GPU 에서의 고속 스테레오 정합을 위한 메모리 효율적인 Belief Propagation

최영규, 윌리엄, 박인규
인하대학교 영상미디어연구실
ykchoi@cs.ucla.edu, williem_060689@hotmail.com, pik@inha.ac.kr

Memory-Efficient Belief Propagation for Stereo Matching on GPU

Young-kyu Choi, Williem, In Kyu Park
Image Media Laboratory, Inha University

요 약

Belief propagation (BP) is a commonly used global energy minimization algorithm for solving stereo matching problem in 3D reconstruction. However, it requires large memory bandwidth and data size. In this paper, we propose a novel memory-efficient algorithm of BP in stereo matching on the Graphics Processing Units (GPU). The data size and transfer bandwidth are significantly reduced by storing only a part of the whole message. In order to maintain the accuracy of the matching result, the local messages are reconstructed using shared memory available in GPU. Experimental result shows that there is almost an order of reduction in the global memory consumption, and 21 to 46% saving in memory bandwidth when compared to the conventional algorithm. The implementation result on a recent GPU shows that we can obtain 22.8 times speedup in execution time compared to the execution on CPU.

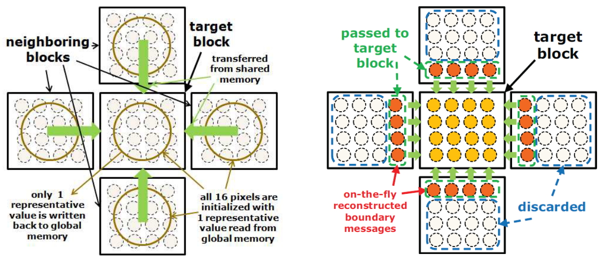
1. Introduction

Stereo matching is one of the most fundamental techniques that can be used for 3-dimensional reconstruction. The problem can be solved by well-established global energy minimization algorithms such as belief propagation (BP). However, one of the bottlenecks of such global algorithms is that they require large memory bandwidth and space. This problem becomes more severe in BP algorithm because it requires large memory space that increases linearly with both the number of pixels (N) and the number of labels (L). Thus, our aim is to devise an efficient BP algorithm which is suitable for memory-efficient MRF-based energy minimization. We concentrate on two aspects: reducing data size and transfer bandwidth.

In this paper, we propose a novel memory-efficient algorithm for BP. The data size and transfer bandwidth is significantly reduced by storing only a part of the whole message. The experimental result shows that the proposed method requires less memory space and bandwidth than conventional algorithms. It also shows that we can obtain high speed-up when the proposed method is implemented on GPU.

This paper is organized as follows. In Section 2, the proposed method is explained in detail. The experimental results are shown in Section 3. Finally, we give a conclusive remark in Section 4.

2. On-the-fly Reconstruction of Local Messages



(a) (b)

Fig 1. Proposed algorithm. (a) Data transfer in the proposed algorithm. (b) On-the-fly boundary message reconstruction

In this paper, we take advantage of local memory usage to reduce memory bandwidth and the global memory size. Thus, we propose a new method of reconstructing the local messages using shared memory in GPU. The local messages are reconstructed with the block's representative message and each pixel's matching cost. Note that the matching cost is computed on-the-fly using left/right image intensity values, as suggested in [1].

In order to reconstruct messages, first, all pixels within the block are initialized with the representative value. Next, we iteratively perform message passing within the block until convergence. However, simply performing message passing within the block will not work correctly, because there is no information about the messages from outside the block. Thus, we also have to fetch messages from the outside the block. As a result, a total of 5 representative value (up, down, left, right, center), along with the intensity of each pixel, is fetched from the global memory. This process is depicted in Fig. 1(a)

The problem with the proposed method explained in the previous section is that the representative message from neighboring block only has low-resolution information. Thus, we propose on-the-fly reconstruction of the neighboring blocks' pixel-wise messages, in addition to the reconstruction of messages of the current block. The neighboring blocks' message reconstruction process is similar to that of the current block - the message passing is performed until convergence. After convergence, only the message to the current block is saved - the rest of the messages are discarded, as shown in Fig. 1(b).

3. Experimental Results

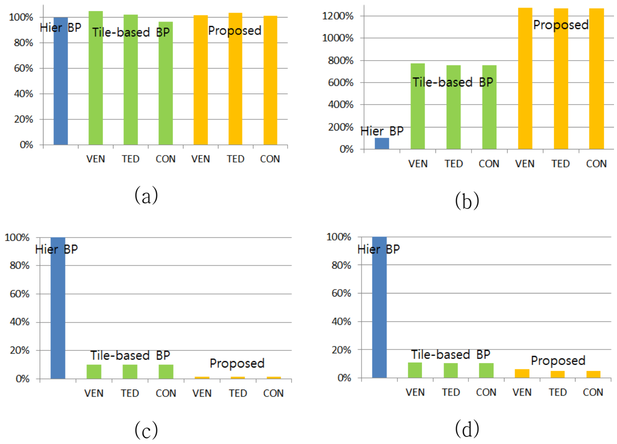


Fig 2. Comparison between hierarchical BP-M, tile based BP, and the proposed method on Tsukuba, Venus, Teddy, Cones image. The block size 8x8. (a) Minimized energy. (b) Number of message computation. (c) Consumed global memory. (d) Global↔local data transaction.

For experiment, we use Tsukuba, Venus, Cones, Teddy stereo test images from Middlebury website [2]. Fig. 2 shows the detail result for energy minimization performance, memory resource consumption, and computational complexity. The result shows that the proposed method has comparable performance to the original hierarchical BP-M and the tile-based BP. As an example, the disparity maps for Tsukuba image is shown in Fig. 3. We have implemented the proposed algorithm on NVIDIA GTX580 GPU running at 3.91GHz with 3GB global memory. For comparison, we implemented hierarchical BP on Intel CPU i7 2600 [3]. Experiment shows that the execution time for Tsukuba testset is 4.38 and 0.192 seconds on CPU and GPU, respectively.

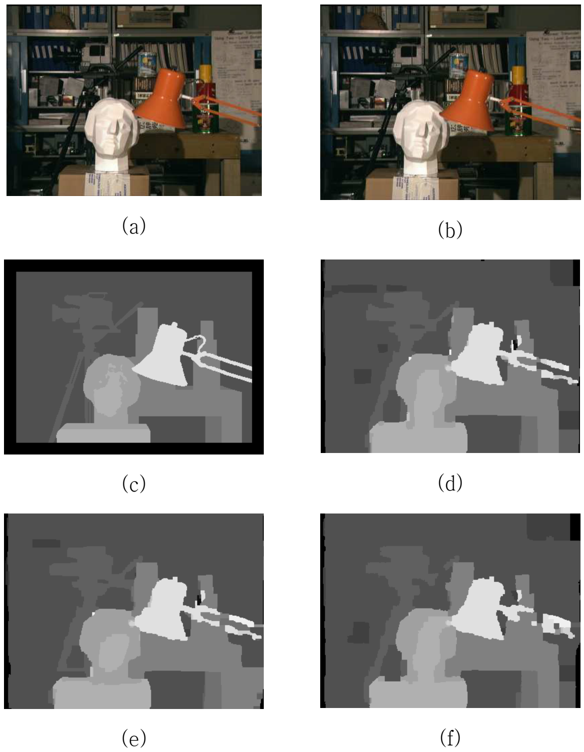


Fig 3. Depth map obtained for Tsukuba testset. (a) Left image. (b) Right image. (c) Ground truth. (d) Hierarchical BP-M. (e) Tile-based BP. (f) Proposed algorithm.

4. Conclusion

In this paper, we have described a new memory-efficient algorithm of BP in stereo matching. We have proposed on-the-fly local message reconstruction method using matching cost and reconstructed boundary message.

감사의 글

이 논문은 2012 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2012R1A1A2009495).

참고문헌

[1] C. K. Liang, C. C. Cheng, Y. C. Lai, L. G. Chen, and H. H. Chen, " Hardware-efficient belief propagation," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 525-537, May 2011

[2] D. Scharstein and R. Szeliski, " A taxonomy and evaluation of dense two frame stereo correspondence algorithms," *Int. J. Computer Vision*, vol. 47, no. 1-3, pp. 7-42, April 2004

[3] P. Felzenszwalb and D. Huttenlocher, " Efficient belief propagation for early vision," *Int. J. Computer Vision*, vol. 70, no. 1, pp. 41-54, October 2006.