# Fast Binocular Depth Inference via Bidirectional Motion Based Interpolation

Wenjing Geng, Yang Yang, Ran Ju, Tongwei Ren, Gangshan Wu*
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, China
{jenngeng, charlie.yang.nju}@gmail.com, juran@smail.nju.edu.cn, {rentw, gswu}@nju.edu.cn

## ABSTRACT

Depth information provides fundamental supports to multimedia applications for both images and videos. Depth acquisition for stereo images has drawn much attention while few approaches are proposed for stereo videos. Conducting stereo matching frame-by-frame is time consuming and the result is temporally inconsistent. As a matter of fact, the redundancy shared by frame sequences may cause extra computational cost. Inspired by rapidly acquiring stereo video depth for some specific applications, we propose a novel bidirectional motion-based interpolation framework, which avoids frame-by-frame matching through making use of the motion estimation and the redundancy between frames. Firstly, comparable accurate depth maps are generated for self-adaptive selected frames via stereo matching. Then rough depth sequences inbetween are calculated using bidirectional motion-based interpolation. To improve the depth accuracy for non-selected frames, we propose a refinement approach to handle cracks and holes. The evaluation on both computer rendered and real world captured datasets show that our approach is competent for fast and accurate binocular video depth acquisition.

## Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: 3D/stereo scene analysis; I.4.8 [**Scene Analysis**]: Depth cues

## General Terms

Algorithm, Experimentation, Performance

## Keywords

Depth Acquisition, Stereo Videos, Stereo Matching, Motion Estimation, Motion-based Interpolation

## 1. INTRODUCTION

Although stereo matching for binocular images is widely studied in the field of multimedia applications and computer vision, acquiring depth sequences from binocular videos has seldom been discussed. As a matter of fact, depth information can serve as a significant cue in many applications, such as 3D reconstruction [11], stereo video coding [9] and scene understanding [4]. Nowadays, stereo media increase rapidly, which leads to an emergent demand for further processing on depth data.

The depth acquisition on stereo videos has two major differences from that on stereo images which makes it a challenging problem to be solved:

- Adjacent frames for a natural video are highly correlated which makes applying stereo methods frame-by-frame time consuming. How to employ the content redundancy and consistency between frames for reducing computation time turns out to be a problem.
- Frames in one shot have implicit continuity which guarantees consistent changes when playing videos and should be preserved in depth sequences for further processing, e.g. stereo video coding and object tracking.

Due to the difficulty of fast computing reliable and consistent depth in long sequences, few studies have well solved the above problems. Wedel et al. [10] computed depth sequentially assuming that the depth in previous frames is known, which obviously contains computing redundancy. Valgaerts et al. [8] estimated motion field in the four-frame configuration which makes temporal consistency still be a problem. In order to acquire temporal consistency preservation depth maps, Hung et al. [5] proposed a depth and image scene flow estimation method using motion-depth temporal consistency constraint. Although this method can generate very accurate and smooth depth videos, the time complexity is definitely high because of a bunch of constraints and optimization, which makes it inappropriate for applications needed to be finished on time.

Differed from these methods, we try to leverage the inter-frame redundancy to reduce computational time and preserve consistency. As a natural video comprises of several screenshots and the frames inside a shot are highly correlated, we propose to perform stereo matching only for selected frames and infer depth maps for the remaining using a bidirectional motion-based interpolation. The inter-frame interpolation is much faster than global optimization methods. Compared to the previous work [10, 8, 5], our approach is more efficient due to the employment of inter-frame redundancy. Furthermore, owing to the bidirectional motion estimation, the proposed approach works well in preserving continuity between frames.

Figure 1 shows an overview of our approach. Considering the redundancy between frames, we first adaptively select a few frames based on PSNR threshold to ensure that error propagation maintains within a proper range. The complete depth computation is only applied to selected

*ICIMCS'14,* July 10–12, 2014, Xiamen, Fujian, China.
Copyright 2014 ACM 978-1-4503-2810-4/14/07 ...$15.00.

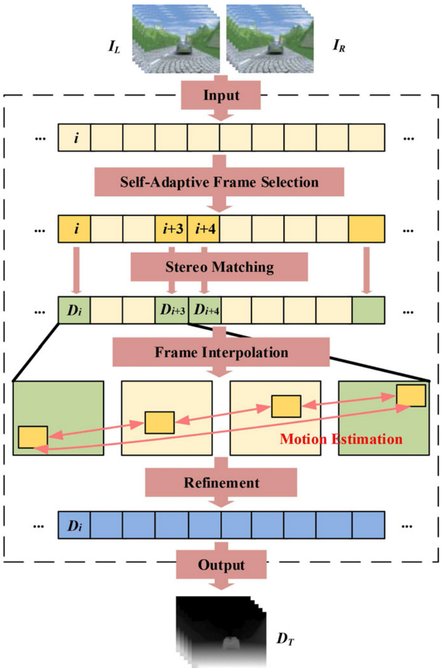


**Figure 1: The framework of the proposed approach**

frames. For the remaining frames, we calculate the depth maps by a bidirectional motion-based interpolation, which shows both promising efficiency and accuracy for highly similar frames. Finally, we apply a hole-filling to the depth maps for refinement.

In brief, our approach contributes in the following aspects. A fast binocular depth inference framework for applications of time requirement has proposed. First, a self-adaptive strategy for selecting frames is employed to control the interpolation error propagation within a proper range. Then an interpolation method based on bidirectional motion estimation is presented to guarantee the consistency and accuracy of the adjacent frames between selected frames. Besides, our method is comparable to the global optimization methods in accuracy while being much faster.

The remaining of the paper is organized as follows. In Section 2 we give a detailed description of the proposed approach. Then the experiments and discussion are shown in Section 3. Finally, we conclude the paper in Section 4.

# 2. DEPTH INFERENCE BY BIDIRECTIONAL MOTION-BASED INTERPOLATION

Given a stereo video, we aim to generate temporal consistent depth maps in time. As shown in Figure 1, the framework can briefly partition into self-adaptive frame selection, stereo matching, and motion-based interpolation.

## 2.1 Frame Selection

The contents between sequential frames are highly correlated in natural videos. In video coding society the property is widely adopted to improve coding efficiency. We follow a similar way to eliminate the redundant computation for stereo video depth acquisition. On one hand, not every pair frame need to be matched because of the temporal redundancy. On the other hand, error propagation caused by motion-based interpolation should be controlled within a proper scope. We manage to select several frames and leverage the information loss. And the depth values for the remaining can be inferred from those selected frames. In our work, we introduce a self-adaptive algorithm based on peak signal-to-noise ratio (PSNR). The PSNR measure is commonly used to evaluate the quality of reconstruction of loss compression codecs between two images. With the help of quantitative measurement to inter-frame reconstruction, we apply PSNR to predict the difference between neighboring frames. The rule for selecting frames is defined as:

$$decision(i) = \begin{cases} 1, & \|P_i^b - P_i^a\| > \tau \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\tau$ is a threshold determined based on the quality of stereo videos, $P_i^b$ is the PSNR difference between $i$th frame and $(i-1)$th frame while $P_i^a$ is the PSNR difference between $i$th frame and $(i+1)$th frame. Each comparison involves three frames which ensures that scene change can also be detected and handled. And PSNR is defined as:

$$PSNR = 10 \times \log_{10}(\frac{(2^n-1)^2}{MSE}) \tag{2}$$

$$MSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}(f(x,y)-g(x,y))}{M \times N} \tag{3}$$

where $n$ represents using $n$ bits per sample, $f(x, y)$ and $g(x, y)$ are grayscale values of adjacent frames and $M \times N$ is the size of each frame. In this paper, we use 8 bits.

A relative high PSNR difference indicates that the current frame has a sharp change between its previous and next frames, which means that the loss would be high if the current frame was used to reconstruct its neighborhoods. Hence this frame is regarded as a selected frame. Depth sequences between selected frames are highly similar and thus can be inferred by motion-based interpolation.

Next we calculate the depth maps on selected frames. For depth acquisition, we adopt Sun's [7] optical flow method for its accuracy and robustness, which achieved a good rank in the Middlebury optical flow Benchmark [1]. Considering that in calibrated stereo videos the flow (disparity) only occurs in horizontal direction, we add a constraint to Sun's model to eliminate the vertical displacement.

## 2.2 Depth Interpolation

For the content variation between selected frames is very small, it is possible to infer accurate depth by inter-frame interpolation. To save computational time and keep the depth maps temporal continuous, we utilize a motion-based linear interpolation, which is defined as:

$$D_K(x+\alpha u, y+\alpha v) = \alpha D_m(x,y) + (1-\alpha)D_n(x+u, y+v) \tag{4}$$

$$\alpha = \frac{k-m}{n-m} \tag{5}$$

where $D_m$ and $D_n$ indicate the depth of selected frames and $D_k$ is a depth frame inbetween. $(u, v)$ is the horizontal and vertical motion vector estimated between frames $m$ and $n$. To refine the flow vector, a bidirectional motion filed is calculated, as well as depth interpolation.

It is worth noting that inter-frame motion is much smaller than left-right view motion (disparity) for the following reasons. First, the frames between selected frames are highly similar. Second, for static shots most of the frame contents like background keep unchanged. Besides, in most cases the objects move slowly, which leads to a very small flow field between neighboring frames. In comparison, the left-right view motion depends only on the baseline distance and scene depth range. This leads to relative large disparities due to scene depth variation even for static shots. According to the above analysis, we perform accurate but time consuming global optimization for left-right view disparity estimation, and utilize coarse but fast motion estimation for sequential adjacent frames. In this work we apply Liu's implementation for its efficiency [6]. The core of the algorithm is based on [2, 3]. By using the successive over-relaxation (SOR), the code runs much faster and the accuracy goes near to the other time consuming methods.

In fact, the disparity and motion estimation method is not limited to the above. Any stereo matching and motion estimation methods satisfying accuracy and efficiency requirements as we mentioned can be applied to our framework.

## 2.3 Depth Refinement

Inter-frame interpolation may generate cracks and holes in depth maps because of inevitable vanishment and occlusion. An example is shown in the fifth column of Figure 2. To overcome the problem, we apply a linear interpolation to fill the cracks and holes based on small changes between selected frames. First, we detect the unassigned pixels by Eq.(6), where $d_i$ is the disparity of pixel $i$ and $j$ is the surrounding pixels within a fixed window. $\epsilon$ is the experiential threshold to control the influence of the neighborhood. $I(x)$ determines whether the pixel has disparity or not. Then we fill the unassigned pixels ($assignment(i) = 0$) for the current frame $k$ by Eq.(8), where $m$ is the previous selected frame and $n$ is the next selected frame.

$$assignment(i) = \begin{cases} 0 & d_i = 0 \text{ and } \sum\limits_{j \in N(i)} I(d_j) < \epsilon \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

$$I(x) = \begin{cases} 1 & x = 0 \\ 0 & x > 0 \end{cases} \tag{7}$$

$$D_k(x,y) = \beta D_m(x,y) + (1-\beta) D_n(x,y) \tag{8}$$

$$\beta = \frac{k-m}{n-m} \tag{9}$$

# 3. EXPERIMENTS

## 3.1 Dataset and Experimental Settings

To comprehensively evaluate and compare our proposed method with the others, we selected two kinds of datasets, a synthesized dataset[1] including TrafficScene1 and TrafficScene2, and Karlsruhe Dataset[2]. The former contains synthesized (gray-level and color) sequences with ground truth for stereo and motion analysis rendered by computer and the resolution is 480×640 of 100 and 396 sequences. And the latter contains high-quality stereo sequences captured in real world and the resolution is 1344×372 of 112 sequences. For the TrafficScene1, the selection threshold $\tau$ is set to 0.03 and 0.05 for TrafficScene2. While $\tau$ is set to 0.2 for the Karlsruhe dataset. It is notable that very few methods in the literature reported error statistics on video sequences based on [5]. For now, only [5] is the state-of-the-art with leveraging long-range temporal information. That is the reason why only this method is used for comparison.

**Table 1: Comparison of running time (in minutes)**

| | Total | Per frame |
|---|---|---|
| Hung et al. (single thread) | 2772 | 28 |
| Hung et al. (multi thread) | 138 | 1.4 |
| Ours (single thread) | 51.606 | 0.516 |

**Table 2: Running time (in minutes) of our approach**

| | OFBSM | MBI | Ref. | Total | Per Frame |
|---|---|---|---|---|---|
| Scene1 | 44.437 | 7.126 | 0.043 | 51.606 | 0.516 |
| Scene2 | 294.025 | 16.675 | 0.099 | 310.799 | 0.785 |
| Scene3 | 82.434 | 12.170 | 0.061 | 94.665 | 0.845 |

*(note: OFBSM is optical flow based stereo matching, MBI is motion based interpolation, and Ref. is short for refinement)*

In order to evaluate the performance of the proposed algorithm quantitatively, the mean absolute error (MAE), the same as in [5], is employed to measure the errors between the inferred depth $\hat{D}$ and the ground truth disparity $D$. Let $\Omega$ be the pixels in an image, the MAE is calculated as follows:

$$MAE = \frac{1}{|\Omega|} \sum_{\Omega} |\hat{D} - D| \tag{10}$$

The experiments are implemented in Matlab on a machine with a 3.4GHZ Intel i7-4770 CPU and 16GB memory.

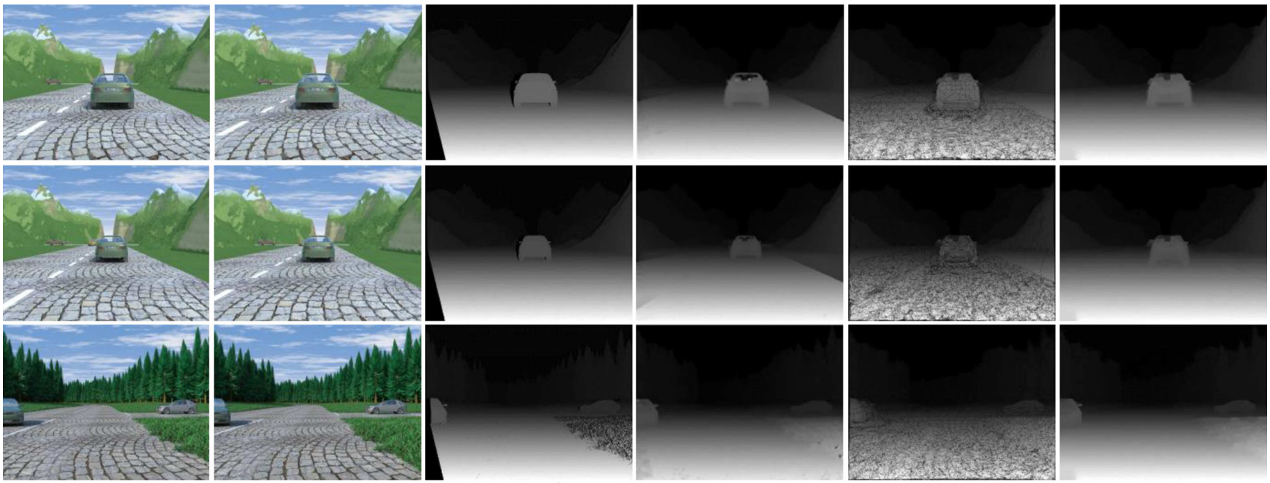
## 3.2 Results and Discussion

A few results of our approach compared to [5] are illustrated in Figure 2. We choose the same frame shown in [5]. It is noteworthy that our depth maps for these frames are generated by interpolation. Obviously both of our approach and [5] can generate satisfactory depth maps. However, our approach is much more efficient due to the employment of inter-frame redundancy. The time evaluation on fixed selective threshold mentioned in Section 3.1 of our approach is given in Table 2. OFBSM means running time of optical flow based stereo matching while MBI is motion based interpolation. And Ref. is short for refinement. Although implemented without parallelization and computation speedup, the average processing time of per frame is close to 0.5 minutes and the accuracy is also satisfactory at the same time. Furthermore, by setting reasonable selective threshold, the number of frames to be stereo matched can be reduced which means processing time of per frame can decrease sharply. We excerpt the running time from [5] and list our result on the same dataset in Table 1 for comparison. We also give a few results on the real world Karlsruhe dataset in Figure 3. It can be seen that even for real world videos our approach can generate promising results due to the reasonable strategy for frame selection and inter depth interpolation. And this suggests that our approach is competent for fast and accurate depth acquisition tasks.

At last we give the mean absolute error (MAE) curve of TrafficScene1 and TrafficScene2 in Figure 4. The first row is the MAE curve on two dataset of our method and the second is the comparison with [5]. On account of coarse motion field estimation and interpolation, error would accumulate in some frames, it is the reason why there are some peaks. It can be eliminated by certain post-processing or refinement which is not considered here. It can be seen that even the computing time is greatly reduced, the accuracy nearly maintains the same level. This indicates that our approach can generate accurate depth maps for stereo videos while preserving the inter-frame continuity well.

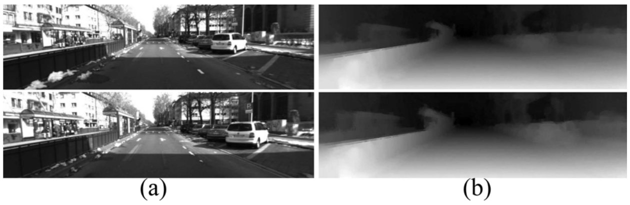
[1] http://ccv.wordpress.fos.auckland.ac.nz/eisats/set-2/

[2] http://www.cvlibs.net/datasets/karlsruhe_sequences/



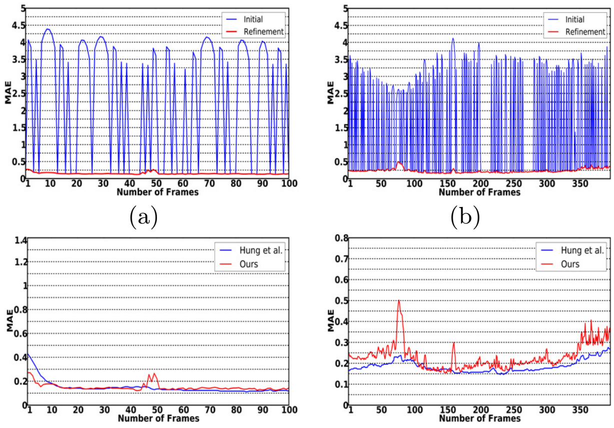
(a) $I_{L,t}$ (b) $I_{R,t}$ (c) $D_t^*$ (d) $D_t$ of [7] (e) $D_t^I$ (f) $D_t$

**Figure 2: Example of disparity estimation. (a)-(b) Color stereo images. (c) The ground truth disparity map. (d) The estimated disparity map from Hung's [5]. (e)-(f) Our initial and final inference results.**



(a) (b)

**Figure 3: Our inference disparity map from Karlsruhe dataset. (a) Left frame from stereo video. (b) Inference results of the proposed approach.**



(a) (b) (c) (d)

**Figure 4: Mean absolute error. (a)-(b) Initial and refinement results of our method. (c)-(d) Comparison with the curve in [5].**

## 4. CONCLUSION

We have proposed a fast depth inference method via bidirectional motion-based interpolation. In contrast to previous optimization techniques, our method utilize frame redundancy to save computing time. By setting out rational strategy for frame selection, increasing number of estimated depth maps meanwhile minimizing error propagation. We evaluate our approach on two open datasets and the experiments show that our approach can generate comparatively accurate and consistent depth efficiently. In the future, we look forward to improving the framework and applying our method to further multimedia applications, such as 3D reconstruction, object tracking and scene understanding.

## 5. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China (No.61321491, 61202320), Research Project of Excellent State Key Laboratory (No.61223003), Natural Science Foundation of Jiangsu Province (No.BK2012304), and National Special Fund (No.2011ZX05035-004-004HZ).

## 6. REFERENCES

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011.

[2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36. Springer, 2004.

[3] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *IJCV*, 61(3):211–231, 2005.

[4] A. Flint, D. Murray, and I. Reid. Manhattan scene understanding using monocular, stereo, and 3d features. In *ICCV*, pages 2228–2235. IEEE, 2011.

[5] C. H. Hung, L. Xu, and J. Jia. Consistent binocular depth and scene flow with chained temporal profiles. *IJCV*, 102(1-3):271–292, 2013.

[6] C. Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis.* PhD thesis, MIT, 2009.

[7] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, 106(2):115–137, 2014.

[8] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *ECCV*, pages 568–581. Springer, 2010.

[9] W. Wang, J. Zhao, W. J. Tam, F. Speranza, and Z. Wang. Hiding depth map into stereo image in jpeg format using reversible watermarking. In *ICIMCS*, pages 82–85. ACM, 2011.

[10] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *ECCV*, pages 739–751. Springer, 2008.

[11] H. Zheng, J. Yuan, and R. Gu. A novel method for 3d reconstruction on uncalibrated images. In *ICIMCS*, pages 138–141. ACM, 2011.