A Stereo Depth Recovery Method Using Layered Representation of the Scene

Tarkan Aydin and Yusuf Sinan Akgul

GIT Vision Lab, http://vision.gyte.edu.tr/
Department of Computer Engineering, Gebze Institute of Technology
Gebze, Kocaeli 41400 Turkey
taydin@gyte.edu.tr
akgul@bilmuh.gyte.edu.tr

Abstract. Recent progresses in stereo research imply that performance of the disparity estimation depends on the discontinuity localization in the disparity space which is generally predicated on discontinuities in the image intensities. However, these approaches have known limitations at highly textured and occluded regions. In this paper, we propose to employ a layered representation of the scene as an approximation of the scene structure. The layered representation of the scenes was obtained by using partially focused image set of the scene. Although self occlusions are still present in real aperture imaging systems, our approach does not suffer from the occlusion problems as much as stereo and focus/defocus based methods. Our disparity estimation method is based on synchronously optimized two interdependent processes which are regularized with a nonlinear diffusion operator. The amount of diffusion between the neighbors is adjusted adaptively according to information in the layered scene representation and temporal positions of the processes. The system is initialization insensitive and very robust against local minima. In addition, it accurately handles the depth discontinuities. The performance of the presented method has been verified through experiments on real and synthetic scenes.

1 Introduction

Recovering 3D information of a scene from 2D images is one of the most fundamental tasks in computer vision. A wide range of image cues have been used to accomplish this task (e.g degree of focus/defocus, stereo correspondence, etc.). Among them, stereo methods try to estimate spatial shifts in the images captured from different views by establishing the visual correspondences between them. The main difficulty of the correspondence problem is the ambiguity due to the image noise, repeated texture, and occlusions that make the mathematical formulation of the problem ill-posed. Therefore, a regularization strategy should be employed by imposing prior assumptions about scene geometry or by including additional information about the scene. A commonly adopted approach is to formulate the problem in an energy minimization framework by explicitly introducing a smoothness term which allows retrieval of piecewise smooth disparity maps. Although considerable progress has been achieved in minimization of the energy functional, solutions with lower energy values does not necessarily result in higher performance values [1]. Therefore researchers tend to include additional information from images of the scene related to its geometry.

One of the most meaningful information for the depth estimation methods is the possible locations of depth discontinuities. Many state of the art stereo algorithms utilize image intensity variations to align depth discontinuities with the intensity discontinuities. Local methods use this information to shape support windows adaptively [2]. Recently, several global stereo algorithms have been proposed that match segments rather than pixels in the optimization process using graph cuts and belief propagation methods [3, 4]. Similarly, diffusion based methods include edge information in their anisotropic operators to supervise flow between neighbors [5].

Another serious challenge for stereo is the occlusion problem which means that some scene points are not visible from all views. Because it is very hard to extract depth values of occluded regions using only their intensity values, use of active illumination were proposed [6, 7]. In order to overcome the limitations associated with the intensity based discontinuity localization and to detect surface structure of the occluded regions without using any illumination setup, we propose to employ more reliable and practical information from other cues.

This paper presents a system that represents the scene in a layered form in which layers are ordered according to their distance to the image plane. Layered form of the scene is extracted from the image set of the scene which are captured from the same view by focusing to the virtual layers in the scene. The layered scene structure is employed by our stereo method to determine the possible depth discontinuity locations. Our approach has partial biological motivations because it is known that human vision uses both stereo and focus for depth estimation. Pentland [8] has reported that human perception of depth is strongly influenced by the gradient of focus as a useful source of depth information.

In order to extract the layers, we establish correspondences between the images in the set and the all-focus image of the scene from the same view. Our approach is closely related to shape from focus (SFF) methods that recover depth of a scene from intensity variations in the images by searching sharpest image sections. However, the occlusion problem inherent to real aperture lenses makes the SFF results ambiguous [9]. With the employment of the all focus image of the scene, we formulate the layer extraction as a correspondence problem. As a result, our method does not suffer from the occlusion problem as much as SFF.

Our previous work introduced synchronous optimization processes for the stereo depth estimation [10]. The optimizations are expressed as two separate but dependent processes which are iteratively minimized to deform two initial surfaces towards each other using a gradient descent minimization. Although gradient descent method does not guarantee optimality and it highly depends on initial settings, due to the interaction between the optimization processes, the overall result of our system is always better than the results achievable by a single optimization process. Reliable convergence is ensured by starting each process with different initial positions.

In this paper, we also introduce a novel nonlinear diffusion operator which is specifically designed to utilize both layered representation and temporal information in the synchronous processes. It performs isotropic smoothing and anisotropic smoothing adaptively around inhomogeneous regions. Unlike the previous diffusion techniques, the proposed operator adapts itself to meta-states of both processes allowing depth discontinuities to emerge during the optimization process.

The proposed method does not need any calibration procedures other than the stereo rectification. In other words, it does not need registration between stereo disparity values and layer depth ordering because we use the layering information only for the depth discontinuity detection. The alternative of focus setting-disparity registration would be too difficult to achieve for real life situations due to complex calibration routines required [11].

The rest of this paper is organized as follows. Section 2 explains the layered form of the scene and synchronous energy functional. Section 3 describes the system validation and experiments. Finally, we provide concluding remarks in Section 4.

2 Method

2.1 Layered Representation of The Scene

The visible surface of a scene from one view can be approximated by non-overlapping fronto parallel layers. Availability of this information will be very helpful to estimate possible locations of depth discontinuities to align them across the layer boundaries. Many state of the art stereo methods have taken advantage of this approach by simply applying color segmentation on the images[3, 4]. However, excessive number of layers may be produced for highly textured images. In addition, relative distances or ordering of the layers cannot be established with color segmentation.

In our system, we approximate the surface of the scene as a composition of ordered layers. Layers are extracted from image set of the scene in which each image is captured from the same view by focusing to the successive virtual planes in the scene. The images are taken with the widest lens aperture setting so that the scene points that lie in the focused layer can easily be detected due to shallow depth of field. Layer assignment to the image points is accomplished by establishing correspondences between images in the set and the all focus image of the scene from the same view that we have already for the stereo system. Note that the system setup and data acquisition method is similar to that of shape from focus (SFF) [12] methods in which depth is reconstructed from multiple images which are taken with different focus settings. However, our layer assignment strategy is completely different from the classical SFF, hence it does not suffer from occlusion problem which is present in finite aperture imaging systems [9].

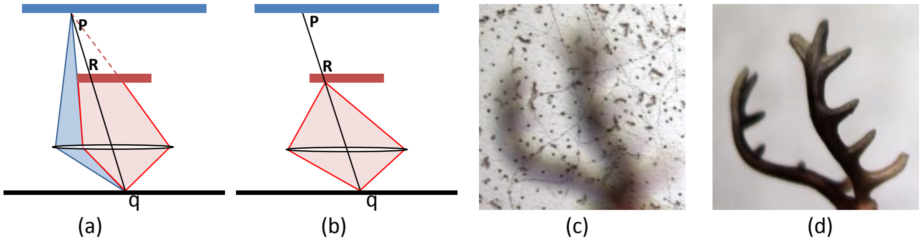

Fig. 1. Image formation process of a scene with two fronto parallel layers where focus is set to (a) background and (b) foreground. The point q may appear focused when focus is set to both foreground or background. Real images of a sample scene taken with focus setting set to (c) background and (d) foreground.

In the presence of the occlusions, image points of the occluding object receive a mixture of light from both focused background and blurred foreground when focus is set to the occluded. Consequently, corresponding image points of occluding object may appear focused, even though it is out of focus. It makes depth estimation for these regions ambiguous [9]. This situation is illustrated in Fig. 1. As seen in Fig. 1.(c) occluding object cause attenuation in the brightness profile of the occluded region[13].

In order to address the ambiguity problem, we establish correspondences between focused images of the virtual planes in the scene and the all focus image of the scene from the same view. Assuming the attenuation in the brightness profile is constant in a small patch, we use the normalized cross-correlation as a similarity measure because it is insensitive to the brightness differences.

$$C_i(x,y) = NCC_{\Omega}(I_p(x,y), I_i(x,y)), \tag{1}$$

where I_p is all focused image, and I_i is the images in the set.

The robustness of the matching score C depends on the size of the patch Ω. Although increasing it also increases the reliability of matching by reducing the effect of noise, it results in shifts at estimated location of discontinuities. In order to increase robustness while preserving location of discontinuities, we aggregate initial matching scores with larger but adaptively weighted support windows [14]. Weights of support windows are assigned by using all focus image of the scene. Computation of the weights is based on similarity and proximity metrics between center pixel and its neighboring pixels that fall inside the support window. Similar and closer pixels get larger weights in the assumption that they probably lie on the same surface. Weights are computed using all-focused image I according to the following formula.

$$\omega_{x_0y_0}(x,y) = e^{-(\Delta d/\gamma_1 + \Delta I/\gamma_2)} \tag{2}$$

where

$$\Delta d = \sqrt{(x-x_0)^2 + (y-y_0)^2}, \tag{3}$$

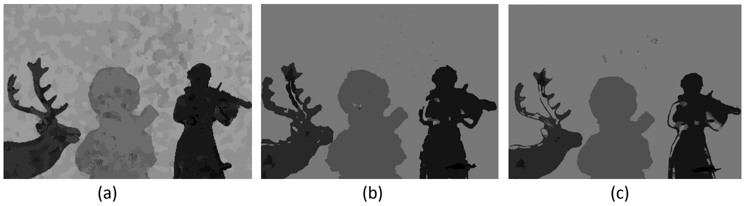

Fig. 2. Our layered representation of the scene (a), results from adaptive shape from focus [14] (b), and shape from focus method [12](c). Although our layered representation does not hold correct depth values, it approximates the structure of the scene more accurately than shape from focus methods.

and

$$\Delta I = \| I(x,y) - I(x_0,y_0) \| \tag{4}$$

Δd and ΔI are euclidian distance in spatial domain and in color space respectively. γ_1 and γ_2 are constant parameters to supervise relative weights.

Using the computed correlation values, matching scores are aggregated.

$$C_i^{'}\left(x_0,y_0\right) = \sum_{(x,y)\in \Omega(x_0,y_0)} \omega_{x_0 y_0}\left(x,y\right) C_i(x,y) \quad i = 1..N \tag{5}$$

where N is the number of virtual planes. Then each image point is assigned to a layer with maximum correlation value

$$I_l(x,y) = \underset{1\leq i\leq N}{\arg\max}\, C_i^{'}(x,y) \tag{6}$$

Figure 2 shows the layers of image of a sample scene and depth image of the scene found by traditional SFF[12] and [14].

2.2 Synchronous Energy Formulation For Stereo

Our stereo energy formulation is built upon synchronous optimization processes which employs segmented image to estimate depth discontinuities in the scene [10]. Starting from the classical stereo energy functional, synchronous optimization processes was formulated by introducing E_{tns} term as in following equations which are intended to be optimized synchronously and to produce two different disparity maps D_1 and D_2.

$$E\left(D_1\right) = \alpha E_{data}\left(D_1\right) + \beta E_{smth}(D_1) + \lambda_t E_{tsn}\left(D_1, D_2\right) \tag{7}$$

$$E\left(D_2\right) = \alpha E_{data}\left(D_2\right) + \beta E_{smth}(D_2) + \lambda_t E_{tsn}\left(D_2, D_1\right) \tag{8}$$

As in the classical stereo energy formulation, the data term E_{data} is for satisfying the image similarity requirement.

$$E_{data} = \int \phi(D) dp \tag{9}$$

The similarity measure is calculated by using normalized cross correlation (NCC) values between the left and right image regions due to its robustness against any brightness differences between the left and right images. In order to reject outliers in data and increase convergence time of the method, data space should be pre-smoothed while preserving the discontinuities recover them accurately. Therefore, we pre-smooth data space with the bilateral filter [15] whose kernels derived from left image of stereo pairs.

The term E_{smth} enforces smoothness to desired disparity map.

$$E_{smth} = \int c \left|\nabla D\right|^2 dp, \tag{10}$$

where c is the diffusion coefficient, which inhibits the smoothing across the marked discontinuities.

The tension energy E_{tsn} is for the synchronization of the two optimization processes and it is the core idea of the synchronous optimization method.

$$E_{tsn}\left(D_1, D_2\right) = \int \left(D_1 - D_2\right)^2 dp, \tag{11}$$

The main function of the tension term is to make the disparity values of two surfaces D_1 and D_2 get close to each other by pushing the optimization process with the worse data term towards the other process.

Minimization of the energy functionals defined in Equation 7 and 8 with the tension terms yields following equations:

$$\frac{\partial D_1}{\partial t} = \gamma\left(\alpha\frac{\partial \phi_{D_1}}{\partial D_1} + \beta\nabla\cdot(c\nabla D_1) + \lambda_t(D_1 - D_2)\right) \tag{12}$$

$$\frac{\partial D_2}{\partial t} = \gamma\left(\alpha\frac{\partial \phi_{D_2}}{\partial D_2} + \beta\nabla\cdot(c\nabla D_2) + \lambda_t(D_2 - D_1)\right). \tag{13}$$

The introduced tension energy enforces both processes to converge the same solution. However, continually forcing both process to pull each other may result in convergence to an irrelevant local minima. In order to prevent processes to force each other symmetrically, we set λ_t to a spatially and temporarily varying coefficient. The coefficient for the first process is computed as

$$\lambda_{t_1} = \begin{cases} 1 - e^{(\frac{\Delta\phi(D_1)}{\lambda})^2} & \Delta\phi(D_1) \geq 0 \\ 0 & otherwise \end{cases}, \tag{14}$$

where $\Delta\phi(D_1) = \phi(D_1) - \phi(D_2)$ and λ is a constant. The same coefficient will be computed for the second process analogously.

Note that, the tension is not symmetric anymore and it is heavily dependent on the local positions of the processes, hence it computes a different value for each process. The process with lower data term has the zero coefficient and its optimization is not affected from the other process. On the other hand, if the process has higher data energy than the other, it will be pulled by the tension term towards the other process.

2.3 Discontinuity Preserving Nonlinear Regularization

Anisotropic regularizer can be used to prevent diffusion between inhomogeneous regions, i.e. across the discontinuities to prevent surface discontinuities to be oversmoothed. Consequently, an anisotropic disparity regularization process requires prior information about possible location of discontinuities. One practical and reasonable prediction can be made by analyzing intensity variations in the stereo images. Assuming that depth discontinuities overlap with some intensity discontinuities in the image, Alvarez [5] adjusted the amount of diffusion among the neighboring elements according to the intensity difference between them. One negative consequence of this assumption is oversmoothing of depth discontinuities that have small intensity variations at corresponding positions in the image.

The diffusion constant c is defined as a function of gradient of image I in anisotropic smoothing as

$$c(x,y,t) = g\left(\nabla I\right), \tag{15}$$

where g is called edge stopping function.

In order to increase the accuracy of depth recovery around discontinuities, the intensity based discontinuity estimation step should be replaced by a more reliable and robust estimation. We propose to employ our layered representation described in section 2.1 to take advantage of its superior performance at discontinuity localization. Using layered image I_l, we define the edge stopping function g as,

$$g\left(\nabla I_l\right) = e^{-\left(|\nabla I_l|/\kappa_1\right)^2}, \tag{16}$$

where κ_1 is a constant. If we minimize the resulting equations using diffusion coefficient in Equation 16, the processes turn out to be sensitive to local minima, especially around noisy and small surface patches in the estimated depth map because they cannot get sufficient flow from the neighbors. Note that this situation does not occur in isotropic regularization in which diffusion coefficient c is taken as a constant so that flow is allowed between all regions. As a result, it is very robust against local minima but it oversmooths the discontinuities.

In order to take advantage of both isotropic and anisotropic regularization, we introduce a novel diffusion operator which adapts itself to the meta-states of the synchronous processes and performs isotropic or anisotropic regularization adaptively. The operator utilize the temporal position information of synchronous processes. The temporal distance between the processes is given as

$$\mathit{\Delta} d(x,y,t) = \left|D_1\left(\mathrm{x,y,t}\right) - D_2\left(\mathrm{x,y,t}\right)\right|. \tag{17}$$

Initially, one of the synchronous processes is started from minimum disparity values and the other is started from maximum disparity values. Therefore, $\mathit{\Delta} d$ has the maximum value that is possible. At this time, the regularization should be isotropic to avoid getting stuck to a local minima. During the minimization, the processes get close to each other and $\mathit{\Delta} d$ goes to zero. In order to prevent smoothing of discontinuities, the regularization should behave anisotropically, as the processes approach the desired minimum. At the end ($\Delta d = 0$), the operator should exhibit pure anisotropic behavior.

We define diffusion function $c\,(x, y, t)$ by including the distance between the synchronous processes as

$$c\,(x, y, t) = (1 - h(\Delta d)) + h(\Delta d) \cdot g(\nabla I_l), \tag{18}$$

where

$$h\,(\Delta d) = e^{-(\Delta d/\kappa_2)^2}, \tag{19}$$

where κ_2 is a constant. Initially, Δd has high values and h evaluates to nearly zero. If h is zero, the diffusion coefficient functions as a isotropic diffusion coefficient. When the processes find the same positions ($\Delta d = 0$), the h would be 1 and the diffusion coefficient evaluates to $g\,(\nabla D_f)$ and it functions as a isotropic diffusion coefficient.

3 Experiments

The proposed method has been tested on real scenes and real scenes with synthetically refocused images which are generated with iris filter [16]. In all experiments, matching cost is pre-smoothed with 11x11 pixel sized bilateral filter. The scenes with known depth maps are obtained from Middlebury [17] image base where ground truth information for stereo pairs are available for benchmarking. 16 refocused images are produced for Venus and Cones data sets which have sharp depth discontinuities. Error rates of the proposed algorithm are computed for non-occluded areas, near discontinuities, and for complete images. Table 1 compares the error rates of disparity map obtained by employing proposed layered from of the scene and segmented image. Figure 3 shows the layered representation, and resulting disparity maps of our algorithm. The results show that our method can robustly recover piecewise smooth surfaces and preserve discontinuities well.

	Venus						Cones					
error threshold	1 pixel			0.5 pixels			1 pixel			0.5 pixels		
	vis	all	disc	vis	all	disc	vis	all	disc	vis	all	disc
our method	0.10	0.25	1.41	0.5	0.89	5.30	5.74	12.0	16.4	8.24	15.0	21.3
with Segment[10]	0.32	0.40	3.45	1.23	1.52	9.55	7.08	14.5	19.7	8.34	16.2	22.4

Table 1. Error rates on visible, all, and discontinuous regions for our method using layered representation of the scenes and segmented image as a source of discontinuity estimation.

Experiments on real scenes are performed using 25mm lenses on c-mount cameras. By changing focus setting, only 10 images are captured from only the left view of the stereo setup to reconstruct layered form. Our results with left stereo images are shown in Fig. 4.

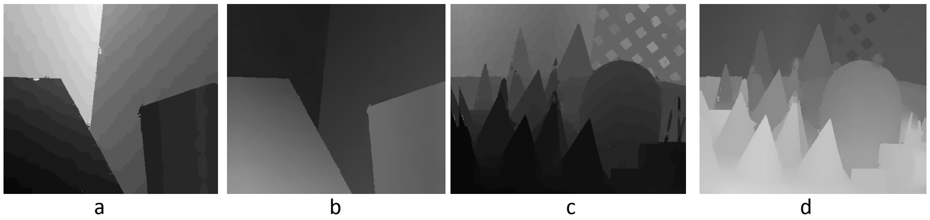

Fig. 3. Layered representation of venus (a) and cones (c) image and their corresponding disparity images (b,d) found by our method.

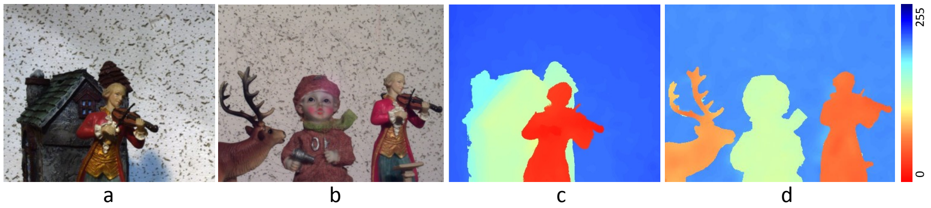

Fig. 4. Left stereo images of the stereo pairs (a,b), and resulting disparity maps from our method(c,d).

4 Conclusions

We proposed a novel system that uses two energy functionals which are optimized by two dependent optimization processes. Unlike the intensity based regularization methods, we utilized the layered representation of the scene as a source of discontinuity estimation. Our layered representation requires focused images of the virtual planes in the scene which can be obtained by changing focus setting of one of the stereo cameras. Consequently, the system can be easily implemented with a simple setup.

We also introduced a novel nonlinear diffusion operator which effectively utilize the layered representation of the scene and temporal positions of the synchronous processes. The operator is capable of adapting itself to meta-states of synchronous processes and perform isotropic or anisotropic regularization accordingly.

Although the proposed system does not include an explicit occlusion mechanism, by using layered representation of the scene, the proposed anisotropic operator propagates disparity values inside homogenous regions and fills values of occluded regions from its neighbors lying in the same surface. This may not be possible in the intensity or segment based diffusion methods, especially when occluded regions have high intensity variation.

Despite the advantages of our method, currently, it cannot handle the situation in which disparity discontinuities are located inside the assigned layers.

5 Acknowledgements

This work was conducted at the Computer Vision Laboratory at Gebze Institute of Technology. It was supported by TUBITAK Career Project 105E097.

References

1. Tappen, M.F., Freeman, W.T.: Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In: International Conference on Computer Vision. (2003) 900–907
2. Yoon, K.J., Kweon, I.S.: Adaptive support-weight approach for correspondence search. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(4) (2006) 650–656
3. Zhang, Y., Kambhamettu, C.: Stereo matching with segmentation-based cooperation. In: ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part II. (2002) 556–571
4. Hong, L., Chen, G.: Segment-based stereo matching using graph cuts. In: IEEE Computer Vision and Pattern Recognition or CVPR. (2004) I: 74–81
5. Alvarez, L., Deriche, R., Sanchez, J., Weickert, J.: Dense disparity map estimation respecting image discontinuities: A pde and scale-space based approach. Journal of Visual Communication and Image Representation **13**(1/2) (March 2002) 3–21
6. Raskar, R., Tan, K.H., Feris, R., Yu, J., Turk, M.: Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. ACM Trans. Graph. **23**(3) (2004) 679–688
7. Zickler, T.E., Belhumeur, P.N., Kriegman, D.J.: Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. Int. J. Comput. Vision **49**(2-3) (2002) 215–227
8. Pentland, A.P.: A new sense for depth of field. IEEE Trans. Pattern Anal. Mach. Intell. **9**(4) (1987) 523–531
9. Schechner, Y.Y., Kiryati, N.: Depth from defocus vs. stereo: How different really are they? Int. J. Comput. Vision **39**(2) (2000) 141–162
10. Aydin, T., Akgul, Y.: Stereo depth estimation using synchronous optimization with segment based regularization. Technical report, Gebze Institude of Technology, Kocaeli, Turkey (2008)
11. Ahuja, N., Abbott, A.: Active stereo: Integrating disparity, vergence, focus, aperture and calibration for surface estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(10) (1993) 1007–1029
12. Nayar, S., Nakagawa, Y.: Shape from focus. PAMI **16**(8) (August 1994) 824–831
13. Asada, N., Fujiwara, H., Matsuyama, T.: Seeing behind the scene: analysis of photometric properties of occluding edges by the reversed projection blurring model. Computer Vision, IEEE International Conference on **0** (1995) 150
14. Aydin, T., Akgul, Y.: A new adaptive focus measure for shape from focus. In: BMVC08. (2008)
15. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV. (1998) 839–846
16. Sakurai, R.: Irisfilter. "http://www.reiji.net/" (2004)
17. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47**(1-3) (April 2002) 7–42