Real-Time Scan-Line Segment Based Stereo Vision for the Estimation of Biologically Motivated Classifier Cells

M. Salah E.-N. Shafik and Bärbel Mertsching

GET Lab, University of Paderborn
Pohlweg 47 - 49, 33098 Paderborn, Germany
[shafik, mertsching]@get.uni-paderborn.de
http://getwww.uni-paderborn.de

Abstract. In this paper we present a real-time scan-line segment based stereo vision for the estimation of biologically motivated classifier cells in an active vision system. The system is challenged to overcome several problems in autonomous mobile robotic vision such as the detection of incoming moving objects and estimating its 3D motion parameters in a dynamic environment. The proposed algorithm employs a modified optimization module within the scan-line framework to achieve valuable reduction in computation time needed for generating real-time depth map. Moreover, the experimental results showed high robustness against noises and unbalanced light condition in input data.

1 Introduction

During the last decade, the importance of depth estimation from stereo images has been extensively increased active vision applications e.g. autonomous mobile robots. The research presented in this paper is a building block of a comprehensive system of multi-object motion analysis in robotic vision which is intended to be included in active vision applications [1, 2]. In such systems the challenge of avoiding multiple moving obstacles overcoming the ego-motion problem of the mobile robot is raised. In this context, priority is given to the fast detection and 3D motion estimation of incoming and outgoing objects towards the mobile robot. Hence, a biologically motivated classifier approach presented in [3] has been improved to integrate depth information generated from a developed stereo matching algorithm. The need for developing a real-time stereo approach is not only constrained by the available computation time in the system, but also by the available hardware and its power consumption, especially if there is no specialized hardware or graphical processing unit (GPU) at hand. Hence, every millisecond counts.

Stereo algorithms can be classified based on the taxonomy proposed by Scharstein and Szelinski [4] into two major categories: local methods and global methods. While most of the global methods attempt to minimize an energy function computed for the entire image area, local methods determine the stereo correspondence between two frames by minimizing a matching cost using an aggregation window. Local algorithms in general are suitable for real-time applications. The main problem is to determine the suitable size and shape of the aggregation window in order to avoid crossing depth discontinuities while covering sufficient intensity variation. Such problems results in noisy

disparities in texture-less regions and blurred object boundaries. On the other hand, global methods such as Dynamic Programming (DP) [5], Belief Propagation (BP) [6] and Graph Cut (GC) [7] make explicit smoothness assumptions on the disparity map. A major problem in DP approaches is the requirement of enforcing the monotonicity, i. e., the relative ordering of pixels on a scan-line remains the same between two frames, which cause errors in case of narrow foreground objects. Although BP and GC algorithms gives impressive results by enforcing the optimization in two-dimensions, they are computationally expensive and are not suitable for real-time applications as most of the global algorithms. Between local and global approaches, there are some algorithms which are based on the minimization of an energy function computed over a subset of the whole image area or along the epipolar lines. The implemented minimization strategy is usually based on Dynamic Programming or Scan-line Optimization (SO) [8] techniques and recently on line segmentation [9].

In this paper, we developed fast SO-based optimization technique for a line segment stereo approach. The generated depth map is used to update and optimize an active weight function $\omega(p)$ in order to estimate biologically motivated classifier cells ($c - cells$).

The remainder of the paper is organized as follows: section 2 gives an account of the related work, while section 3 describes the proposed algorithm and the structure of the $c - cells$. Section 4 discusses the results of experiments and evaluates the proposed method, while finally, section 5 concludes the paper.

2 Related Work

Recent stereo algorithms have significantly advanced the state-of-the-art in terms of quality. However, in terms of speed, the best stereo algorithms typically take from several seconds to several minutes to generate a single disparity map ([10] gives a performance evaluation of cost aggregation strategies proposed for stereo matching). There are many interesting applications in which depth maps at video rate are needed such as (autonomous mobile robots, augmented-reality and automatic vehicle guidance). Hence, the importance of real-time stereo algorithms increases as in [11] where an adapted recursive formulation is proposed to reduce the computing cost of SAD cost function of a local approach which in turn inherits the ambiguity problem from the local algorithms. As a solution to overcome this problem, they implement a post processing filter application at the last phase of the algorithm which is considered as an overhead to the computation time. On the other hand, [12] overcomes the ambiguity problem in low textured areas by replacing estimates in texture-less regions with fitting planes. The algorithm starts with window-based multi-view stereo matching followed by the application of consistency fusion module. Afterword, a plane-fitting phase is applied by using color segmentation, where a plane is adjusted for each segment. In order to correct the potential errors and enhance the overall computation time, a GPU hierarchical recursive belief propagation approach is implemented. In spite of using a GPU, the BP-based plane-fitting stereo pipeline runs at about 1 frame per second. The use of a GPU has been introduced before in global approaches with hierarchical BP [13] and DP based on adaptive cost aggregation [14]. As the use of a GPU due to hardware con-

straints is not applicable on some platforms, solving the low texture problem using an effective variable support based on image segmentation within the SO framework has been addressed in [15]. While the result is promising, the performance is far from being real-time (i. e. some minutes). The computational time has been improved by using line segment techniques and tree dynamic programming as in [9]. The segmentation module there contains three steps: computing the initialization marks, repositioning marks, and removing isolated marks. In order to extract linear planes, a parameter estimation approach is used for fitting planes on sparse correspondence. Afterward, dynamic programming is used on the constructed tree to minimize the energy function. The algorithm has performed well with an Intel Pentium IV 2.4 GHz processor (processing time for "tsukkuba" [4] is about 160ms). However, the requirement of enforcing the monotonicity inherited from the DP techniques still cause the thin foreground objects problems. On the other hand, a biologically motivated classifier can be derived from neurophysiological studies [16], which found that there are two classes of neurons in the middle temporal area (MT) of the mammalian visual cortex which are considered to encode motion information from the retinal signal: direction selective D-cells and velocity selective S1-cells. Both cells covering small receptive fields of three types of neurons which have been identified in the medial superior temporal area (MTS): R-cells detecting rotation, S-cells for expansion, and D-cells for translation. Furthermore, cells which combine all three motion types give evidence for the composition of elementary motions into more complex motion patterns similar to motion templates [17]. In this module, motion-sensitive cells for the preferential direction will be constructed in order to measure the response of the sensitive cells to a corresponding motion which helps predicting hazardous scenarios such as collisions. The generation of the corresponding map requires calculating a radial symmetric weight function for each cell which is in term of computation time is very expensive. Developing a fast connection weight function based on the depth information reduce the time consumption dramatically without the use of a GPU power as in [18] where a biologically motivated classifier and feature descriptors are designed for execution on single instruction multi data hardware using the programmable GPU.

3 Proposed Algorithm

The goal of stereo algorithms is to establish pixel correspondences between the left image I_l and the right image I_r. In order to achieve reasonable results, two geometric constraints are used: first on the imaging systems, i. e. the input stereo images are *rectified* where the epipolar lines are aligned with corresponding scan-lines. And second on the scene, i. e. the *smoothness assumption* where the disparity map is smooth almost everywhere except at the border of objects assuming that scene is composed of smooth structures which in the case of autonomous mobile robots applications is not granted.

The first step of the proposed technique is the line segmentation of the reference image I_r, in which the epipolar line epl_y is segmented into different labels $l_i(epl_y) \in \Gamma$ in the label space Γ based on the color differences between a seed pixel $P_s^{l_i}$ and the neighborhood pixels of the same epipolar line:

$$P_r(x+k,y) \in l_i(epl_y) \qquad \forall\, |P_s^{l_i} - P_r(x+k,y)| < \tau_c \tag{1}$$

where τ_c is the Euclidean color distance threshold. The correspondence problem is formulated as an energy minimization function between segments of the input images.

$$E(d_\Gamma) = argmin(E_{data}^{l_i}(d_\Gamma) + E_{smooth}^{l_{(i,d)\in K}}(d_\Gamma)) \tag{2}$$

where $E(d_\Gamma)$ is the disparity map of line segments and K is an adjacent segment in a particular epipolar line epl_y. The data term $E_{data}^{l_i}(d_\Gamma)$ of the energy function is the matching cost between a segment $l_i(epl_y)$ in the reference image and the opponent segments $l_{i,d}(epl_y)$ in the target image. The smoothness term $E_{smooth}^{l_{(i,d)\in K}}(d_\Gamma)$ encodes the smoothness assumption.

3.1 Matching Cost and Optimization

In order to reduce the complexity of calculations, a matching cost C_M based on the absolute color difference between the points of the current segment in the reference image and all disparity hypotheses is used to evaluate the data term

$$C_M(P_r, P_{l,d}) = \sum_{c\in\Re} |P_r^c(x,y) - P_{l,d}^c(x+d,y)| \tag{3}$$

where $\Re$ is the RGB color space, $P_r(x,y) \in I_r$, $P_{l,d}(x,y) \in I_l$, and d is the hypothesized disparity value.

The data term is computed from the sum of the matching cost along the segment points

$$E_{data}^{l_i}(d_\Gamma) = \sum_{P=P_s}^{P_r^N} C_M(P_r, P_{l,d}) \qquad \forall\, P_r \in l_i(epl_y) \tag{4}$$

where P_s is the starting seed pixel of a line segment and P_r^N is the last point in the same segment.

In order to enhance the optimization performance, we propose an effective and simplified smoothness term within the SO-framework.

$$E_{smooth}^{l_{(i,d)\in K}}(d_\Gamma) = \lambda(\ell_{l_i}) \,.\, |d_{l_i} - d_{l_K}| \tag{5}$$

where $\lambda(\ell_{l_i})$ is an ascending function to the length of the current segment ℓ_{l_i} used to penalize depth discontinuities. The concept behind the function is to balance the relation between the disparity of a segment and the sum of the matching cost of the segment points. While the matching cost is affected by the length of the segment, only one disparity value is assigned to all the segment points and the best value is chosen within a winner take all (WTA) scheme. Considering the inter-scan-line smoothness resulting from line segmentation leads to overcome the ambiguity problem without the use of a recursive smoothing function as in BP approaches or facing narrow front objects problem as in DP algorithms.

3.2 Biologically Motivated Classifier

In [3], a model neuron for the biologically motivated classifier cell $c_i(p_0)$ represents the instantaneous velocity in the preferential direction of the cell

$$c_i(p_0) = \frac{1}{|V|} \sum_{p \in V} \frac{\omega(p_0 - p)\psi_i^L(p)}{|\psi_i^L|^2} v(p) \tag{6}$$

where $v(x, y)$ is a motion vector (MV) at an image point (x, y) generated by a motion template ψ_i^L which represents the six infinitesimal generators of a 2D vector field (translation in x,y,z and rotation around x,y,z axis). V denotes the set of MVs detected in the image. Instead of using a radial symmetric weight function $\omega(p) = \frac{1}{2\pi\sigma^2} exp(-\frac{1}{2}\frac{|p|^2}{\sigma^2})$ to create a local neighborhood around p_0, we use a connection weight function $\varpi(p)$ based on the depth information generated from the previous stereo module

$$\varpi(p_0 - p) = \begin{cases} 1 & \forall \ p \in \Im(p_0) \\ 0 & otherwise \end{cases} \tag{7}$$

where $\Im(p_0)$ is the segment label off point p_0. The new connection weight function enhances the overall computation time as well as it overcomes the blurring effect of the $\omega(p)$ function especially at the edges of an object. Moreover, considering only the points that belongs to the same depth level improves the estimation process overcoming the ambiguous interpretation problem.

Generating a response map requires a validation function to measure how well is the correspondence of a *c-cell* to a preferential motion.

$$\xi_i(p_0) = \frac{1}{|V|} \sum_{p \in V} |\varpi(p_0 - p)(v(p) - c_i(p_0)\psi_i^L(p))| \tag{8}$$

where $c_i(p_0)$ is the estimated *c-cell*, and $\varpi(p_0 - p)$ the updated connection weight function between a point p_0 and its neighbor point p.

4 Results and Discussion

In this section, the results of applying the proposed algorithm on different images will be presented and discussed. In order to analyze the output of our approach in a scalable complex scene, a virtual environment for simulating a mobile robot platform (SIMORE) is used [19, 20]. The simulator provides a possibility to test the outcome of the proposed algorithm in different ideal scenarios, where complex 3D motions of multiple objects are not easily available in a real-time. In this environment, the simulated robot is moving forward and steering towards the left in front of a stable cube, a moving cone, and a size-changing ball. The experiments have been performed on a P4 personal computer (PC) with 3.00 GHz speed and 2 GB RAM.

In an effort to reduce the overall computation time, the depth from the stereo approach is applied without a refinement step depending on the enhancement done by the modified smoothing function. Fig. 1 represents a qualitative comparison of the proposed

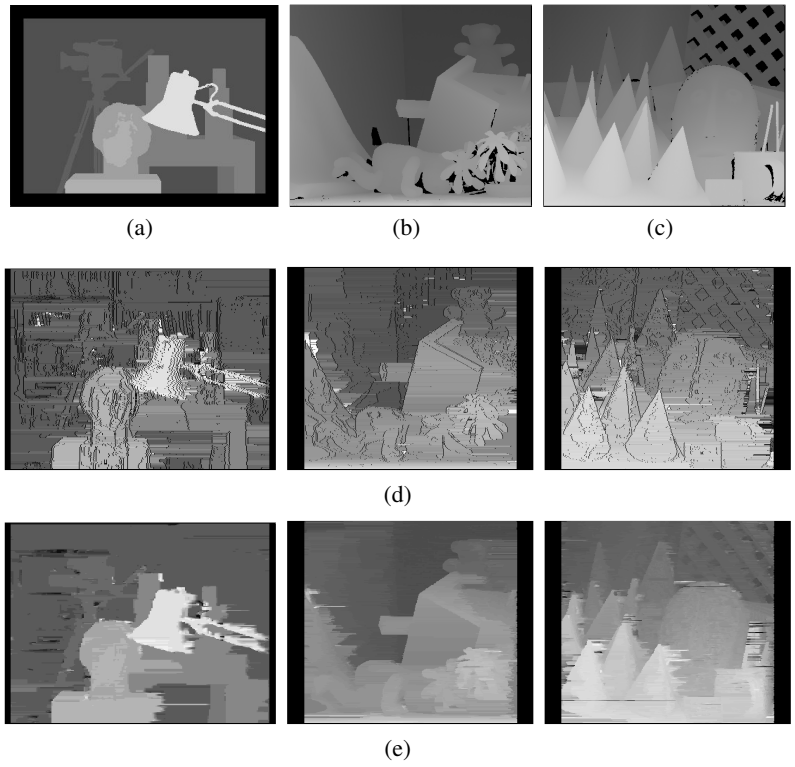


Fig. 1. Qualitative comparison of the generated depth map: (a-c) Ground truth data for the three images from the Middle-bury data-set (Tsukkuba, Teddy, and Cones). Result of the proposed line segment based stereo algorithm (d) without and (e) with the use of the modified smoothing function.

algorithm to the ground truth of the Middlebury data-set [4, 21]. The second row depicts the generated depth map without the use of the smoothing function, while the third row represents the result of the stereo approach using the smoothing function. The result shows that when the smoothing function is applied, it provides a better quality. However, the use of the smoothing function increases the processing time by about 5$\sim$9 ms which is not very critical for real-time application. On the other hand, the result without the smoothing function is affected by noise but this is not very critical to the depth perception.

The second result as shown in fig. 2 represents the response of the motion sensitive cells using image sequence from SIMORE. The generated depth maps are implemented in the estimation of the *e cells* process, where the first and the second row represent the response of the motion sensitive cells in the preferential direction of the translation in x, y, and z and the rotation around x, y, and z axis, respectively. The response maps are valuable in tackling typical problems in autonomous mobile robotic vision such as the detection and estimation of a specific motion pattern of multiple moving objects in

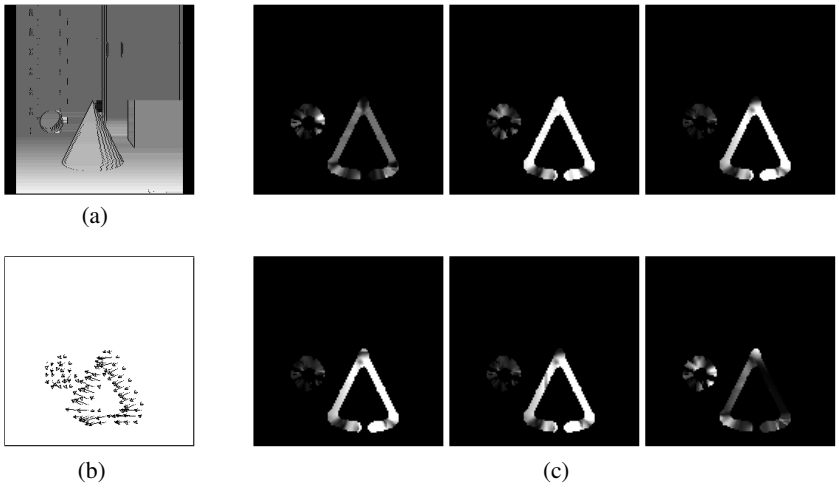


Fig. 2. Response maps of motion sensitive neuron using image sequence from SIMORE. (a) Generated depth map. (b) Generated MVF. (c) Response map of (*e cell*)in the preferential directions, up: in the translation in the x, y, and z axis, down: around the x, y, and z axis.

a dynamic environment, and the ego-motion problem resulted from the motion of the mobile robot.

The overall computation time used for generating the proposed line segment based depth approach from the Tsukkuba image is about 60∼70 ms without any use of the GPU due to the energy constraint of the real platform (tele-operated mobile robot). The computation time enhancement provided by the proposed algorithm is vital to speed-up the overall estimation process of the biologically motivated classifier cells, which is intended to be implemented in the tele-sensory mobile robot.

5 Conclusion

In this paper, we have presented a real-time scan-line segment based stereo vision employing a modified optimization module within the scan-line framework. The achievement of reducing the computation time provides the ability to implement the generated depth map into the estimation process of biologically motivated classifier. Moreover, updating the connection weight function of the motion sensitive cells using the depth information helps in the fast detection of incoming objects in a dynamic environment.

References

1. Shafik, M., Mertsching, B.: Enhanced motion parameters estimation for an active vision system . Pattern Recognition and Image Analysis **18**(3) (September 2008) 370 – 375
2. Aziz, Z., Mertsching, B.: Fast and robust generation of feature maps for region-based visual attention. IEEE Trans. on Image Processing (5) (January 2008) 633 – 644

3. Massad, A., Jesikiewicz, M., Mertsching, B.: Space-variant motion analysis for an active-vision system. In: Advanced Concepts for Intelligent Vision Systems, Ghent, Belgium (2002)
4. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. Journal of Computer Vision **47**(1/2/3) (June 2002) 7 – 42
5. Lei, C., Selzer, J., Yang, Y.: Region-tree based stereo using dynamic programming optimization. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition **2** (2006) 2378 – 2385
6. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. 18th Int. Conf. on Pattern Recognition, ICPR 2006 **3** (2006) 15 – 18
7. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions using graph cuts. In: Proc. Int. Conf. Computer Vision (ICCV). (2001) 508 – 515
8. Hirschmuller, H.: Stereo vision in structured environments by consistent semi-global matching. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (2) (June 2006) 2386 – 2393
9. Deng, Y., Lin, X.: A fast line segment based dense stereo algorithm using tree dynamic programming. In: ECCV. (2006) 201 – 212
10. Tombari, F., Mattoccia, S., Stefano, L.D., Addimanda, E.: Classification and evaluation of cost aggregation methods for stereo correspondence. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Florida, USA (December 2008)
11. Foggia, P., Limongiello, A., Vento, M.: A real-time stereo-vision system for moving object and obstacle detection in avg and amr applications. In: Proc. of the Seventh Int. Workshop on Computer Architecture for Machine Perception (CAMP), Washington, DC, USA (2005) 58 – 63
12. Yang, Q., Engels, C., Akbarzadeh, A.: Near real-time stereo for weakly-textured scenes. In: British Machine Vision Conference (BMVC), Leeds, UK (2008)
13. Yang, Q., Wang, L., Yang, R., Wang, S., Liao, M., Nistér, D.: Real-time global stereo matching using hierarchical belief propagation. In: BMVC. (2006) 989 – 998
14. Wang, L., Liao, M., Gong, M., Yang, R., Nister, D.: High-quality real-time stereo using adaptive cost aggregation and dynamic programming. Int. Symposium on 3D Data Processing Visualization and Transmission (2006) 798 – 805
15. Mattoccia, S., Tombari, F., Stefano, L.D.: Stereo vision enabling precise border localization within a scanline optimization framework. In: ACCV. (2007) 517 – 527
16. Albright, T.D., Desimone, R., Gross, C.G.: Columnar organisation of directionally selective cells in visual area mt of the macaque. J. Neurophysiol **51** (1984) 16 – 31
17. Duffy, C.J., Wurtz, R.H.: Sensitivity of mst neurons to optic flow stimuli. i. a continuum of response selektivity to large-field stimuli. J. Neurophysiol **65** (1991) 1329 – 1345
18. Woodbeck, K., Roth, G., Chen, H.: Visual cortex on the GPU: Biologically inspired classifier and feature descriptor for rapid recognition. CVPRW (June 2008) 1 – 8
19. Mertsching, B., Aziz, Z., Stemmer, R.: Design of a simulation framework for evaluation of robot vision and manipulation algorithms. In: Proceedings of Asia Simulation Conference (ICSC). (2005) 494–498
20. Kutter, O., Hilker, C., Simon, A., Mertsching, B.: Modeling and simulating mobile robots environments. In: 3rd Int. Conf. on Computer Graphics Theory and Applications, Funchal, Madeira, Portugal (January 2008)
21. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition **1** (June 2003) 195 – 202