Review of stereo matching algorithms for 3D vision

Lazaros Nalpantidis[1], Georgios Ch. Sirakoulis[2] and Antonios Gasteratos[1]
[1]Democritus University of Thrace, Department of Production and Management Engineering, GR-67 100 Xanthi, Greece
[2]Democritus University of Thrace, Department of Electrical and Computer Engineering, GR-67 100 Xanthi, Greece
lanalpa@pme.duth.gr; gsirak@ee.duth.gr; agaster@pme.duth.gr
Tel: +302541079359, Fax: +302541079343

ABSTRACT

Stereo vision, resulting in the knowledge of deep information in a scene, is of great importance in the field of machine vision, robotics and image analysis. As a result, in order to address the problem of matching points between two images of a stereo pair several algorithms have been proposed so far. In this paper, an explicit analysis of the existing stereo matching methods, up to date, is presented in full detail. The algorithms found in literature can be grouped into those producing sparse output and those giving a dense result, while the later can be classified as local (area-based) and global (energy-based). The presented algorithms are discussed in terms of speed, accuracy, coverage, time consumption and disparity range. Comparative test results concerning different image sizes as well as different stereo data sets are presented. Furthermore, the usage of advanced computational intelligence techniques such as neural networks and cellular automata in the development and application of such algorithms is also considered. However, due to the fact that the resulting depth calculation is a computationally demanding procedure, most of the presented algorithms perform poorly in real-time applications. Towards this direction, the development of real-time stereo matching algorithms, able to be efficiently implemented in dedicated hardware is of great interest in the contexts of 3D reconstruction, simultaneous localization and mapping (SLAM), virtual reality, robot navigation and control. Some possible implementations of stereo-matching algorithms in hardware for real-time applications are also discussed in details.

1. INTRODUCTION

Since the excellent taxonomy presented by Scharstein and Szeliski [1] many new methods have been proposed. Latest trends in the field mainly pursue real-time execution speeds, as well as decent accuracy. As indicated by this survey the algorithms' theoretical matching cores are quite well established leading the researchers towards innovations resulting in more efficient hardware implementations.

The issue of stereo correspondence is of great importance in the field of Machine Vision. It concerns the matching of points, or any other primitive, between a pair of pictures of the same scene. Assuming a calibrated stereo setup, matching points reside on corresponding horizontal lines. The disparity is calculated as the distance of these points when one of the two images is projected onto the other. The disparity values for all the image points comprise the disparity map. Once the stereo correspondence problem is solved the depth of the scenery can be estimated. This issue is of interest in

the contexts of 3D reconstruction, virtual reality, robot navigation and many other aspects of production, security, defense, exploration and entertainment.

Matching methods can be grouped into those producing sparse output and those giving a dense result. Feature based methods stem from human vision studies and are based on matching segments or edges between two images, thus resulting in a sparse output. This disadvantage, dreadful for many purposes, is counterbalanced by the accuracy and speed obtained. In order to categorize and evaluate the stereo correspondence algorithms that produce dense output a context has been proposed [1]. According to this, dense matching algorithms are classified in local and global ones. Local methods (area-based) trade accuracy for speed. The disparity computation at a given point depends only on intensity values within a finite window. Global methods (energy-based) are time consuming but very accurate. Their goal is to minimize a global cost function that combines data and smoothness terms. Of course, there are many other methods that are not strictly included in either of these two broad classes [15,16]. The issue of stereo matching has recruited a variation of computation tools. Recent attempts to confront the problem involve neural networks that are being trained according to the circumstances [4]. Others report the use of cellular automata in order to make the algorithm more adaptable to each problem [7]. Such advanced computational intelligence techniques are very interesting and promiscuous.

Most of the work done involves a theoretical description of the algorithm, a software development stage and finally the testing of the algorithm with the use of a general purpose personal computer. This methodology results in considerable running times. However, this is not the case when the objective is the development of SLAM or virtual reality systems. Such tasks require real-time, efficient performance and demand dedicated hardware and, consequently, specially developed and optimized algorithms. The evolution of FPGAs has made them an appealing choice due to the small prototyping times, their flexibility and their good performance.

2. RECENT STEREO MATCHING ALGORITHMS

The contemporary research in stereo matching algorithms is reigned by the test bench of Scharstein and Szeliski available at the www.middlebury.edu/stereo . Most of the next results are based on the standard image sets and test provided there. The most common image sets are shown in figures 1-4.

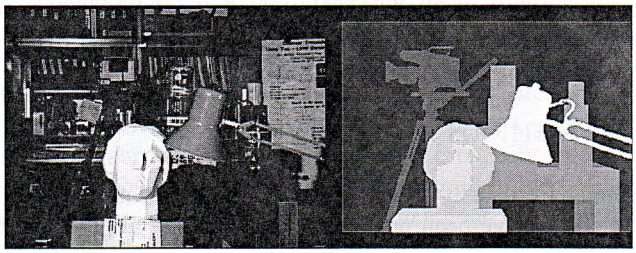

Fig. 1. Left image of the Tsukuba stereo pair (left) and ground truth (right).

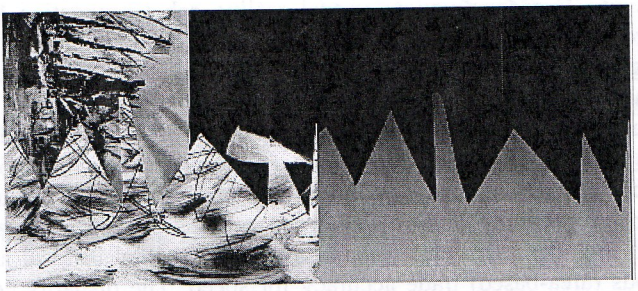

Fig. 2. Left image of the Sawtooth stereo pair (left) and ground truth (right).

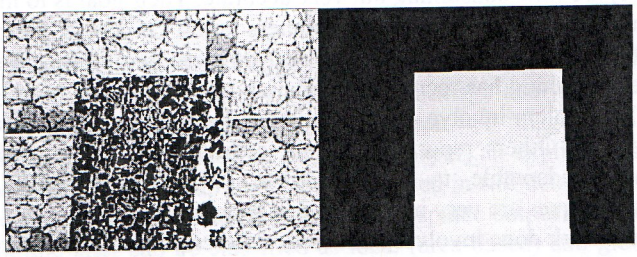

Fig. 3. Left image of the Map stereo pair (left) and ground truth (right).

Fig. 4. Left image of the Venus stereo pair (left) and ground truth (right).

2.1 DENSE DISPARITY ALGORITHMS

Methods that produce dense disparity maps gain popularity as the computational power grows. Efforts towards this direction are being reported much more frequently than towards the direction of sparse results, during the latest years.

2.1.1 LOCAL METHODS

Local methods are usually fast and can at the same time produce descent results. Several new methods have been presented. Muhlmann et al [2] describe a method that uses the SAD correlation measure for RGB color images. It achieves high speed and reasonable quality. It makes use of the left to right consistency and uniqueness constraints and applies a fast median filter to the results. It can achieve 20 fps for 160x120 pixels image size, making this method suitable for real-time applications.

A fast area-based stereo matching algorithm is presented in [3]. It is referred to as Single Matching Phase (SMP). Based on the uniqueness constraint, it rejects previous matches as soon as better ones are detected. In contrast to bidirectional matching algorithms this one performs only one matching phase, having though similar results. It uses the Sum of Absolute Differences (SAD) as the error function, but any other could be used. The results obtained are tested for reliability and sub-pixel refined. It produces a dense disparity map in real-time using a PC. It achieves 39.59 fps speed for 320x240 pixels and 16 disparity levels and the root mean square error for the standard Tsukuba pair is 5.77 .

A more advanced method can be found in [4]. This method is based on the use of the zero mean normalized cross correlation (ZNCC) as matching cost, integrated within a neural network model, which uses the least-mean-square delta rule for training. This neural network decides on the proper window shape and size for each support region. The results obtained are satisfactory but the running time needed for standard image sets renders this method not suitable for real-time applications.

Ogale and Aloimonos in [5] take into consideration the shape of the objects depicted. It demonstrates the importance of the vertical and horizontal slanted surfaces. The authors propose the replacement of the standard uniqueness constraint referred to pixels with a uniqueness constraint referred to line segments along a scanline. So the method performs interval matching instead of pixel matching. The slants of the surfaces are computed along a scanline, a stretching factor is then obtained and the matching is performed based on the absolute intensity difference. The object is to achieve minimum segmentation. The experimental results indicate 1.77%, 0.61%, 3.00% and 7.63% error percentages for the Tsukuba, Sawtooth, Venus and Map stereo pairs respectively. The execution time of the algorithm varies from 1 to 5 seconds.

Another local method that presents almost real-time performance is reported in [6]. It makes use of a refined implementation of the SAD method and a left right consistency check. The errors in the problematic regions are reduced using different sized correlation windows. Finally a median filter is used in order to interpolate the results. The performance of the algorithm is promising since it computes 7 fps for 320x240 pixels images and 32 disparity levels.

The use of cellular automata is exploited in [7]. This work presents a hardware architecture for real-time extraction of disparity maps. It is capable of processing 1Mpixels image pairs at more than 40 fps. The core of the algorithm relies on matching pixels of each scanline using a one-dimensional window and the absolute intensities differences. This method involves a pre-processing mean filtering step and a post-processing cellular automaton based filtering one. The system uses parallel structures and can easily be embodied to FPGAs.

A window-based method for correspondence search is presented [8] that uses varying support-weights. The support-weights of the pixels in a given support window are adjusted based on color similarity and geometric proximity to reduce the image ambiguity. The difference between pixel colors is measured in the CIELab color space because the distance of two points in this space is analogous to the stimulus perceived by the human eye. The running time for the Tsukuba image pair with a 35x35 pixels support window is about one minute. The error ratio is 1.29%, 0.97%, 0.99%, and 1.13% for the Tsukuba, Sawtooth, Venus and Map image sets accordingly. These figures can be further improved through a left-right consistency check.

2.1.2 GLOBAL METHODS

Global methods on other hand produce very accurate results but are very time and computational demanding due to their iterative nature. The work of Torra and Criminisi [9] presents a unified framework that allows the fusion of any partial knowledge about disparities, such as matched features and known surfaces within the scene. It combines the results from corner, edge and dense stereo matching algorithms to impose constraints that act as guide points to the standard dynamic programming method. The result is a fully automatic dense stereo system with up to four times faster run time and greater accuracy.

A method based on the Bayesian estimation theory, with a prior Markov random field model for the assigned disparities is described in [10]. The continuity, coherence and occlusion constraints as well as the adjacency principal are taken into account. The optimal estimator is computed using a Gauss-Markov random field model for the corresponding posterior marginals, which results in a diffusion process in probability space. The results are encouraging in terms of accuracy but are not suitable for real-time applications, since it takes a few minutes to process a 256x255 stereo pair with up to 32 disparity levels.

A method that uses the concept of image color segmentation is reported in [11]. An initial disparity map is calculated using an adapting window technique. The segments are combined in larger layers iteratively. The assignment of segments to layers is optimized using a global cost function. The quality of the disparity map is measured by warping the reference image to the second view and comparing it with the real image and calculating the color dissimilarity. For the 384x288 pixel Tsukuba and the 434x383 pixel Venus test set, the algorithm needed approximately 20s to produce results. For the 450x375 pixel Teddy image pair, the running time increased to 100s due to the increased scene complexity. The root mean-square error obtained is 0.73 for the Tsukuba, 0.31 for the Venus and 1.07 for the Teddy image pair.

The work done by Kim and Sohn [12] presents a two-stage algorithm consisting of hierarchical dense disparity estimation and vector field regularization. The dense disparity estimation is accomplished by a region dividing technique that uses a Canny edge detector and a simple SAD function. The results are refined by regularizing the vector fields by means of minimizing an energy functional. The root-mean-squared error obtained from this method is 0.9278 and .9094 for the "Head and Lamp" and Sawtooth image pairs. The running times are 6.765s and 9.496s respectively.

A completely uncommon measure is used in [13]. This wok describes an algorithm which is focused on achieving contrast invariant stereo matching. It relies on multiple spatial frequency channels for local matching. The measure for this stage is the deviation of phase difference from zero. The global solution is found by a fast non-iterative left right diffusion process. Occlusions are found by enforcing the uniqueness constraint. The algorithm is able to handle significant changes in contrast between the two images and can handle noise in one of the frequency channels. The implementation of the algorithm needs 2 to 4 seconds to process standard image pairs.

Another algorithm that generates high quality results in real time is reported in [14]. It is based on the minimization of a global energy function comprising of a data and a smoothness term. The hierarchical belief propagation iteratively optimizes the smoothness term but it achieves fast convergence by removing redundant computations involved. In order to accomplish real-time operation authors take advantage of the

parallelism of graphics hardware. Experimental results indicate a 16 fps performance for 320x240 pixel images and 16 disparity levels.

2.1.3 OTHER METHODS

There are of course some other methods that could not be placed in either of previous categories. Such a method, based on the continuous wavelet transform (CWT) is found in [15]. It makes use of the redundant information that results from the CWT. Using one-dimensional orthogonal and biorthogonal wavelets as well as two-dimensional orthogonal wavelet the maximum matching rate obtained is 88.22% for the Tsukuba pair. Upsampling the pixels in the horizontal direction by a factor of two, through zero insertion, further decreases the noise and the matching rate is increased to 84.91%.

Another work [16] presents an algorithm based on non-uniform rational B-splines (NURBS) curves. The curves replace the edges extracted with a wavelet based method. The NURBS are projective invariant and so they reduce false matches due to distortion and image noise. Stereo matching is then obtained by estimating the similarity between projections of curves of an image and curves of another image. A 96.5% matching rate is reported for this method.

2.2 SPARSE DISPARITY ALGORITHMS

Algorithms resulting in sparse, or semi-dense, disparity maps tend to be less attractive. Though, very interesting ideas flourish in this direction. Veksler presented an algorithm [17] that detects and matches dense features between the left and right images of a stereo pair, producing a semi-dense disparity map. A dense feature is a connected set of pixels in the left image and a corresponding set of pixels in the right image such that the intensity edges on the boundary of these sets are stronger than their matching error. They are computed during the stereo matching process. The algorithm computes 1 fps with 14 disparity levels for the Tsukuba pair producing 66% density, 0.06% absolute average error and 0.36 total error in the non occluded regions.

Another method [18] developed by Veksler is based on the same basic concepts as the former one. The main difference is that this one uses the graph cuts algorithm for the dense feature extraction. As a consequence this algorithm produces semi-dense results with significant accuracy in areas where features are detected. The results are significantly better considering density and error percentage but require longer running times. For the Tsukuba pair the density obtained is 75%, the total error in the non occluded regions 0.36% and the required time 6 seconds. For the Sawtooth pair the corresponding results are 87%, 0.54% and 13 seconds.

Gong and Yang in their paper [19] propose a dynamic programming (DP) algorithm, called reliability-based DP (RDP) that uses a new measure to evaluate the reliabilities of matches. According to this the reliability of a proposed match is the cost difference between the globally best disparity assignment that includes the match and the one that does not include it. The interscanline consistency problem, common to the DP algorithms, is reduced through a reliability thresholding process. The result is a semi-dense unambiguous disparity map. It achieves 76% density, 0.32% error rate and 16 fps for the Tsukuba and 72% density, 0.23% error rate and 7 fps for the Sawtooth image

pair. Accordingly, the results for Venus and Map pairs are 73%, 0.18%, 6.4 fps and 86%, 0.7%, 12.8 fps. It is seen that the required times are encouraging for real-time operation if a semi-dense disparity map is acceptable.

A near-real-time stereo matching technique is presented in another paper [20], which is based on the reliability based dynamic programming algorithm. This algorithm can generate semi-dense disparity maps. Two orthogonal RDP passes are used to search for reliable disparities along both horizontal and vertical scanlines. Hence, the interscanline consistency is explicitly enforced. It takes advantage of the computational power of programmable graphics hardware, which further improves speed. It results in 85% dense disparity map with 0.3% error rate at 23.8 fps for the Tsukuba pair, 93% density, 0.24% error rate at 12.3 fps for the Sawtooth pair, 86% density, 0.21% error rate at 9.2 fps for the Venus pair and 88% density, 0.05% error rate at 20.8 fps for the Map image pair. If needed, the method can also be used to generate more dense disparity maps deteriorating the execution time.

3. REAL-TIME HARDWARE IMPLEMENTATIONS

As we have already mentioned there are plenty of applications that could use real-time disparity map extraction. The need for dedicated hardware is more evident in the case of autonomous units. Moreover conventional computers find it difficult to handle all the computations needed in real-time speed. These tasks generally demand sparse output.

The hardware implementation of a global algorithm is not very appealing. As stated above, global methods are extremely time and computational demanding because of their iterative nature. This is also the reason that prevents them from being implemented with parallel structures. In addition, global algorithms require odd, rather than simple and straightforward, implementations. Parallelism and simplicity are key factors, available in dedicated hardware implementations, that can reduce the required running times.

In contrast, local methods could be greatly benefited by the use of such parallel and straightforward structures. There are several works [21–24] that describe local methods custom implemented on hardware. What most of these have in common is that they choose to implement a rather simple local algorithm (SAD) and make extensively use of parallelism of computations. Performance is refined by custom choices during the hardware architecture development phase.

All of the aforementioned methods can achieve real-time operation. The use of FPGAs is now the most convenient and reasonable choice for hardware development. They are cheap and perform extremely well. The variety of available tools make the prototyping times very short. Another advantage is that the resulting hardware implementation is open for further upgrades. Thus, FPGA implementations are very flexible and fault tolerant. On the other hand ASIC implementation [24] is an option as well, but it is more expensive, except of the case of massive production. The prototyping times are considerable longer and the result is highly process-dependent. Any further changes are difficult and additionally time and money consuming. Their performance supremacy does in most cases not justify choosing ASICs. All these reasons make FPGA implementations preferable. Table 1. presents the main characteristics of the FPGA implemented works.

Author	Matching Cost	Aggregation	Image Size	Disparity Levels	Window Size	Speed (fps)	Device
Arias-Estrada et al [21]	SAD	fixed window	320×240	16	7x7	71	Xilinx Virtex XCV800HQ 240-6
Lee et al [22]	SAD	fixed window	640x480	64	32x32	30	Xilinx Virtex-II XC2V8000
Hariyama et al [23]	SAD	adaptive window	64x64	64	8x8 (max)	5263	Altera APEX20KE

Table 1. FPGA implementations' characteristics.

4. CONCLUSIONS

The stereo correspondence problem remains an active area for research. More and more modern applications demand not only accuracy but real-time operation as well. It seems that both area and energy based methods walk towards this objective with satisfactory results. But when it comes for hardware implementation it is the simpler local algorithms that outperform any others. The use of FPGAs facilitates the development and is preferred by most of the active researchers.

5. ACKNOWLEDGEMENTS

The work is supported by the E.U. funded project "View-Finder", FP6-IST-2005-045541.

REFERENCES

[1] D. Scharstein, R. Szeliski, "A taxonomy and evaluation of dense two frame stereo correspondence algorithms", IJCV 7, 1/3, 2002.

[2] K. Muhlmann, D. Maier, J. Hesser and R. Manner, "Calculating Dense Disparity Maps from Color Stereo Images, an Efficient Implementation", IJCV 47, 1/2/3, pp. 79–88, 2002.

[3] Luigi Di Stefano, Massimiliano Marchionni, Stefano Mattoccia, "A fast area-based stereo matching algorithm", Image and Vision Computing 22, pp. 983–1005, 2004.

[4] Elisabetta Binaghi, Ignazio Gallo, Giuseppe Marino, Mario Raspanti, "Neural adaptive stereo matching", Pattern Recognition Letters 25, pp. 1743–1758, 2004.

[5] Abhijit S. Ogale and Yiannis Aloimonos, "Shape and the Stereo Correspondence Problem", IJCV 65 , 3, pp. 147–162, 2005.

[6] Sukjune Yoon, Sung-Kee Park, Sungchul Kang, Yoon Keun Kwak, "Fast correlation-based stereo matching with the reduction of systematic errors", Pattern Recognition Letters 26, pp. 2221–2231, 2005.

[7] L. Kotoulas, A. Gasteratos, G.Ch. Sirakoulis, C. Georgoulas, and I. Andreadis , "Enhancement of Fast Acquired Disparity Maps using a 1-D Cellular Automation Filter", Proceedings of the Fifth IASTED International Conference on Visualization, Imaging and Image Processing, Benidorm, Spain, September 7-9, 2005.

[8] Kuk-Jin Yoon and In So Kweon, "Adaptive Support-Weight Approach for Correspondence Search", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, April 2006.
[9] P.H.S. Torra, A. Criminisi, "Dense stereo using pivoted dynamic programming", Image and Vision Computing 22, pp. 795–806, 2004.
[10] Salvador Gutierrez, Jose Luis Marroquin, "Robust approach for disparity estimation in stereo vision", Image and Vision Computing 22, pp. 183–195, 2004.
[11] Michael Bleyer, Margrit Gelautz, "A layered stereo matching algorithm using image segmentation and global visibility constraints", ISPRS Journal of Photogrammetry & Remote Sensing 59, pp. 128– 150, 2005.
[12] Hansung Kim, Kwanghoon Sohn, "3D reconstruction from stereo images for interactions between real and virtual objects", Signal Processing: Image Communication 20, pp. 61–75, 2005.
[13] Abhijit S. Ogale and Yiannis Aloimonos, "Robust Contrast Invariant Stereo Correspondence", Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, April 2005.
[14] Qingxiong Yang, Liang Wang, Ruigang Yang, Shengnan Wang, Miao Liao, David Nister, "Real-time Global Stereo Matching Using Hierarchical Belief Propagation", The British Machine Vision Conference (BMVC), 2006.
[15] Xiaodong Huang, Eric Dubois, "Dense disparity estimation based on the continuous wavelet transform", CCECE 2004 - CCGEI 2004, Niagara Falls, May 2004.
[16] C. Liu, W. Pei, S. Niyokindi, J. C. Song and L. DWang, "Micro stereo matching based on wavelet transform and projective invariance", Meas. Sci. Technol. 17, pp. 565–571, 2006.
[17] Olga Veksler, "Dense Features for Semi-Dense Stereo Correspondence", IJCV 47, 1/2/3, pp. 247–260, 2002.
[18] Olga Veksler, "Extracting Dense Features for Visual Correspondence with Graph Cuts", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.
[19] Minglun Gong and Yee-Hong Yang, "Fast Unambiguous Stereo Matching Using Reliability-Based Dynamic Programming", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, June 2005.
[20] Minglun Gong and Yee-Hong Yang, "Near Real-time Reliable Stereo Matching Using Programmable Graphics Hardware", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
[21] Miguel Arias-Estrada, Juan M. Xicotencatl, "Multiple Stereo Matching Using an Extended Architecture", FPL 2001, LNCS 2147, pp. 203-212, 2001.
[22] Sung Hwan Lee, Jongsu Yi, and JunSeong Kim, "Real-Time Stereo Vision on a Reconfigurable System", SAMOS 2005, LNCS 3553, pp. 299-307, 2005.
[23] Masanori Hariyama, Yasuhiro Kobayashi, Haruka Sasaki and Michitaka Kameyama, "FPGA Implementation of a Stereo Matching Processor Based on Window-Parallel-and-Pixel-Parallel Architecture", IEICE Trans. Fundamentals, vol.E88-A, no.12, December 2005.
[24] Masanori Hariyama, Haruka Sasaki and Michitaka Kameyama, "Architecture of a Stereo Maching VLSI Processor Based on Hierarchically Parallel Memory Access", IEICE Trans. Inf. & Syst., vol.E88-D, no.7 July 2005.