Preprint. Presented at The 1st International Workshop on Live Dense Reconstruction with Moving Cameras, ICCV 2011, Barcelona, Spain.

Real-time Global Prediction for Temporally Stable Stereo

Martin Dobias and Radim Sara
Center for Machine Perception
Department of Cybernetics
Czech Technical University in Prague
{dobiama5,sara}@cmp.felk.cvut.cz

Abstract

We present a method for calculation of disparity maps from stereo sequences. Disparity map from previous frame is first transferred to the new frame using estimated motion of the calibrated stereo rig. The predicted disparities are validated for the new frame and areas where prediction failed are matched with a traditional stereo matching algorithm. This method produces very fast and temporally stable stereo matching suitable for real-time applications even on non-parallel hardware.

1. Intro

Last decade marked an increasing interest of researchers in stereo matching of image sequences. This task comes up mainly in automotive industry, 3D TV technologies and robotics.

Using classical stereo algorithms (designed for standalone image pairs) on stereo sequences is not sufficient for these cases because of several issues. First of all, the resulting disparity maps are not temporally consistent – most methods exhibit unwanted flicker between the frames of the sequence. Second, it is desired to lower the computational complexity to achieve higher processing framerates. Typically this means that algorithms have higher error rate because they search a smaller volume in disparity space or do other simplifications. Finally, additional temporal information such as ego-motion or 3D scene flow may be extracted.

2. Related Work

2.1. Spacetime Stereo

Early work on spacetime stereo [19, 5] proposed extensions of spatial windows used for computation of matching costs to spatiotemporal windows, however they do not perform well with dynamic scenes. Their main advantage is that existing algorithms can be easily adapted to handle temporal dimension. Recent work from Richardt *et al*. [14] use spatiotemporal windows for the temporal variant of their algorithm (with addition of per-frame weights).

Temporally stable stereo proposed in [15] considers image sequences as space-time volumes and the matching cost is based on similarity of spatiotemporal elements called *stequels* – optical flow is not explicitly computed to recover motion.

A different approach is taken in [2] where temporal smoothing is applied as a post-processing of disparity maps using a median filter for each pixel over few adjacent frames. In order to cope with motion they compute optical flow between frames and trace the pixels over time.

Min *et al*. [12] achieve temporal stability by adding a coherence function to the stereo matching cost to lower the matching cost in areas with small changes between frames.

2.2. Stereo and Scene Flow

The concept of scene flow has been introduced in [17] as an extension of optical flow to temporal dimension. Some algorithms are designed to take advantage of joint calculation of disparity maps and disparity (scene) flow. For example, in [9] disparities are computed either using WTA or DP strategy, then the disparity flow is calculated using previous frame. Disparity prediction is done for the next frame and matching costs are updated to ensure temporal smoothness. In a similar fashion [3] describe a stereo algorithm with joint estimation of scene flow based on seed growing. The scene flow is grown from stereo matches from previous frame, then the scene flow is used to predict matching in the next frame.

Conversely, [18, 13] present a framework for computation of scene flow that separates stereo matching from scene flow computation. They argue that it is an advantage since the user is free to choose stereo and optical flow algorithm with best properties.

2.3. Real-time Stereo

Some of the above mentioned methods are capable of real-time or near real-time performance by implementing

the proposed algorithms on GPU. There are further real-time algorithms performing neither temporal smoothing nor any extraction of temporal information. Of these the most related to this work is [7], who employ a technique involving triangulation that helps limit the search space for correspondences. Similar to our algorithm the parallel hardware is not required for real-time operation.

Middlebury Stereo Evaluation page[1] lists more than a dozen of real-time algorithms. Most methods leverage the power of GPU (e.g. [14]), while only a few achieve real-time performance on CPU (e.g. [6]).

3. Algorithm Description

Our initial targets were two-fold: (1) implement a real-time stereo based on an existing algorithm; and (2) explore possibilities how to decrease the amount of disparity search space and achieve temporal stability.

The proposed Real-time Prediction (RTP) algorithm is based on the observation that scenes observed by vehicles or robots are typically mostly static and only few objects in a scene are moving. We could therefore predict how the scene would look like in the next frame (assuming static scene and considering ego-motion) and only validate that the prediction has been successful. Only the areas where the prediction fails shall be matched with traditional stereo algorithm, possibly avoiding a lot of computation.

3.1. Global Prediction and Validation

We propose a technique we call *disparity transfer*. It involves (1) prediction of disparity map from the last frame and (2) validation of predicted disparity values in the new stereo frame. The prediction assumes that stereo rig calibration is known. Matched points in the disparity map represent points in the 3D space. In case the scene is static and the stereo rig moves, the 3D points representing rigid objects should only be transformed using a global rotation and translation. If the rig is not moving then a disparity map identical to the last one should be predicted.

We assume that $\mathtt{P}_1$ and $\mathtt{P}_2$ are camera projection matrices of the stereo rig in one frame and the internal calibration $\mathtt{K}$ is the same for both cameras:

$$\mathtt{P}_1 = \mathtt{K}\left[\mathtt{I} \mid \mathbf{0}\right] \quad \text{and} \quad \mathtt{P}_2 = \mathtt{K}\left[\mathtt{I} \mid \mathbf{b}\right] \tag{1}$$

In the next frame the rig is transformed using a rotation $\mathtt{R}$ and a translation $\mathbf{t}$, inducing new matrices $\mathtt{P}'_1$ and $\mathtt{P}'_2$.

$$\mathtt{P}'_1 = \mathtt{P}_1\mathtt{T} \text{ and } \mathtt{P}'_2 = \mathtt{P}_2\mathtt{T} \text{ where } \mathtt{T} = \begin{bmatrix} \mathtt{R} & \mathbf{t} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix} \tag{2}$$

It is possible to show that if $\mathbf{x_d} = (x, y, w, d)$, where (x, y, w) are homogenous coordinates of a point in the image plane of $\mathtt{P}_1$ and d is disparity assigned to that point, then we can transform the point $\mathbf{x_d}$ to the new frame as $\mathbf{x}'_\mathbf{d} = \mathtt{M}\mathbf{x_d}$ using a homography

$$\mathtt{M} = \begin{bmatrix} \mathtt{K}\mathtt{R}\mathtt{K}^{-1} & \frac{\mathtt{K}\mathbf{t}}{f\|\mathbf{b}\|} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix} \tag{3}$$

where $\mathtt{K}$ is the calibration matrix, f focal length, $\mathbf{b}$ baseline and $\mathtt{R}$, $\mathbf{t}$ are rotation and translation between the frames.

The transformed points $\mathbf{x}'_\mathbf{d}$ in disparity space are then aligned to integer image coordinates in the new disparity map. Finally the predicted (real) disparity values are validated by calculating the similarity in the new frame and assigning either $\lfloor d \rfloor$, $\lceil d \rceil$ or no disparity if the similarity is below a threshold.

In dynamic scenes the validation will fail (i.e. no disparity is assigned) on objects which undergo a motion different from the estimated ego-motion. Such objects will be matched using traditional stereo (Sec. 3.3).

3.2. Estimation of Ego-motion

The prediction requires an estimate of ego-motion between consecutive frames. This is a relatively simple task when a stereo camera is employed and we are aware of previous methods [1, 11, 8]. Here we propose a simple yet roboust algorithm.

The task is to find relative rotation $\mathtt{R}_1$ and translation $\mathbf{t}_1$ of a stereo rig between consecutive frames from stereo image and disparity map from last frame (I_0^L, I_0^R, D_0) and stereo image from current frame (I_1^L, I_1^R). Calibration matrix $\mathtt{K}$ for both cameras and baseline $\mathbf{b}$ are known.

The estimation proceeds as follows:

1. Extract Harris interest points $\{\mathbf{x}_i\}$ from image I_0^L. Use the associated disparity map D_0 to determine disparities $\{d_i\}$ for the points $\{\mathbf{x}_i\}$ and discard the points where the disparity is not known.

2. For each point $\mathbf{x}_i = (u_i, v_i)^\mathsf{T}$ and its disparity d_i compute corresponding 3D point $\mathbf{X} = (x, y, z, w)^\mathsf{T}$ by using camera calibration $\mathtt{K}$ and baseline $\mathbf{b}$.

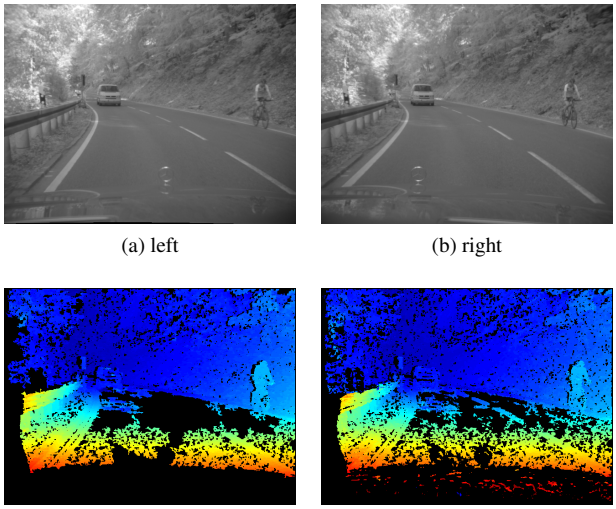
3. Predict position of points $\{\hat{\mathbf{x}}_i\}$ in the current frame I_1^L from points $\{\mathbf{X}_i\}$ using ego-motion estimate from the last frame $\mathtt{R}_0$, $\mathbf{t}_0$ (initially $\mathtt{R}_0 = \mathtt{I}$, $\mathbf{t}_0 = \mathbf{0}$).

4. Run Lucas-Kanade tracker for corresponding pairs $\{\mathbf{x}_i\}$ in I_0^L and $\{\hat{\mathbf{x}}_i\}$ in I_1^L to get subpixel estimation $\{\tilde{\mathbf{x}}_i\}$ in new frame I_1^L of features $\{\mathbf{x}_i\}$ from the last frame.

5. Find relative camera rotation $\mathtt{R}_1$ and translation $\mathbf{t}_1$ using P3P algorithm in the RANSAC scheme by minimizing the total solution error

$$e = \sum_i \min\left\{e_i, e_{\text{thr}}\right\} \text{ where } e_i = ||f_{\mathtt{R}_1,\mathbf{t}_1}(\mathbf{X}_i) - \tilde{\mathbf{x}}_i||$$

[1] http://vision.middlebury.edu/stereo/eval/



(a) left (b) right

(c) RTG (d) RTP

Figure 1: Source images and output disparity maps

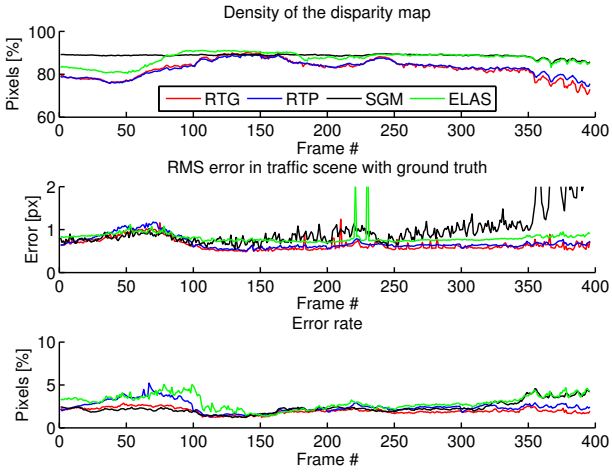


Figure 2: Comparison with other algorithms

3.3. Real-time Growing (RTG) Algorithm

For bootstrapping and matching of areas where prediction failed we have produced a very efficient implementation of seed-growing method that is capable of real-time performance. In principle, other fast stereo algorithms could be used, but we have not tested them.

Our implementation is based on the Baseline Method from [4]. The algorithm starts with a sparse set of robustly matched correspondences called *stereo seeds*. The seeds are first inserted into a queue, then in each step their neighborhood is analyzed. The best matches from neighborhood are inserted into the queue for further processing and put into the output disparity map. The matching is greedy (once a disparity has been assigned, it will not be changed in future) and terminates when the newly encountered correspondences stay below the minimal similarity threshold. The threshold effectively determines the ratio between matching density and error rate.

The original paper proposes also a more robust variant (called GCS). That approach produces fewer errors in the exchange of density of the disparity map. We have chosen the Baseline Method in favor of the robust variant because of speed. The baseline method is several times faster since it searches a smaller part of the disparity space and omits a final optimization step. We plan to develop a combined method in future.

The advantage of the seed growing is that by controling the seed queue one can indirectly limit the search space. This has a similar effect as the adaptive search range in [7], although the means to achieve it are completely different.

4. Evaluation

Figure 1 shows a sample of disparity maps produced from real-world sequences [16]. RTP maintains a better temporal stability compared to RTG as it greatly reduces the flicker. The improved stability comes from the fact that disparity maps from RTP are based on the ones from previous frames. The differences can be seen in the online resource[2]. RTP generally produces more errors than RTG because sometimes wrong disparity values are validated.

The processing speed on Intel Core i5 processor and VGA resolution is about 10-15 fps for RTG and 15-20 fps for RTP — it depends on the density of disparity maps. Our implementation uses two threads.

Next we compare the performance of RTP, RTG to ELAS [7] and SGM algorithms [10] in three aspects: disparity map density, RMS error of valid disparities and error rate ($|d - d_{GT}| > 2$). The tests were run on a synthesized sequence with ground truth [16] – Sequence 2. The results are shown on per-frame basis in Figure 2. The RMS error is low for both RTP and RTG. The end of the sequence is more challenging: the density of RTP/RTG decreases keeping stable error, while ELAS and SGM increase the error rate and keep the density.

5. Demo Content

The work will be presented as a live demo with Bumblebee2 stereo camera. We will demonstrate real-time 3D visualization of the scene as a point cloud and automatic registration of point clouds from estimated ego-motion. Figure 3 shows an example of a live demonstration.

[2] http://cmp.felk.cvut.cz/~dobiama5/stereort/

Figure 3: 3D scanning demonstration

Acknowledgements

This work was supported by the Czech Ministry of Education under Project MEB111006.

References

[1] H. Badino. A robust approach for ego-motion estimation using a mobile stereo platform. In *Proceedings of the 1st international conference on Complex motion*, IWCM'04, pages 198–208, 2007.

[2] M. Bleyer and M. Gelautz. Temporally consistent disparity maps from uncalibrated stereo videos. In *ISPA*, pages 383–387, sept. 2009.

[3] J. Cech, J. Sanchez-Riera, and R. Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR*, pages 3129–3136, june 2011.

[4] J. Cech and R. Sara. Efficient sampling of disparity space for fast and accurate matching. In *CVPR BenCOS*, june 2007.

[5] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: a unifying framework for depth from triangulation. *IEEE PAMI*, 27(2):296–302, feb. 2005.

[6] S. Gehrig and C. Rabe. Real-time semi-global matching on the CPU. In *CVPR Workshops*, pages 85–92, june 2010.

[7] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In R. Kimmel, R. Klette, and A. Sugimoto, editors, *ACCV*, volume 6492 of *Lecture Notes in Computer Science*, pages 25–38. Springer, 2010.

[8] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV)*, pages 963–968, june 2011.

[9] M. Gong. Real-time joint disparity and disparity flow estimation on programmable graphics hardware. *CVIU*, 113:90–100, January 2009.

[10] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30(2):328–341, February 2008.

[11] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *IROS*, pages 3946–3952, sept. 2008.

[12] D. Min, S. Yea, and A. Vetro. Temporally consistent stereo matching using coherence function. In *3DTV-CON*, pages 1–4, june 2010.

[13] C. Rabe, T. Müller, A. Wedel, and U. Franke. Dense, robust, and accurate motion field estimation from stereo image sequences in real-time. In *ECCV*, pages 582–595, 2010.

[14] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *ECCV*, pages 510–523, 2010.

[15] M. Sizintsev and R. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In *CVPR*, pages 493–500, june 2009.

[16] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behavior on synthetic and real-world stereo sequences. In *23rd International Conference of Image and Vision Computing New Zealand (IVCNZ '08)*, pages 1–6, 2008.

[17] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV*, volume 2, pages 722–729 vol.2, 1999.

[18] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. Stereoscopic scene flow computation for 3d motion understanding. *IJCV*, 95:29–51, 2011.

[19] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: shape recovery for dynamic scenes. In *CVPR*, volume 2, pages II – 367–74 vol.2, june 2003.