https://doi.org/10.2352/ISSN.2470-1173.2017.19.AVM-013

Accelerated Stereo Matching for Autonomous Vehicles using An Upright Pinhole Camera Model

Chen Chen[1], *Jiangbo Lu*[2], *Do-Kyoung Kwon*[3], *Darnell Moore*[3] *and Minh N. Do*[1]
[1] *University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA*
[2] *Advanced Digital Sciences Center, Singapore 138632*
[3] *Texas Instruments, Dallas, TX 75243 USA*

Abstract

In this paper, we propose a new method for accelerating stereo matching in autonomous vehicles using an upright pinhole camera model. It is motivated by that stereo videos are more restricted when the camera is fixed on the vehicles driving on the road. Assuming that the imaging plane is perpendicular to the road and the road is generally flat, we can derive the current disparity based on the previous one and the flow. The prediction is very efficient that only requires two multiplications per pixel. In practice, this model may not hold strictly but we still can use it for disparity initialization. Results on real datasets demonstrate the our method reduces the disparity search range from 128 to 61 with only slightly accuracy decreasing.

Introduction

Autonomous vehicles have been an active area of research in the recent years, which are able to navigate without human interaction. In order to make a high-level driving decision, it is important to accurately measure the detailed 3D representation of the environment outside the vehicle in real-time. The surroundings can be detected using a variety of techniques such as radar, lidar, GPS, odometry, *etc*. Instead of installing expensive hardware such as LIDAR sensors on the vehicles, computer vision based techniques using stereo matching could be a great substitution, in which the depth of the scene can be estimated accurately using consumer cameras. The key problem of stereo vision is to find the corresponding pixels from different viewpoints, and then the depth can be estimated based on these matching points.

Existing stereo algorithms can often be classified into two groups: local or global algorithms [1]. Global algorithms often optimize a global cost function to solve the disparity at all pixels. Typical methods include dynamic programming and graph cut. In local algorithms, the computation of the disparity only depends on the pixels within a local patch. The global methods usually produce better results [1] but are generally too slow to be used in autonomous vehicles. Summaries of existing stereo correspondence algorithms can be found in [2, 1]. The state-of-the-art stereo methods achieve promising results on the benchmark dataset like KITTI [3], but they are still computationally expensive or require powerful devices such as GPU that can be difficult to be used in real-time in autonomous vehicles. One of the most costly steps by these methods is to compute the matching cost of each pixel over a large range of possible disparities.

In this paper, we are interested in stereo matching for *videos* taken from cameras mounted on ordinary vehicles. While many of previous works considering temporal information to improve the results of matching, we believe that stereo can be accelerated by using the temporal consistency between consecutive frames. We observe that the cameras on consumer vehicles are fixed and approximately upright to the ground if the road is generally flat. Motivated by this observation, we propose an efficient method to accelerate video stereo matching using such a constraint in the pinhole camera model. Based on this model, the disparity of the current frame can be directly derived from the disparity of the previous frame and the optical flow between these two frames. In practice, the optical flow can be estimated by FPGA-based real-time systems [4, 5] and the disparity prediction can be performed very efficiently with only two multiplications. When the optical flow and disparity of the previous frame are accurate, our model can faithfully predict the correct solution excepted occluded pixels. On real datasets like KITTI, we may have errors in both the estimated optical flow and the disparity of previous frame. In this case we can use such predicted disparity values as initialization to accelerate the state-of-the-art stereo matching algorithms. We evaluate this strategy in the fast cost-volume filtering based framework [6, 7] on the KITTI dataset. The results demonstrate our method can reduce the disparity search range from 128 to 61 with only slightly accuracy decreasing.

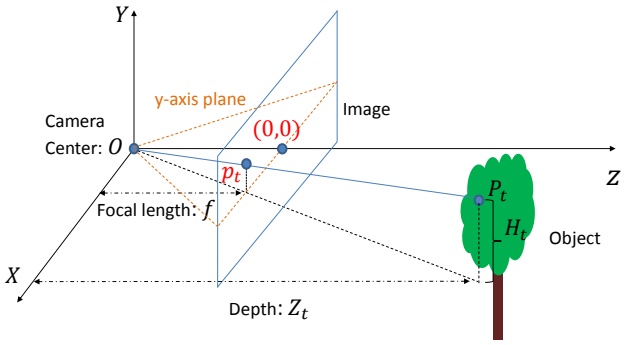
Related Work on Stereo Video

There are many existing works developed to improve stereo matching using temporal information. Zhang *et al.* [8] proposed to extend a local spatial window to the spatiotemporal window to compute the matching cost. This method does not work well when the scene is dynamic. Jiang *et al.* [9] proposed to predict the current disparity map based on the previous disparity. They divided the scene into moving parts and static background, where the disparity of static background is predicted by a rigid camera motion model. This method is computationally expensive due to the segmentation of motions in the scene. Jaafari *et al.* [10] used dynamic programming to match edge points of stereo images, where the disparity search range can be reduced if an edge point can be associated with the one in the previous frame. However, the edge association is very challenging when the scene is dynamic and complex. Dobias *et al.* [11] proposed to transfer the disparity of the previous frame to the current frame by estimating the motion of a calibrated stereo rig. The transformation is assumed to be linear and results in large errors on real data. Our method is also for disparity prediction. Compared with existing methods, our proposed method does not have any assumption about the motion of cameras as well as the moving objects in the scene. It is more general and can be applied to complex and dy-

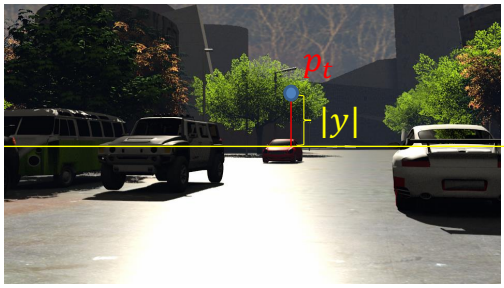
namic scenes without time consuming processes such as motion detection, edge association.

Method

For classical stereo video matching, the motion of cameras can be arbitrary. That is why many existing methods have to model the scene in 3D, in order to find the correspondence in two adjacent frames. For autonomous vehicles, the stereo cameras can always be fixed on the car and are perpendicular to the ground. We can use such addition information for disparity prediction.



(a)



(b)

Figure 1. *(a) The upright pinhole camera model. (b) The corresponding point and distance to y-axis on the image.*

Fig. 1 illustrates the upright pinhole camera model. When the imaging plane is perpendicular to the ground, the y-axis plane in the image is the horizontal vanishing line. Let O be the camera center and the focal length is f. For a point P in the scene at time t, we denote the point as P_t and the corresponding point in the image is $p_t = (x,y)^T$. Let $(0,0)$ represent the principle point. The distance of P_t to the y-axis plane is H_t and this distance in the image is $|y|$. The depth of P_t is Z_t. With these notations, we can have:

$$H_t/|y| = Z_t/f. \tag{1}$$

For the next frame, we assume the ground is flat between the two frames, so that the cameras' height is the same and the y-axis plane keeps the same. This means that $H_{t+1} = H_t$ no matter what direction the cameras move. Let $p_{t+1} = (x+\Delta x, y+\Delta y)^T$ be the corresponding point in the next frame and $(\Delta x, \Delta y)$ denotes the flow. Similar as Eq. (1), we also have:

$$H_{t+1}/|y+\Delta y| = Z_{t+1}/f. \tag{2}$$

Combining the above two equations, we can obtain:

$$|y|Z_t = |y+\Delta y|Z_{t+1} \tag{3}$$

Suppose that the baseline of the cameras is B. $\mathbf{d}_t(x,y)$ represents the disparity map at point (x,y) of frame t. A similar notation is used for frame $t+1$. We have:

$$\mathbf{d}_t(x,y) = Bf/Z_t \tag{4}$$

$$\mathbf{d}_{t+1}(x+\Delta x, y+\Delta y) = Bf/Z_{t+1} \tag{5}$$

Combining the above equations, we obtain:

$$\mathbf{d}_{t+1}(x+\Delta x, y+\Delta y) = |y+\Delta y|\mathbf{d}_t(x,y)/|y| \tag{6}$$

Or we can use the backward flow:

$$\mathbf{d}_{t+1}(x,y) = \mathbf{d}_t(x-\Delta x, y-\Delta y)|y|/|y-\Delta y| \tag{7}$$

which indicates that we can predict $\mathbf{d}_{t+1}$ at pixel (x,y) based on the results on the previous frame. Given the flow between the two frames, the prediction can be made with only two multiplications and two additions for each pixel.

In practice, the optical flow can be estimated by FPGA-based real-time systems [4, 5] and the disparity prediction can be made very efficiently. Since our prediction is based on the optical flow and the disparity of the previous frame, it may propagate the errors from the previous frame. Therefore, our prediction can be used as the disparity initialization instead of the final result. Now we introduce how our method can be used in existing stereo matching frameworks.

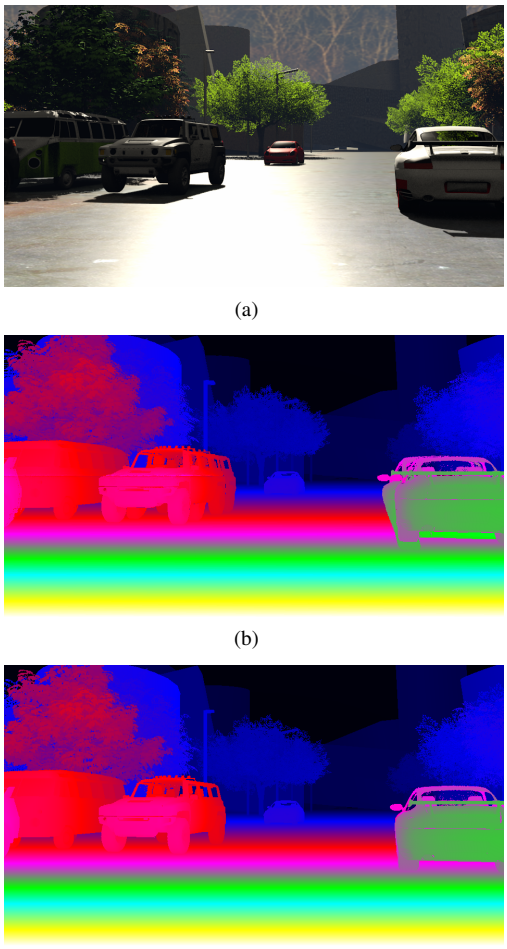
Stereo algorithms generally perform (subsets of) the following four steps [1]: (1) matching cost computation; (2) cost aggregation; (3) disparity optimization; (4) disparity refinement. We follow the scheme of the multi-block matching (MBM) algorithm [7]. In the first step, normalized cross correlation is used to compute the matching cost volume. After that, box filtering with multiple block shapes is used for cost aggregation, which is followed by a winner-takes-all process. Finally, disparity errors are removed by consistency check and refined by slanted plane smoothing [12].

For the first step, an accurate choice of the disparity search range is crucial. Without exploiting any prior information from the temporal frames, this method [7] has to search a large range to ensure the match, which is computationally expensive and may degrade the quality of the disparity map. With our prediction from temporal information, the search range can be significantly decreased. More clearly, supposing that our predicted disparity is $\tilde{\mathbf{d}}_{t+1}$ using Eq. (7), our search range can be constrained to $[\,\tilde{\mathbf{d}}_{t+1} - N, \tilde{\mathbf{d}}_{t+1} + N]$ while the original search range is $[0, M]$ without the prediction. Here M, N are scalars and $M > 2N+1$. We will show the results in the next section on synthetic and real data.

Results

First, we evaluate our upright camera model on synthetic stereo data [13], which provides ground-truth disparity and optical flow for every frame. Given the ground-truth disparity of the previous frame and the optical flow, our prediction is based on Eq. (7), which does not require the actual matching with the right

image. If our assumption holds, our method should nicely predict the result since there is no error in the input. Fig. 2 shows our prediction on one example of the driving stereo set. The error is measured by the percentage of outliers for a 3-pixels tolerance. The result demonstrates our model is very accurate that we only have small errors for the pixels occluded by the motion of cars. It validates the effectiveness of proposed model.



(a)

(b)

(c)

Figure 2. *(a) The RGB image. (b) The predicted disparity d_{t+1} using the ground-truth disparity of the previous frame and optical flow. The error is 7.3% with a 3-pixels tolerance. (c) The ground truth disparity d_{t+1}.*

The proposed method is then validated on the driving dataset KITTI [3]. We use two flow algorithms SIFT flow [14] and EpicFlow [15] for the flow estimation. Generally, EpicFlow is one of the state-of-the-art flow estimation methods and outperforms SIFT flow in most cases. From those two methods, we could observe how the accuracy of flow could influence our disparity prediction. The baseline method for comparisons is the same method without temporal information and flow.

One may notice that the denominator of Eq. (7) is the distance of a pixel to the y-axis plane. When the pixel is very close to the y-axis plane, the error caused by optical flow or disparity of the previous frame will be amplified a lot. Instead of directly applying the model, we interpolate the disparity for a few rows (*e.g.*

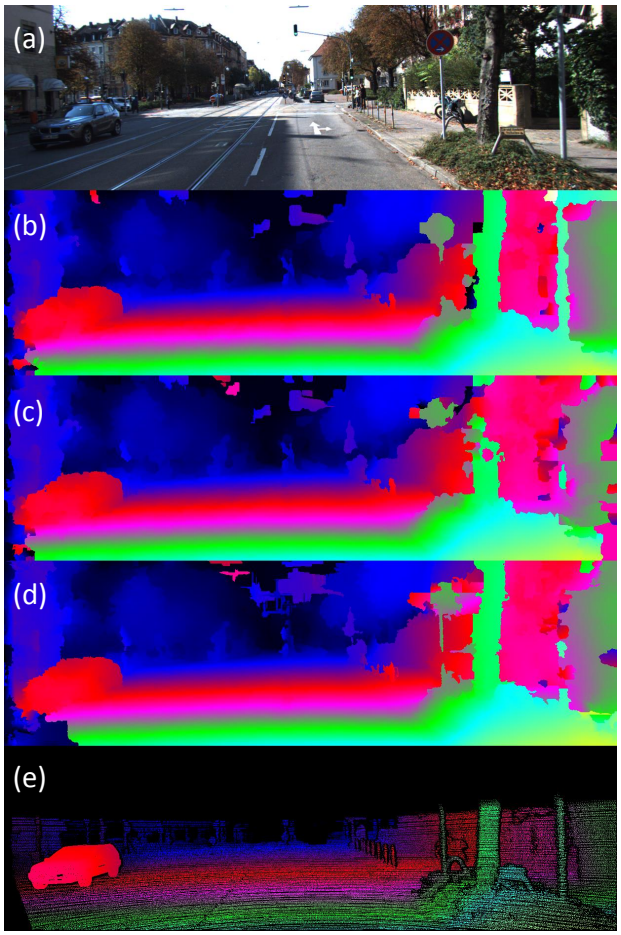


Figure 3. *(a) The RGB image. (b) The disparity map using the baseline method. The error is 3.94%. (c) The disparity map using SIFT flow in our prediction. The error is 9.77%. (d) The disparity map using EpicFlow in our prediction. The error is 5.11%. (e) The ground-truth disparity.*

3) closest to the center. As our method can reduce the disparity search range, the baseline method searches the full range [0,128] while this range is reduced to [-30,30] given our prediction as the initialization.

Our method does not require any training data and is parameter-free. Therefore, we use the training set of the KITTI dataset for result evaluation. It contains 200 groups of driving scene frames. Each group has four RGB images: left and right images of two adjacent frames. One example is shown in Fig. 3. It demonstrates that the baseline method with [0,128] has the lowest error. Our result with EpicFlow is much better than the one with SIFT flow, in terms of both visual appearance and the error rate, which shows the importance of the flow estimation in our prediction.

The quantitative results on the whole training set are shown in Table 1. If the optical flow can be estimated properly, our method can reduce the disparity search range by half with only slightly accuracy decrease (1.85%). With a better flow method, it is very possible that our result can be further improved.

Table 1: the average error on KITTI training dataset.

Method	Baseline	Proposed (SIFT flow)	Proposed (EpicFlow)
Error	7.04%	10.71%	8.89%

Disucssion and Conclusion

In this paper, we have proposed a new stereo matching method for autonomous vehicles. It is motivated by that the cameras on autonomous vehicles can always be fixed and perpendicular to the ground. Therefore the structure of the captured videos is more restricted and can be used to infer disparity based on temporal information. In practice, the proposed model may not hold strictly for every pixel, but our method can be used as an initialization to reduce the disparity search range in existing methods. Results on the KITTI dataset demonstrate that our method can reduce the disparity search range from 128 to 61 with only slight accuracy decreasing.

While the results are promissing, our work is still an initial study in this direction and there are some limitations for the current work. First, although the proposed method reduces the disparity search range by half, it may not accelerate the whole stereo matching by two times since there are many post-processing steps which are not accelerated. Second, we use SIFT flow or EpicFlow to estimate the optical flow instead of the FPGA based flow methods. The advantage of our method on real hardware platform needs to be further investigated, which is planned as future work. Finally, our predition is used in the MBM framework [7] that computes the dense matching cost. With some sparse matched points [16, 17], some recent methods can interpolate the whole disparity map efficiently and accurately [18]. Another future direction would be to combine our prediction in this framework.

Acknowledgment

The authors would like to thank Benjamin Chidester for helpful discussions. This project was supported by Texas Instruments Inc.

References

[1] D. Scharstein and R. 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.

[2] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[3] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] J. Díaz, E. Ros, F. Pelayo, E. M. Ortigosa, and S. Mota, "FPGA-based real-time optical-flow system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 274–279, 2006.

[5] G. K. Gultekin and A. Saranli, "An FPGA based high performance optical flow hardware design for computer vision applications," *Microprocessors and Microsystems*, vol. 37, no. 3, pp. 270–286, 2013.

[6] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[7] N. Einecke and J. Eggert, "A multi-block-matching approach for stereo," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 585–592.

[8] L. Zhang, B. Curless, and S. M. Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, 2003, pp. II–367.

[9] J. Jiang, J. Cheng, B. Chen, and X. Wu, "Disparity prediction between adjacent frames for dynamic scenes," *Neurocomputing*, vol. 142, pp. 335–342, 2014.

[10] I. El Jaafari, M. El Ansari, L. Koutti, A. Mazoul, and A. Ellahyani, "Fast spatio-temporal stereo matching for advanced driver assistance systems," *Neurocomputing*, vol. 194, pp. 24–33, 2016.

[11] M. Dobias and R. Sara, "Real-time global prediction for temporally stable stereo," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011, pp. 704–707.

[12] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *European Conference on Computer Vision*. Springer, 2014, pp. 756–771.

[13] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[15] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow," in *Computer Vision and Pattern Recognition*, 2015.

[16] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1706–1721, 2014.

[17] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1385–1392.

[18] Y. Li, D. Min, M. N. Do, and J. Lu, "Fast guided global interpolation for depth and motion," in *European Conference on Computer Vision*, 2016, pp. 717–733.