.g. [28, 22, 7]. More recently, low-cost structured-light sensors such as the Microsoft Kinect [30] have become mainstream. Newcombe et al. [21] continually register the pose of a Kinect to achieve super-resolution. Kuster et al. [15] combine two Kinects and three synchronized cameras for depth-image-based rendering. We use depth sensors as one possible source of depth data.

Geometric models. Using geometric models as a means for estimating image correspondences is a wide research area, and an exhaustive survey is outside the scope of this paper. Seitz et al. [29] performed a survey of static multi-view stereo algorithms. Large-scale reconstruction of static outdoor scenes can be based on point or patch models [8, 11]. To capture the motion and structure of a scene, the notion of scene flow was introduced by Vedula [33]. Multiple precomputed optical flow fields can be merged [36, 33], static 3D reconstructions at discrete time steps can be registered to recover the motion data [35, 24, 25], or unsynchronized cameras can be used for offset-tolerant reconstruction [13]. We use geometric models as another possible source of depth data.

View interpolation. Common approaches to view interpolation include depth-image based rendering [6] or pure image-based rendering [5]. The first excels for clear object boundaries and Lambertian scenes, while the second copes well with amorphous structures such as a fireball and other effects that are hard to model geometrically. Germann et al. [10] represent soccer players as a collection of articulated billboards. Ballan et al. [2] use image-based view interpolation with billboards to represent a moving actor. Lipski et al. [17] employ multi-image morphing. Image interpolation has been used in various movies, e.g. in "2 Fast 2 Furious" or "Swimming pool" [26], with all corrections of visual artifacts performed manually in the image domain.

3. STAGE 1: APPROXIMATE DEPTH

In the first stage of our process, we acquire approximate image correspondences from depth sensors or geometric proxies. These approximate correspondences are later used as an uncertain prior.

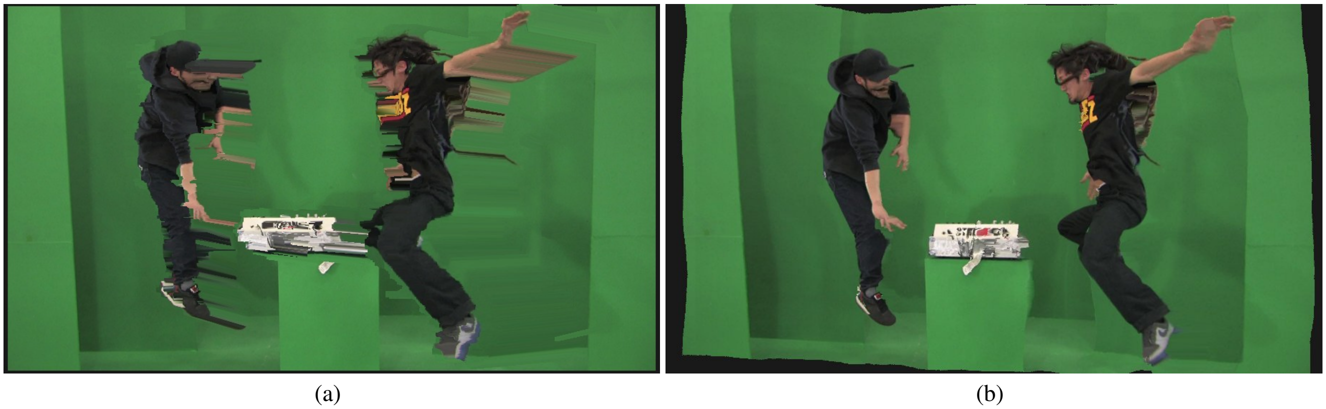
For a depth sensor scenario, we use the data set by Lipski et al. [16], which includes two HD cameras and two Kinects at wide baselines. A large distance between the depth sensors increases 3D scene coverage and decreases interference of the structured light patterns, down to a depth pixel loss of less than 10 percent, see Fig. 1 (b).

Cameras are calibrated by sparse bundle adjustment [31], and depth images can be included e.g. by using a mirror-diffuse checkerboard [3] or coded markers [15]. As the scene calibration is only accurate up to scale, we use the known physical distance between the cameras to determine the metric scale of the scene.

Following the ROS project evaluation [27], we convert the 11-bit disparity values supplied by the Kinects to metric depth values $z(d) = (8\ b\ f)(d_{\text{off}} - d)^{-1}$ with $b \approx 7.5$cm as baseline, f the focal length, and $d_{\text{off}} \approx 1090$ the disparity offset. We reduce the number of depth points by background subtraction. We then use the intrinsic matrix $\mathbf{K}_i$ and extrinsic matrix $\mathbf{M}_i$ obtained by the sparse bundle adjustment to project the depth points of device i into world space:

$$\mathbf{X} = (\mathbf{K}_i \mathbf{M}_i)^{-1}\, \mathbf{x} \tag{1}$$

where $\mathbf{X} = [X\ Y\ Z\ 1]^T$ is a homogenous coordinate in world space and $\mathbf{x} = [x\, y\, z\ 1]^T$ a homogeneous depth point coordinate in the local space of a depth sensor i. A projection example in Blender



(a)

(b)

Figure 3: Stage 1 vs. 2: Approximate prior vs. estimation guided by approximate prior. (a) source image warped directly by approximate prior (b) source image warped after dense image correspondence estimation guided by the approximate prior. the large-displacement, occlusion and low-texture matching properties have been preserved while detail errors are much less present.

is shown in Fig. 1 (c). We perform the reverse transformation in order to reproject the world space points back into different cameras j:

$$\tilde{\mathbf{x}} = (\mathbf{K}_j \mathbf{M}_j)\ \mathbf{X} \tag{2}$$

with $\tilde{\mathbf{x}} = [\tilde{x}\ \tilde{y}\ \tilde{z}\ 1]^T$ being a depth point in the local space of camera j. To estimate approximate (non-dense) image correspondences $\tilde{\mathbf{u}} = [\tilde{u}\ \tilde{v}]^T$ between frames from cameras j and k, we calculate the 2D difference between reprojected points $\tilde{\mathbf{x}}_j$ and $\tilde{\mathbf{x}}_k$:

$$\tilde{\mathbf{u}}_{j \to k} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_j \tag{3}$$

For a geometric proxy scenario, we use the data set by Hasler et al. [12], which includes two HD cameras at wide baselines. We manually construct a coarse geometry, then start with world coordinates and perform the same procedure beginning at Eq. 2.

Due to either low resolution, coarse proxy geometry, subframe temporal offset or calibration inaccuracies, the resulting flow field $\tilde{\mathbf{u}}_{j \to k}$ is not perfect, as shown in Fig. 2. Regions and borders have only been coarsely matched, as evident in (b). Consequently, a fully warped source image is imprecise in many locations, see (c) vs. (d). However the main objective, namely resolving large displacements, occlusions and low-textured regions, has been generally solved well.

4. STAGE 2: DEPTH AS GUIDE

In the second stage of our process, we take the not-necessarily-dense approximate depth-based flow $\tilde{\mathbf{u}}$ as a prior and strive to use it for refinements inside a dense image correspondence algorithm.

We base our algorithm on the optical flow by Zach et al. [34], a variational formulation using a coarse-to-fine image pyramid. Our source image is denoted by I_1 and the target image by I_0 with coordinates $\mathbf{x} = [x, y]^T$, and we strive to attain the flow $\mathbf{u} = [u_x, u_y]$ such that $I_1(\mathbf{x} + \mathbf{u}(\mathbf{x})) = I_0(\mathbf{x})$, where $I_{\{0,1\}}(\mathbf{x})$ is a brightness value and $\mathbf{u}(\mathbf{x})$ the two-dimensional displacement vectors. The employed image pyramid has levels $L \in [0..n_L]$, with 0 the finest (highest) and n_L the coarsest (lowest) image resolution.

We minimize the following overall energy (Eq. 12 in [34]):

$$E = \int_\Omega |\nabla \mathbf{u}| + \frac{1}{2\theta}(\mathbf{u} - \mathbf{v})^2 + \lambda |\rho(\mathbf{v})| d\mathbf{x} \tag{4}$$

where $\mathbf{u}$ and $\mathbf{v}$ both represent the flow to be computed and are auxiliaries to each other, the regularizer $|\nabla \mathbf{u}|$ (with gradient ∇) rewards smoothness of the flow field, the residual $|\rho(\mathbf{v})|$ rewards adherence to the brightness constancy (data term) and λ is a weight relating data and smoothness term.

The coupling term $\frac{1}{2\theta}(\mathbf{u} - \mathbf{v})^2$ enforces closeness of $\mathbf{u}$ and $\mathbf{v}$, allowing the algorithm to perform alternate updates to $\mathbf{u}$ and $\mathbf{v}$ (Eq. 13 and 15 in [34]), with θ a small constant. Following convergence, $\mathbf{u}$ is equal or very close to $\mathbf{v}$.

Considering the data term in more detail, ρ is defined as the difference between the warped source image I_1 and the target image I_0. In order to make the function locally convex, a first order Taylor expansion is applied:

$$\begin{aligned} \rho(\mathbf{u}) &= I_1(\mathbf{x} + \mathbf{u}) - I_0(\mathbf{x}) \\ &\approx I_1(\mathbf{x} + \mathbf{u_0}) + \langle \nabla I_1(\mathbf{x}), \mathbf{u} - \mathbf{u_0} \rangle - I_0(\mathbf{x}) \end{aligned} \tag{5}$$

For this, the flow $\mathbf{u}$ is subdivided into a fixed part $\mathbf{u_0}$ and a differentiable part $\mathbf{u} - \mathbf{u_0}$ which is optimized pointwise along ∇I_1 (with $\langle\rangle$ being the scalar product). Since Taylor expansion is only valid for small distances, a coarse-to-fine warping scheme is employed where $\mathbf{u_0}$ is the upsampled flow from a coarser level. The smoothness term $|\nabla \mathbf{u}|$ is already a convex function, so no further modification is required.

This gives us two avenues to integrate our approximate depth prior $\tilde{\mathbf{u}}$ on any given level L: Either we modify the formulation of ρ, or we modify $\mathbf{u_0}$.

The straightforward idea is an integration into the data term; however, the hidden uncertainties inside the prior make this infeasible. We start with the observation that while the prior accuracy has high uncertainty at full resolution $L = 0$ and the same uncertainty at lower levels, the uncertainty *when measured in pixels* decreases as L increases. Therefore, it is more favorable to include the prior at the bottom of the image pyramid at lower resolutions to reduce the effect of uncertain priors.

The subpixel-precise data term update is performed in a thresholding step on ρ (Eq. 15 in [34]):

$$\mathbf{v} = \mathbf{u} + \begin{cases} +\lambda\theta\nabla I_1 & \text{if } \rho(\mathbf{u}) < -\lambda\theta|\nabla I_1|^2 \\ -\lambda\theta\nabla I_1 & \text{if } \rho(\mathbf{u}) > +\lambda\theta|\nabla I_1|^2 \\ -\rho(\mathbf{u})\frac{\nabla I_1}{|\nabla I_1|^2} & \text{otherwise} \end{cases} \tag{6}$$

where $\pm\lambda\theta|\nabla I_1|^2$ is the radius within which small update steps

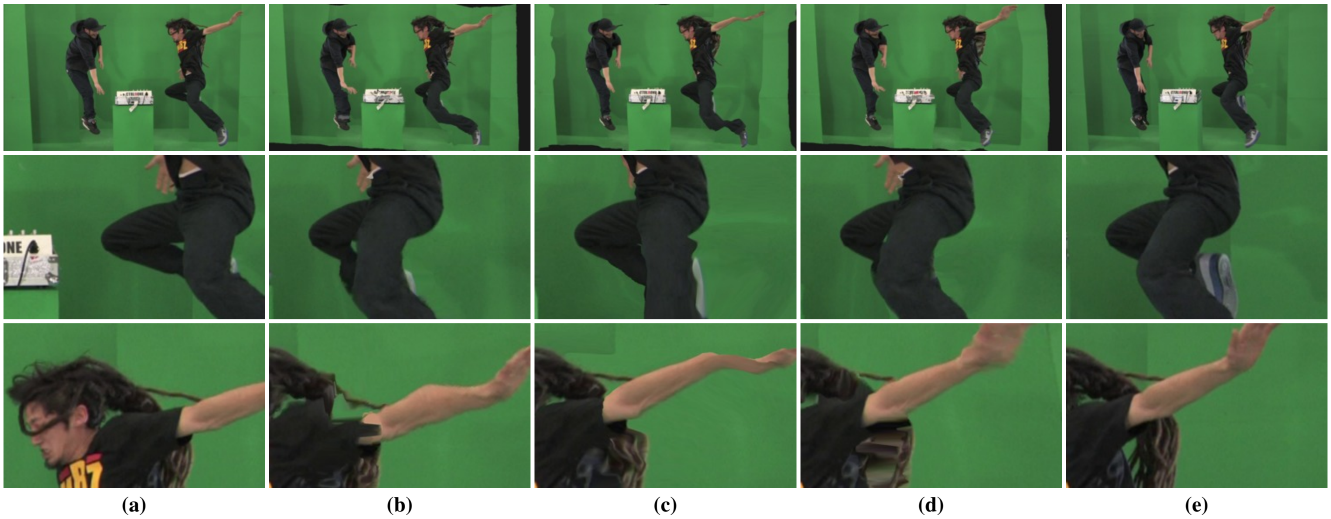


Figure 4: "Who Cares" scene from Lipski et al. [16] (a) source image I_1 (b) unguided TV-L1 (c) large displacement optical flow [4] (d) approximate-depth guided TV-L1 (e) target image I_0. Both unguided algorithms are unable to resolve occlusions and large displacements correctly. Approximate depth hints coupled with automated refinement are able reduce artifacts greatly.

are taken, and outside which larger steps are necessary.

Since we have no way to determine the uncertainty per prior pixel, an integration into the data term and thus the thresholding step potentially leads the data update astray. In contrast, when modifying an initialization $\mathbf{u_0}(\mathbf{x})$ at a level L, all subsequent data update iterations on that level can refine that initialization. The prior $\tilde{\mathbf{u}}(\mathbf{x})$ can have a maximum uncertainty of $\pm\lambda\theta|\nabla I_1|^2$ that the thresholding step can compensate in small steps (option 3 in Eq. 6).

We therefore choose the latter option, performing reprojections onto $\tilde{\mathbf{u}}$ for each pyramid level separately as in Eq. 3, then applying the prior locally for a selection described by a mask $\mathbf{m_p}$.

$$\mathbf{u_0}(\mathbf{x}) = \begin{cases} \tilde{\mathbf{u}}(\mathbf{x}) & \text{if } \mathbf{m_p}(\mathbf{x}) \neq 0 \\ \mathbf{u_0}(\mathbf{x}) & \text{otherwise} \end{cases} \tag{7}$$

where $\mathbf{m_p}$ is a boolean mask requiring the prior $\tilde{\mathbf{u}}(\mathbf{x})$ to exist locally, and where an optional prior mask $\mathbf{m_{\tilde{u}}}$ can also be incorporated. Additionally, we integrate one more uncertainty-reducing measure: Acknowledging that the data term ρ is, in good cases, more accurate than the prior $\tilde{\mathbf{u}}$, we omit the replacement when the residual is already very low. In total, $\mathbf{m_p}$ is defined as:

$$\begin{aligned} \mathbf{m_p}(\mathbf{x}) &= \tilde{\mathbf{u}}(\mathbf{x}) \neq 0 \\ &\wedge \mathbf{m_{\tilde{u}}}(\mathbf{x}) \neq 0 \\ &\wedge |\rho(\mathbf{x})| > \rho_t \end{aligned} \tag{8}$$

with ρ_t a small user-defined constant, default $\rho_t = 0.01$.

Fig. 3 shows an example comparing direct application of a depth-based prior $\tilde{\mathbf{u}}$ against the results of a dense image correspondence estimation merely guided by $\tilde{\mathbf{u}}$. The characteristics of the depth-based prior – correct correspondences for large displacements, occlusions and low texture (a) – have been retained, while detail errors have been considerably reduced (b).

5. RESULTS

We demonstrate our approach with scenarios including either depth sensors or geometric proxies. The scenes are challenging for optical flow algorithms, exhibiting wide camera baselines, low texture, occlusions and large displacements of small, deformable objects.

In the figures, warped images have been rendered using a dense mesh with one vertex per pixel, with the vertices having been displaced according to the flow field. Though a blend of forward and backward warping produces more compelling results, we only use forward warping for a more meaningful visual inspection.

Fig. 4 shows the "Who Cares" data set from Lipski et al. [16]. The two HD cameras are 10 degrees and 1 meter, the two Kinects 20 degrees and 2 meters apart. The cameras and depth sensors are not subframe synchronous. Since masks are available, we use them for prior selection as in Eq. 8, but for evaluation purposes not in the rendered frames. Priors are integrated into the lower half of the image pyramid only, from level $\frac{n_L}{2}$ until the coarsest level n_L. We compare our approximate-depth guided TV-L1 to an unguided TV-L1 as in Zach et al. [34] and to a large displacement optical flow (LDOF) by Brox et al. [4].

Due to the wide baseline, both unguided algorithms have severe issues with the right actor, see Fig. 4 (b) and (c); particularly arm and leg which have very large displacements compared to their size. The knee, with its black-on-black occlusion, is also hard to resolve due to lack of texture. The approximate-depth-guided TV-L1 improves the situation in all regards compared to the unguided approach, see (d). Note however that not all details have been resolved, e.g. the right fingertips of the left actor. Still, the guided approach outperforms the unguided one in terms of visual quality also in this case.

Fig. 5 shows the free climber data set from Hasler et al. [12]. We use two of the four hand-held HD cameras and a coarse geometric proxy from which a depth map is reprojected into the two cameras. No masks are being used, and priors are again integrated into the lower half of the image pyramid from level $\frac{n_L}{2}$ to n_L. We compare TV-L1 and LDOF to a direct proxy-based flow field and to our guided TV-L1 approach.

Again due to the very wide baseline, the unguided algorithms have severe issues with both climber and wall, see Fig. 5 (b) and (c). Applying the approximate depth-based prior directly is as good as the coarse geometric proxy, but exhibits many detail errors, e.g. skinny legs, cut trousers, see (d). The approximate-depth-guided TV-L1 repairs many but not all of the detail errors, see (e). While

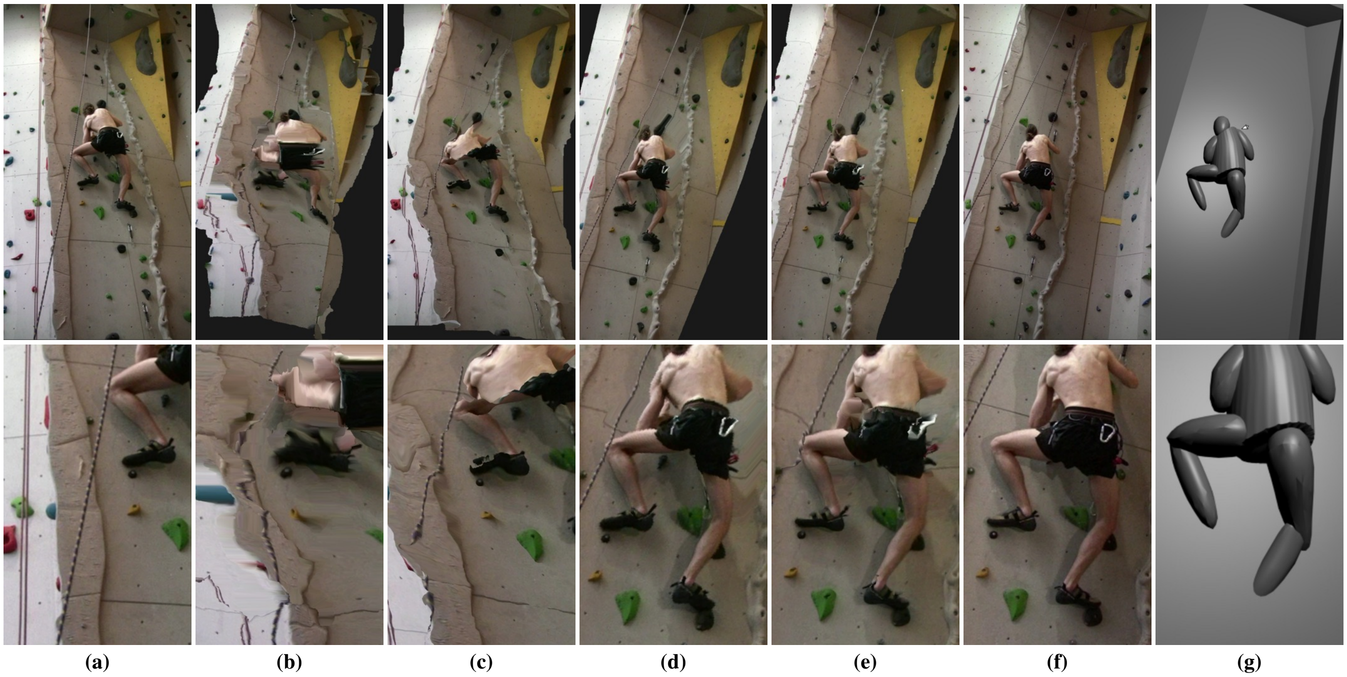


Figure 5: Free climber scene from Hasler et al. [12] (a) source image I_1 (b) unguided TV-L1 (c) large displacement optical flow (d) direct approximate depth (e) approximate-depth guided TV-L1 (f) target image I_0 (g) geometric proxy. In unguided form, both algorithms cannot follow the wide baseline. Direct proxy-depth based warping shows good results, but details such as the legs are not correctly captured. Approximate depth-based priors coupled with TV-L1 refinement eliminates many detail artifacts.

the improvements are not as prominent as in the previous example, the guided approach overall leads to a warped image requiring considerably reduced correction effort.

6. DISCUSSION

In the shown challenging scenes, the approximate-depth guided approach resolves detail issues that unguided algorithms have difficulties with. In scenes with a more moderate challenge level, we found that contemporary optical flow algorithms find good solutions, so that guidance is not required. We combine advantages from both worlds: The global structure qualities of depth data, and the sub-pixel precision of optical flow.

In a visual media production, geometric proxies are often already present and can be used immediately in our guided approach, which refines the missing details automatically. Geometric detail can always be added to improve the results even further. Future high-resolution depth sensors would also be beneficial.

Our approach is limited by the trade-off between depth guide quality and the repair capabilities of the underlying optical flow algorithm. Higher uncertainty means fewer levels where the prior can be applied; this in turn can degrade results.

As the evolution of optical flow algorithms continues, integrating our approach into ever more sophisticated and fault-tolerant formulations becomes another result-improving avenue.

7. CONCLUSIONS

We presented a dense image correspondence estimation approach that integrates approximate depth data, and compensates its inherent uncertainty. Instead of making the input data more accurate, we accept its inaccuracies and apply the resulting prior only where benefits outweigh risk of false input. Our approach improves on the results of unguided algorithms, and offers help in regions that are systematically problematic: occlusions, large displacements, and low texture.

8. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 256941, Reality CG, and from the German Science Foundation DFG MA 2555/1-3. We would also like to thank the GPU4vision project [23] and Zach et al. [34] for making the TV-L1 source code openly available; Brox et al. [4] for openly providing the LDOF binaries; and Hasler et al. [12] for sharing their data set.

9. REFERENCES

[1] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE Computer Society, 2007.

[2] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3):87, July 2010.

[3] K. Berger, K. Ruhl, C. Brümmer, Y. Schröder, A. Scholz, and M. Magnor. Markerless motion capture using multiple color-depth sensors. In *Proc. Vision, Modeling and Visualization (VMV) 2011*, pages 317–324, Oct. 2011.

[4] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:41–48, 2009.

[5] S. E. Chen and L. Williams. View interpolation for image synthesis. In *Proc. of ACM SIGGRAPH'93*, pages 279–288. ACM Press/ACM SIGGRAPH, 1993.

[6] C. Fehn. A 3D-TV approach using depth-image-based rendering (dibr). In *Proc. of VIIP*, volume 3, 2003.

[7] A. Frick and R. Koch. Improving depth discontinuities for depth-based 3dtv production. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pages 1–4. IEEE, 2011.

[8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2009.

[9] FXguide. Art of stereo conversion: 2D to 3D - 2012. `http://www.fxguide.com/featured/art-of-stereo-conversion-2d-to-3d-2012/`.

[10] M. Germann, A. Hornung, R. Keiser, R. Ziegler, S. Würmlin, and M. Gross. Articulated billboards for video-based rendering. *Comput. Graphics Forum (Proc. Eurographics)*, 29(2):585, 2010.

[11] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[12] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H. Seidel. Markerless motion capture with unsynchronized moving cameras. In *Computer Vision and Pattern Recognition, 2009*, pages 224–231. IEEE, 2009.

[13] F. Klose, C. Lipski, and M. Magnor. Reconstructing shape and motion from asynchronous cameras. In *Proc. Vision, Modeling and Visualization (VMV) 2010*, pages 171–177, Siegen, Germany, 2010.

[14] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight sensors in computer graphics. *Eurographics State of the Art Reports*, pages 119–134, 2009.

[15] C. Kuster, T. Popa, C. Zach, C. Gotsman, M. Gross, P. Eisert, J. Hornegger, and K. Polthier. Freecam: A hybrid camera system for interactive free-viewpoint video. In *Vision, Modeling, and Visualization (VMV)*, pages 17–24, 2011.

[16] C. Lipski, F. Klose, K. Ruhl, and M. Magnor. Making of who cares HD stereoscopic free viewpoint video. In *Proc. European Conference on Visual Media Production (CVMP) 2011*, volume 8, pages 1–10, Nov. 2011.

[17] C. Lipski, C. Linz, K. Berger, A. Sellent, and M. Magnor. Virtual video camera: Image-based viewpoint navigation through space and time. *Computer Graphics Forum*, 29(8):2555–2568, 2010.

[18] C. Lipski, C. Linz, T. Neumann, and M. Magnor. High resolution image correspondences for video Post-Production. In *CVMP 2010*, pages 33–39, London, 2010.

[19] C. Liu, W. Freeman, E. Adelson, and Y. Weiss. Human-assisted motion annotation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[20] Microsoft Corporation. Kinect for xbox 360, November 2010. Redmond WA.

[21] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *10th Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136. IEEE, 2011.

[22] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011.

[23] U. of Graz. Gpu4vision project. http://www.gpu4vision.org/.

[24] J. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, 2005.

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

[26] E. Powell. Is frame interpolation important? *Projector Central, Whitepaper*, 2009.

[27] ROS. Kinect calibration guide. `http://www.ros.org/wiki/kinect_calibration/technical`, 2010.

[28] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. High-quality scanning using time-of-flight depth superresolution. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08.*, pages 1–7. IEEE, 2008.

[29] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *Computer Vision and Pattern Recognition, IEEE*, 1:519–528, 2006.

[30] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition*, volume 2, page 7, 2011.

[31] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 835–846. ACM, 2006.

[32] F. Steinbruecker, T. Pock, and D. Cremers. Large displacement optical flow computation without warping. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1609–1614, Kyoto, Japan, 2009.

[33] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:475–480, 2005.

[34] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Pattern recognition: 29th DAGM symposium*, volume 29, pages 214–223, 2007.

[35] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 367–374, 2003.

[36] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In *Proc. of CVPR'01*, volume 2, pages 778–785. IEEE Computer Society, 2001.