VARIATIONAL MULTI-IMAGE STEREO MATCHING

Simon Donné, Bart Goossens, Jan Aelterman, Wilfried Philips

Ghent University
iMinds - IPI - UGent
{Simon.Donne, Bart.Goossens, Jan.Aelterman, philips}@telin.ugent.be

ABSTRACT

In two-view stereo matching, the disparity of occluded pixels cannot accurately be estimated directly: it needs to be inferred through, e.g., regularisation. When capturing scenes using a plenoptic camera or a camera dolly on a track, more than two input images are available, and – contrary to the two-view case – pixels in the central view will only very rarely be occluded in all of the other views. By explicitly handling occlusions, we can limit the depth estimation of pixel $\vec{p}$ to only use those cameras that actually observe $\vec{p}$. We do this by extending variational stereo matching to multiple views, and by explicitly handling occlusion on a view-by-view basis. Resulting depth maps are illustrated to be sharper and less noisy than typical recent techniques working on light fields.

Index Terms— stereo, light field, depth estimation

1. INTRODUCTION

Depth maps for scenes can be acquired in several ways. An active measurement of scene depth is possible using time-of-flight cameras; yet these are still limited in spatial resolution. In stereo matching techniques, the disparity $u(x, y)$ links pixels on row y and column x in image I_C to pixels in image I_R representing the same scene point. When the cameras share their viewing plane and are connected by a line segment parallel to their scan-line direction, the images are said to be rectified and the correspondences lie on the same scan line:

$$I_C(x, y) = I_R(x + u(x, y), y) \tag{1}$$

The actual depth of a given pixel is inversely proportional to its disparity (from trigonometry, e.g. as in figure 1).

As technology advances, it becomes more and more feasible to capture scenes with more than two cameras, for example with a plenoptic camera [1], which results in a light field representation of the scene [2]. As more cameras observe the scene, all this information can improve the disparity/depth estimation for the scene.

This work was performed within the iMinds ASPRO+ project and "Multi-camera human behavior monitoring and unusual event detection" (G0398111N). Simon Donné is funded by BOF grant 01D21213 and Bart Goossens is a postdoctoral researcher for FWO.

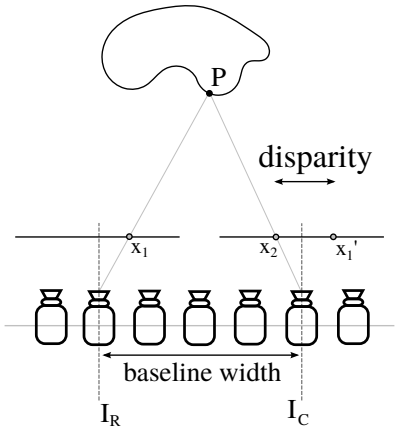

Fig. 1: The central camera and the reference camera observe the same point P at x_1 and x_2. The orthogonal distance from the baseline to point P can be calculated when the baseline width, disparity and focal distance of the cameras are known.

In this paper we evaluate the extension of the variational stereo matching [3] approach to a light field set-up as in figure 1. We start with an overview of existing work. In section 3 we formulate the problem mathematically and describe our proposed method. Finally, results are presented and the conclusion is drawn.

2. EXISTING WORK

Traditional approaches to (two-image) stereo matching consist of computing large cost fields which are aggregated over neighbourhood windows [5, 6] to achieve spatial regularisation. For each pixel, a cost per candidate disparity value is computed. These cost fields are then aggregated over spatial neighbourhoods after which the best disparity for each pixel is chosen. Extensions include the use of image transforms, such as the census transform [7, 3] or a normalized cross-correlation [8], to make the cost field more robust to illumination changes and other influences, e.g. vignetting. Our proposed method will allow for multiple image transforms to be used simultaneously, in order to allow us to complement the strengths of one transform with those of another as in [9].

Fig. 2: Pixels are unlikely to be occluded in all views, because the occlusion occurs in opposite directions in case of a central view. From left to right: the central view, the pixels not visible in the right view and the pixel invisible in the left one. The input is a detail from the dolls sequence of Middlebury [4].

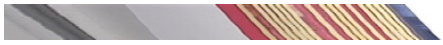

Fig. 3: An example slice through a light field showing the same scan line in all of the views. From top to bottom in the image the camera is moving from left to right. The closer to vertical its trajectory is, the closer a pixel is to the baseline.

Borrowing from optical-flow techniques such as [8], a reformulation of stereo matching results in a more direct optimisation of the disparity map [3] in a coarse-to-fine approach. The authors of [10] adopt a similar approach in order to accurately estimate disparity maps from light fields. Our proposed method will explicitly take occlusions into account.

An alternative approach to computing depth from light fields is presented in [11]: in a light field representation, the depth of a pixel is equivalent to the slope of the line it traces in the light field, illustrated in figure 3. Through visual evaluation we illustrate that our proposed method results in smoother depth maps while respecting image boundaries. This results in more visually pleasing interpolated views: when generating new views, one of the most important factors is the adherence to depth discontinuities in the scene.

While plenoptic cameras generally have a two-dimensional *grid* of cameras (due to the microlenses), the focus of our research lies on sequences recorded with a camera mounted on a track. In such cases, the camera trajectory is restricted to a single dimension. The extension of the work presented here to a two-dimensional grid is relatively straightforward.

3. MULTI-VIEW STEREO

We wish to estimate the disparity map of a central view $I_C(x, y)$ based on a one-dimensional light field comprising the central camera and K additional views $I_k(x, y)$ (see figure 1). Trigonometry shows that the ratio of disparities for two different cameras relative to the central camera is directly proportional to the ratio of these cameras' (signed) distances to the central camera.

This means that, when denoting the location of the k^{th} camera by θ_k, the relation is:

$$I_C(x, y) = I_k(x + \theta_k u(x, y), y), \forall k \in [1, K]\,. \qquad (2)$$

This formulation assumes a one-dimensional light field. In the case of a two-dimensional light field this model should similarly include a vertical shift in correspondence matches. In this paper we will restrict ourselves to a single dimension.

Ideally, the estimate of the disparity map fulfils equation (2) in all pixels of the image. However, this is not the case in areas of occlusion (where two pixels in $I_k(x, y)$ are mapped onto the same pixel in $I_C(x, y)$) or simply because the pixel lies outside of the bounds of $I_k(x, y)$. Two-view methods generally use regularisation to estimate the disparity of such pixels [3].

In the case of a one-dimensional light field, e.g. from a camera dolly on a track, it is unlikely for a pixel to be occluded in all of the views (both to the left and to the right of the central view) as illustrated in figure 2. Generally, each pixel is visible in at least one of the other views. Pixels which are occluded in all views are only rarely caused solely by occlusions within the scene – they are most often caused by pixel correspondences being occluded in part of the views and lying outside of the image bounds in the others.

3.1. Cost Function

Ideally, equation (2) is fulfilled in all of the pixels in the image. To express the quality of an estimate u, we first warp the image with the estimated disparity map and call it $\tilde{I}_k$:

$$\tilde{I}_k(x, y, \hat{u}) = I_k(x + \theta_k \hat{u}(x, y), y). \qquad (3)$$

Per equation (2) this warped version should resemble the central view as much as possible, which we evaluate with the Euclidean norm:

$$d(I_C, I_k, \hat{u}) = \sum_{(x,y)\in\Omega} \|I_C(x, y) - \tilde{I}_k(x, y, \hat{u})\|^2, \qquad (4)$$

where Ω is the set of all of pixels in the image plane.

Fig. 4: Examples of neighbourhood transforms for an input neighbourhood (left). In the middle, the census transform results in either -1,0 or 1 depending on whether it has a lower value, a similar value or a higher value than the central pixel. To the right, the normalization transform subtracts the mean from each pixel and divides by the neighbourhood variance.

In [10] the authors note that the ℓ^2 norm can be sensitive to outliers, for example from occlusions, and instead use the ℓ^1 norm. Because of our explicit occlusion handling, we can use the ℓ^2 norm which has the benefit of being differentiable.

As the authors of [10] state, the comparison between the *ground truth* $I_C(x, y)$ and $\tilde{I}_k$ (the current modelled view for the central camera) is susceptible to illumination changes (e.g. due to vignetting or dynamic scenes). They use a structure-texture decomposition in order to circumvent this. We will instead use various image transforms to achieve illumination-invariance, as was discussed in [9] for optical flow. Examples of such transforms include the census transform [7] and the normalization transform [8] (illustrated in figure 4). The latter results in the normalized cross-correlation when using the Euclidean norm in equation (4), as shown in [8].

Transforming the images in T ways, we denote the t^{th} transformed version of input image k by $\mathcal{F}_t I_k$. Then the data error term for camera k and transform $\mathcal{F}_t$ is $d(\mathcal{F}_t I_C, \mathcal{F}_t I_k, \hat{u})$. The complete data fidelity term is:

$$E_d(\hat{u}) = \frac{1}{K} \sum_{t=1}^{T} \frac{\lambda_t}{|\mathcal{N}_t|} \sum_{k=1}^{K} d(\mathcal{F}_t I_C, \mathcal{F}_t \tilde{I}_k, \hat{u}). \tag{5}$$

The divisions by K (the number of cameras) and $|\mathcal{N}_t|$ (the neighbourhood size for the t^{th} transform) imply that the weighting between data fidelity and regularisation (see later) does not change as the number of input images or image transforms increases.

The λ_t are weighting factors for the various transforms, which serves two purposes. First of all, the weighting factors equalize the range of the transforms so that one transform does not have much more influence than another simply because the transformed values are larger by an order of magnitude. Secondly, the weighting factors also allows to trade off the data fidelity against the regularisation introduced later.

As mentioned earlier, we explicitly take occlusions into account. With the current formulation of the cost function, this is done by limiting the summation in equation (4) to the non-occluded pixels. Additionally, in line with the earlier note on a constant weighting between data fidelity and regularisation, the division by K in equation (5) is revised on a pixel-by-pixel basis, now replaced by the number of views in which each pixel is visible. To decide whether a pixel (x, y) from the central view is visible in a given view, we use a z-buffer based warping: (x, y) is visible in the generated view if and only if it is the pixel closest to the camera that is warped to its corresponding pixel. This number will generally be larger than zero, but on image boundaries it may well be some pixels are visible in no other view. In this case, we disregard data fidelity information for this pixel and infer the disparity solely from regularisation.

We adopt the same regularisation term as [8]. Based on the bilateral filtering, it allows depth discontinuities only in those locations where color discontinuities are present. This assumes that the foreground and background have distinct colours. Denoting the bilateral filter coefficient between pixels (x, y) and (p, q) as $b_{(x,y),(p,q)}$, the regularisation term is defined over the neighbourhoods $\mathcal{N}(x, y)$:

$$E_s(\hat{u}) = \sum_{(x,y)\in\Omega} \sum_{(p,q)\in\mathcal{N}(x,y)} b_{(x,y),(p,q)} \|\hat{u}(x, y) - \hat{u}(p, q)\|_1. \tag{6}$$

3.2. Optimisation

The optimisation of the cost function results in the following minimization problem:

$$\hat{u} = \arg\min_u E_d(u) + E_s(u) \tag{7}$$

Our minimization approach for equation (7) is based on [8], adjusted to our cost function. Non-differentiability of the smoothness term is resolved by defining the linear operator K: $Ku((x, y), n) = u(\mathcal{N}_{(x,y)}(n)) - u(x, y)$, where n is a linear index to the regularisation neighbourhood. Using $F(\vec{y}) = \|\vec{b} \cdot \vec{y}\|_1$, the smoothness term becomes $E_s(u) = F(Ku)$. Finally, the problem is reformulated using primal-dual techniques [13]:

$$\hat{u} = \arg\min_u \max_q E_d(u) + \langle Ku, q\rangle - F^*(q). \tag{8}$$

In this expression, $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbb{R}^{|\Omega|\times|\mathcal{N}|}$, and F^* is the convex conjugate of F:

$$\begin{aligned} F^*(q) &= \begin{cases} 0 & \text{if } q \in Q \\ \infty & \text{elsewhere} \end{cases} \\ Q &= \{q \in \mathbb{R}^{|\Omega|\times|\mathcal{N}|} \mid \\ &\qquad \forall \vec{i} \in \Omega, \vec{s} \in \mathcal{N}_{\vec{i}} : \|q(\vec{i}, \vec{s})\|_1 \le b_{\vec{i},\vec{s}}\} \end{aligned} \tag{9}$$

Now we optimise the problem by alternating gradient descent (respectively gradient ascent) between the variables u and q. In order to make the data term easily differentiable, we use the the first-order Taylor approximation in terms of changes to the disparity map in equation (3):

$$\tilde{I}_k(x, y, u + \Delta u) \approx \tilde{I}_k(x, y) + \theta_k \Delta u(x, y) \frac{\partial}{\partial x} \tilde{I}_k(x, y), \tag{10}$$

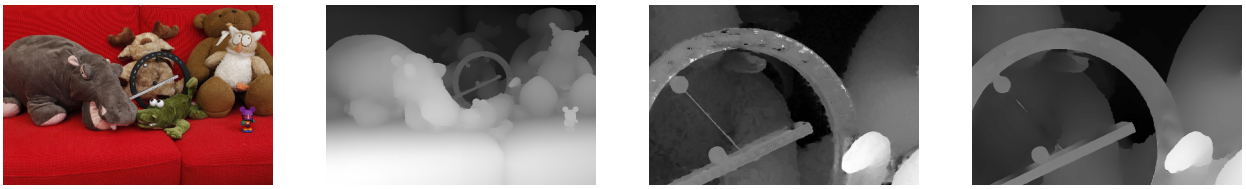

Fig. 5: Depth estimation for the couch dataset from [11] with 10 input frames. From left to right: the input image for the central camera, our depth estimate, a detail of the depth estimate from [11] and our estimate for that same detail.

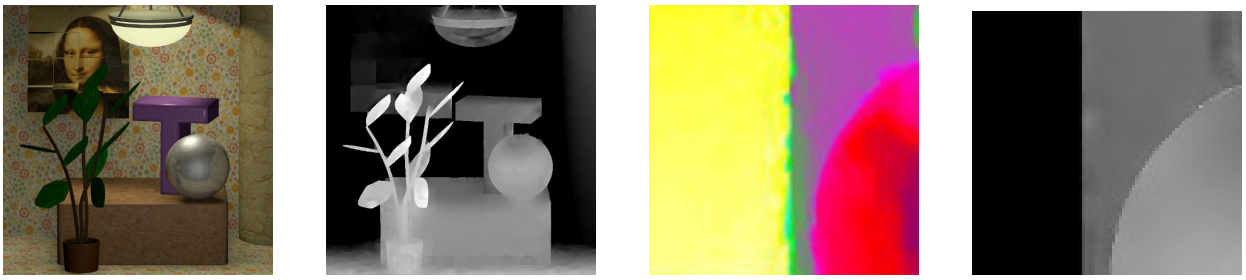

Fig. 6: Depth estimation for the mona dataset from [12]. From left to right: the input image for the central camera, our depth estimate, the depth estimate from [10] for a detail and our estimate for that same detail. Note that we only used the central line in the light field for each pixel, while [10] exploited the entire light field.

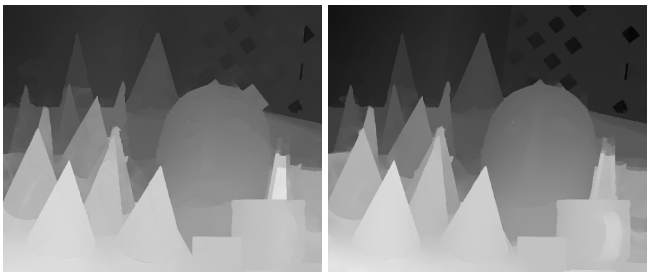

Fig. 7: Illustration for the use of multiple input images: the left image shows estimation using only two images (4 and 5), while the right shows the result using all five. Input images courtesy of the Middlebury 2003 dataset[15].

optimising Δu in each iteration. Because the solution method is based on a first-order Taylor approximation the step Δu should not be too large. In order to enforce this, proximal point terms are added to equation (8) as in [14] (τ and η are tunable step sizes):

$$\frac{1}{2\tau}\|u - u^{(i)}\|^2 - \frac{1}{2\eta}\|q - q^{(i)}\|^2. \qquad (11)$$

4. RESULTS

In figure 7 we illustrate that our estimate from more than two (five in this example) images is more accurate than the estimation based on two images, as one would expect. Secondly, we compare our depth estimate to the results given by [11] and [10] as seen in figures 5 and 6 (best viewed in colour).

The proposed method results in less noisy estimations and our estimate follows the edges better. Remaining problems are structures smaller than their own disparity (the small bar in figure 5) and parts of the image where foreground and background have very similar colours (the painting and the plant in figure 6).

5. CONCLUSION

Here we have presented an approach to depth estimation in multi-camera set-ups inspired by two-view stereo matching. Through explicit occlusion handling we are able to exploit the characteristic of light fields that a pixel is only rarely occluded in *all* of the other views (and then usually just because in some views it lies outside of the image). We use multiple image transforms, complementing the strong suits of one by those of another. When comparing the output to that of existing techniques, we illustrate that our proposed method results in smoother disparity estimates adhering better to image bounds: an important characteristic when using the estimated disparities in view interpolation.

Applying the proposed method to two-dimensional light fields rather than one-dimensional ones is straightforward and will allow even better occlusion handling. Yet, for the proposed method to be applicable in more practical applications, requiring all camera locations to be known accurately should be less stringent: jointly estimating their locations and the disparity may provide the solution. This will be the topic of our future work.

6. REFERENCES

[1] E.H. Adelson and J.Y.A. Wang, "Single lens stereo with a plenoptic camera," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 99–106, Feb 1992.

[2] Marc Levoy and Pat Hanrahan, "Light field rendering," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1996, SIGGRAPH '96, pp. 31–42, ACM.

[3] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof, "Pushing the limits of stereo using variational stereo estimation," in *Intelligent Vehicles Symposium (IV), 2012 IEEE*, June 2012, pp. 401–407.

[4] D. Scharstein and Chris Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[5] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[6] J. Braux-Zin, R. Dupont, and A. Bartoli, "A general dense image matching framework combining direct and feature-based costs," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 185–192.

[7] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer Vision ECCV '94*, Jan-Olof Eklundh, Ed., vol. 801 of *Lecture Notes in Computer Science*, pp. 151–158. Springer Berlin Heidelberg, 1994.

[8] Marius Drulea and Sergiu Nedevschi, "Motion estimation using the correlation transform," *Image Processing, IEEE Transactions on*, vol. 22, no. 8, pp. 3260–3270, 2013.

[9] Li Xu, Jiaya Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012.

[10] Stefan Heber, Rene Ranftl, and Thomas Pock, "Variational shape from light field," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Anders Heyden, Fredrik Kahl, Carl Olsson, Magnus Oskarsson, and Xue-Cheng Tai, Eds., vol. 8081 of *Lecture Notes in Computer Science*, pp. 66–79. Springer Berlin Heidelberg, 2013.

[11] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73:1–73:12, July 2013.

[12] Sven Wanner, Stephan Meister, and Bastian Goldluecke, "Datasets and benchmarks for densely sampled 4d light fields," in *VMV'13*, 2013, pp. 225–226.

[13] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems withapplications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[14] R. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[15] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, June 2003, vol. 1, pp. I–195–I–202 vol.1.