COST AGGREGATION WITH ANISOTROPIC DIFFUSION IN FEATURE SPACE FOR HYBRID STEREO MATCHING

Bumsub Ham[1], Dongbo Min[2], Kwanghoon Sohn[1]

Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea[1]
Advance Digital Science Center, Singapore[2]
khsohn@yonsei.ac.kr

ABSTRACT

In this paper, we present a cost aggregation using anisotropic diffusion on a feature space for hybrid stereo matching. Stereo matching can be classified into two categories: feature-based and area-based approaches. Feature-based approaches generate accurate but sparse disparity maps. On the other hand, area-based approaches generate dense but unreliable disparity maps, especially at depth discontinuities and homogeneous regions. We hence propose a stereo matching algorithm having advantages of both approaches. We study how to design a correspondence algorithm without modeling any depth cues except disparity. A procedure of depth perception is modeled via anisotropic diffusion on the feature space in terms of coherence. Based on the assumption that similar local feature space has similar disparity, we define the feature space and its similarity and then introduce feature confidences into the proposed model. Experimental results show that the performance of the proposed method is comparable to that of the state-of-the-art methods.

Index Terms— Stereo matching, cost aggregation, feature based matching, anisotropic diffusion, feature space analysis.

1. INTRODUCTION

Stereo matching has been one of the most important problems in computer vision task. Many researchers in this field have studied a correspondence problem due to its numerous applications, e.g., image based rendering (IBR), 3D reconstruction, robot vision, surveillance and so on. Solving the problem is, however, still challenging due to matching ambiguities especially at homogeneous (or repeated patterns) and occlusion regions. Many algorithms have been proposed with several constraints in order to relax these ambiguities. A comprehensive review of correspondence algorithms can be found in [1].

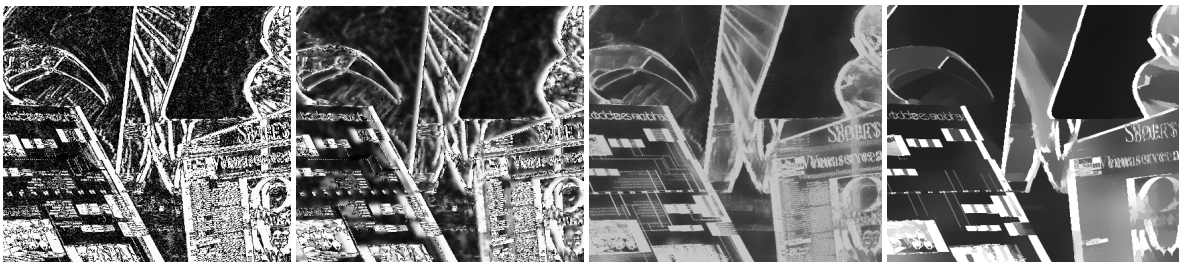
Correspondence algorithms generally can be classified into two categories (feature- and area-based approaches) according to the density of the disparity map [2]. In feature-based approaches, structural information such as edge, corner and texture is used. It produces reliable but sparse disparity maps. The performance largely depends on how many reliable features are detected. It is difficult to find distinct features and match their correspondences due to the outliers. Veksler proposed semi-dense stereo matching algorithm which matches the detected dense features [2]. However, it processes each scanline independently and produces sparse disparity maps. Jawahar and Narayanan proposed a generalized correlation framework which can combine the different image features [3].

Area-based approaches produce dense but unreliable disparity maps. It can be classified into two categories according to modeling of the smoothness assumption: Local and global approaches. Global approaches explicitly leverage the smoothness assumption into the energy model. It defines an energy function which consists of fidelity and smoothness term, and solves it by an inference algorithm such as belief propagation and graph cuts. Local (or window-based) approaches implicitly make smoothness assumption at the cost aggregation step. That is, all pixels in a support window have similar disparities. Therefore, localization, i.e., adaptively selecting window sizes and shapes, is important to aggregate only appropriate costs. Moreover, the foreground fattening may occur in case that there exist multiple disparities in a window. To overcome this problem, Kanade and Okutomi used an adaptive window size in order to localize only relevant disparities [4]. However, this method cannot localize well at depth discontinuities due to its rectangular window shape. Recently, Yoon and Kweon proposed adaptive support-weight approach [5]. This method aggregated costs based on the similarity and proximity weight by modeling the gestalt principles in a fixed window. Although this approach can handle successfully at depth discontinuities, it cannot handle homogeneous regions. In addition, classification and evaluation of the cost aggregation method can be found in [6].

In this paper, we propose a cost aggregation method using anisotropic diffusion on the feature space. It is worthy of note that the proposed method has advantages of both feature- and area-based approaches. The organization of this paper is as follows: In section 2, we study how to design a correspondence algorithm without modeling any depth cues but disparity. Section 3 discusses the proposed cost aggregation with anisotropic diffusion by defining the feature space and its similarity. The experimental results are shown in section 4. Finally, in section 5, we conclude with a brief summary.

2. DESIGN OF CORRESPONDENCE MATCHING

There are a number of cues which can help human perceive depth. Among them, monocular and binocular cues have been widely used for designing a correspondence matching algorithm. It is, however, impossible to perfectly model several depth cues. We pose how to find corresponding points without modeling any depth cues except disparity. Fortunately, it was shown that depth can be perceived even in the absence of monocular cues and binocular cues, which was proved by a picture called 'Random dot stereogram' [7]. It showed that the correspondence matching is



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

Fig. 1. Results of aggregated cost in 'Venus' for (a) initial matching cost, (b) anisotropic diffusion [14], (c) Adaptive weight [5], (d) proposed method when disparity is 0.

based on simple process of finding *connected clusters* formed by neighboring points with similar brightness. This observation is applied to the cost aggregation in stereo matching. A set of these points is defined as a feature space in this paper. Based on the assumption that two points are more likely to have similar depth if two feature spaces are similar, we define a new metric for measuring similarity between feature spaces and then formulate the cost aggregation method based on this metric in the following section.

3. HYBRID COST AGGREGATION

3.1. Feature space

A feature space is defined as a set of points having similar brightness [7]. It is not easy to measure a similarity between feature spaces since they have arbitrary shapes and sizes. Therefore, comparing a representative value only can be an alternative way of measuring the similarity. Local mode can be used as a representative value in feature space since it characterizes the property of the space well. In other words, the local maxima in local histogram of each space are used for measuring the similarity of feature space. From now on, it is called *coherence* which represents the similarity of feature space. Finding local modes has been an important issue in the field of early vision. It was shown that local mode filtering, robust estimation and mean shift are equivalent [8]. In this paper, we find the local modes by using mean-shift analysis [9] as follows.

Let $\mathbf{I}_\mathbf{p}$ be a vector which represents generalized pixel in 5D spaces as follows:

$$\mathbf{I}_\mathbf{p} = (\mathbf{p}^\mathbf{T}, \mathbf{c}_\mathbf{p}^\mathbf{T})^\mathbf{T} \tag{1}$$

where $\mathbf{p} = (x, y)^\mathbf{T}$ and $\mathbf{c}_\mathbf{p} = (L, a, b)^\mathbf{T}$ represent spatial and color information (CIE-Lab color), respectively. Local modes can be found by iteratively computing mean vectors, followed by the translation of the kernel by using the mean shift vector in Eq. (2).

$$\mathbf{m}(\mathbf{I}_\mathbf{p}) = \frac{\sum_{\mathbf{s}\in N_\mathbf{p}} \mathbf{I}_\mathbf{s}\, \mathrm{g}_p\left(\|\mathbf{p}-\mathbf{s}\|\right) \mathrm{g}_c\left(\|\mathbf{c}_\mathbf{p}-\mathbf{c}_\mathbf{s}\|\right)}{\sum_{\mathbf{s}\in N_\mathbf{p}} \mathrm{g}_p\left(\|\mathbf{p}-\mathbf{s}\|\right) \mathrm{g}_c\left(\|\mathbf{c}_\mathbf{p}-\mathbf{c}_\mathbf{s}\|\right)} - \mathbf{I}_\mathbf{p} \tag{2}$$

$\mathrm{g}_p(\cdot)$ and $\mathrm{g}_c(\cdot)$ are the spatial and color kernels with bandwidth p and c, respectively, and $N_\mathbf{p}$ represents the neighborhood of $\mathbf{p}$. Let us denote $\tilde{\mathbf{I}}_\mathbf{p} = (\tilde{\mathbf{p}}^\mathbf{T}, \tilde{\mathbf{c}}_\mathbf{p}^\mathbf{T})^\mathbf{T}$ the convergence point with initial condition $\mathbf{I}_\mathbf{p} = (\mathbf{p}^\mathbf{T}, \mathbf{c}_\mathbf{p}^\mathbf{T})^\mathbf{T}$. A feature space can then be represented as follows.

$$\mathbf{F}_\mathbf{p} = (\mathbf{p}^\mathbf{T}, \tilde{\mathbf{c}}_\mathbf{p}^\mathbf{T})^\mathbf{T} \tag{3}$$

3.2. Cost aggregation via anisotropic diffusion on feature space

Prazdny showed that a disparity gradient is a function of feature similarity, i.e., *more dissimilar features allows larger disparity gradients* [10]:

$$\|\nabla \mathbf{D}\| \propto 1 / h\left(\|\nabla \mathbf{F}\|\right) \tag{4}$$

where $\mathbf{D}$ and $\mathbf{F}$ represents the depth and feature space, respectively. $\|\cdot\|$ represents the norm of vector. $h(x)$ is a monotonically decreasing function which satisfy $h(x) \to 0$ as $x \to \infty$. It coincides with our assumption as discussed in section 2, i.e., the similarity of the feature space is closely related to the similarity of the depth. The role of the function $h(x)$ is the same as the "edge-stopping" function in anisotropic diffusion [11], which enables anisotropic diffusion to be applied to the cost aggregation for stereo matching.

First, we calculate initial matching cost volume by shifting the target image further to the opposite direction of the reference image and then subtracting it from the reference image. We describe the cost aggregation via anisotropic diffusion as in Eq. (5). Let $\mathbf{E}$ be 2D cost plane which is a section of initial 3D cost volume.

$$\frac{\partial \mathbf{E}}{\partial t} = \nabla \cdot \left(\mathrm{g}_f\left(\|\nabla \mathbf{F}^\mathrm{R}\|\right) \mathrm{g}_f\left(\|\nabla \mathbf{F}^\mathrm{T}\|\right) \nabla \mathbf{E} \right) \tag{5}$$

The superscripts R and T represent reference and target images, respectively. $\mathrm{g}_f(\cdot)$ is a monotonically decreasing function on feature space with bandwidth f. This function satisfies $\mathrm{g}_f(x) \to 0$ as $x \to \infty$ in order to stop diffusion across different feature space.

We call $\mathrm{g}_f\left(\|\nabla \mathbf{F}^\mathrm{R}\|\right)$ and $\mathrm{g}_f\left(\|\nabla \mathbf{F}^\mathrm{T}\|\right)$ as *intra-coherence* of the reference and target image, respectively. It represents coherence within an image. That is, a value of function is high if two points have similar feature spaces, which means that they belong to similar depth spaces as in Eq. (4). *Inter-coherence* is defined as a product of intra-coherence of the reference and target images. Therefore, two points are likely to have similar depth spaces as both intra-coherences are high. It also means that they belong to the similar feature space in both images. We discretize Eq. (5) using forward Euler approximation with initial condition $\mathbf{E}^0$ as in Eq. (6).

$$\begin{aligned} &\mathbf{E}^{t+1}(\mathbf{p},\mathbf{d}) = \mathbf{E}^t(\mathbf{p},\mathbf{d}) \\ &+\lambda \sum_{\mathbf{s}\in N_\mathbf{p}} \mathrm{g}_f\left(\|\nabla \mathbf{F}^\mathrm{R}(\mathbf{s},\mathbf{p})\|\right) \mathrm{g}_f\left(\|\nabla \mathbf{F}^\mathrm{T}(\mathbf{r},\mathbf{q})\|\right) \nabla \mathbf{E}^t(\mathbf{s},\mathbf{p}) \end{aligned} \tag{6}$$

where $\mathbf{q} = \mathbf{p} + \mathbf{d}$ and $\mathbf{r} = \mathbf{s} + \mathbf{d}$ with $\mathbf{d} = (d, 0)^{\mathbf{T}}$ in the target images are the corresponding points of $\mathbf{p}$ and $\mathbf{s}$ in the reference image, respectively. λ is a time step which controls the rate of diffusion, t is evolution parameter. $N_{\mathbf{p}}$ represents the neighborhood of $\mathbf{p}$. Note that the neighborhood is extended in order to aggregate appropriate costs more reliably in contrast to conventional anisotropic diffusion, so that the gradient is approximated to the difference between pixels as follows [12]:

$$\nabla \mathbf{F} \approx \mathbf{F}_{\mathbf{s}} - \mathbf{F}_{\mathbf{p}} \equiv \nabla \mathbf{F}(\mathbf{s}, \mathbf{p}) \tag{7}$$

3.3. Feature confidence

$\mathbf{F}$ can be referred to as sets of diffused local mode in a feature space, so that a point $\mathbf{I}_{\mathbf{p}} = \mathbf{F}_{\mathbf{p}} = (\mathbf{p}^{\mathrm{T}}, \tilde{\mathbf{c}}_{\mathbf{p}}^{\mathrm{T}})^{\mathrm{T}}$ which is inherently located at local mode is more reliable. That is, these points can be thought of as distinct features. We hence introduce the feature confidence term into Eq. (6). The final equation is shown in Eq. (8):

$$\begin{aligned}&\mathbf{E}^{t+1}(\mathbf{p}, \mathbf{d}) = \mathbf{E}^{t}(\mathbf{p}, \mathbf{d}) \\ &+\lambda \sum_{\mathbf{s} \in N_{\mathbf{p}}} c_{\mathbf{R}} c_{\mathbf{T}}\, \mathrm{g}_f\left(\left\|\nabla \mathbf{F}^{\mathrm{R}}(\mathbf{s}, \mathbf{p})\right\|\right) \mathrm{g}_f\left(\left\|\nabla \mathbf{F}^{\mathrm{T}}(\mathbf{r}, \mathbf{q})\right\|\right) \nabla \mathbf{E}^{t}(\mathbf{s}, \mathbf{p})\end{aligned} \tag{8}$$

where $c_{\mathbf{R}}$ and $c_{\mathbf{T}}$ are feature confidence of the reference and the target image as shown in Eq. (9) and Eq. (10), respectively.

$$c_{\mathbf{R}} = \mathrm{g}_p\left(\left\|\tilde{\mathbf{s}} - \mathbf{s}\right\|\right) \mathrm{g}_c\left(\left\|\tilde{\mathbf{c}}_{\mathbf{s}}^{\mathrm{R}} - \mathbf{c}_{\mathbf{s}}^{\mathrm{R}}\right\|\right) \tag{9}$$

$$c_{\mathbf{T}} = \mathrm{g}_p\left(\left\|\tilde{\mathbf{r}} - \mathbf{r}\right\|\right) \mathrm{g}_c\left(\left\|\tilde{\mathbf{c}}_{\mathbf{r}}^{\mathrm{T}} - \mathbf{c}_{\mathbf{r}}^{\mathrm{T}}\right\|\right) \tag{10}$$

Finally, disparities are chosen by applying the winner-take-all method (WTA) in 3D cost volume of Eq. (8).

Cost aggregation in the feature space has the following advantages: 1) Similar features are grouped together, which makes costs (or energy) vary smoothly within a same depth level. 2) It enables that different weights are adaptively imposed according to the relative importance of the features, and it provides better discriminative power for different depth levels. 3) A window can move dynamically in constructing the feature space, although it is fixed at cost aggregation step. This dynamic property helps the proposed method propagate the information into neighborhood very well. Consequently, homogeneous regions can be successfully handled with a relatively small window only.

4. EXPERIMENTAL RESUTLS

In this section, we present comparative results of the proposed method with other cost aggregation methods in Middlebury test bed [13]. The proposed method is tested using the same parameters for all the test images. The initial matching costs are calculated using truncated absolute error (TAD) with threshold value, 60. We use the following kernel function in all experiments for the sake of simplicity with bandwidths k .

$$g_k(x) = \exp^{-(x^2/k^2)} \tag{11}$$

The bandwidths (p , c and f) is fixed to 4.0. Time step (λ) is set to 0.5, and the number of iteration (t) is 100. The size of the neighborhood (N) is set to 11. Stereo matching algorithms in [5] [14] are implemented with the same parameters used in the papers. Note that we only compare the proposed method to the cost aggregation with anisotropic diffusion in [14], not the cost aggregation with weighted least square in [14].

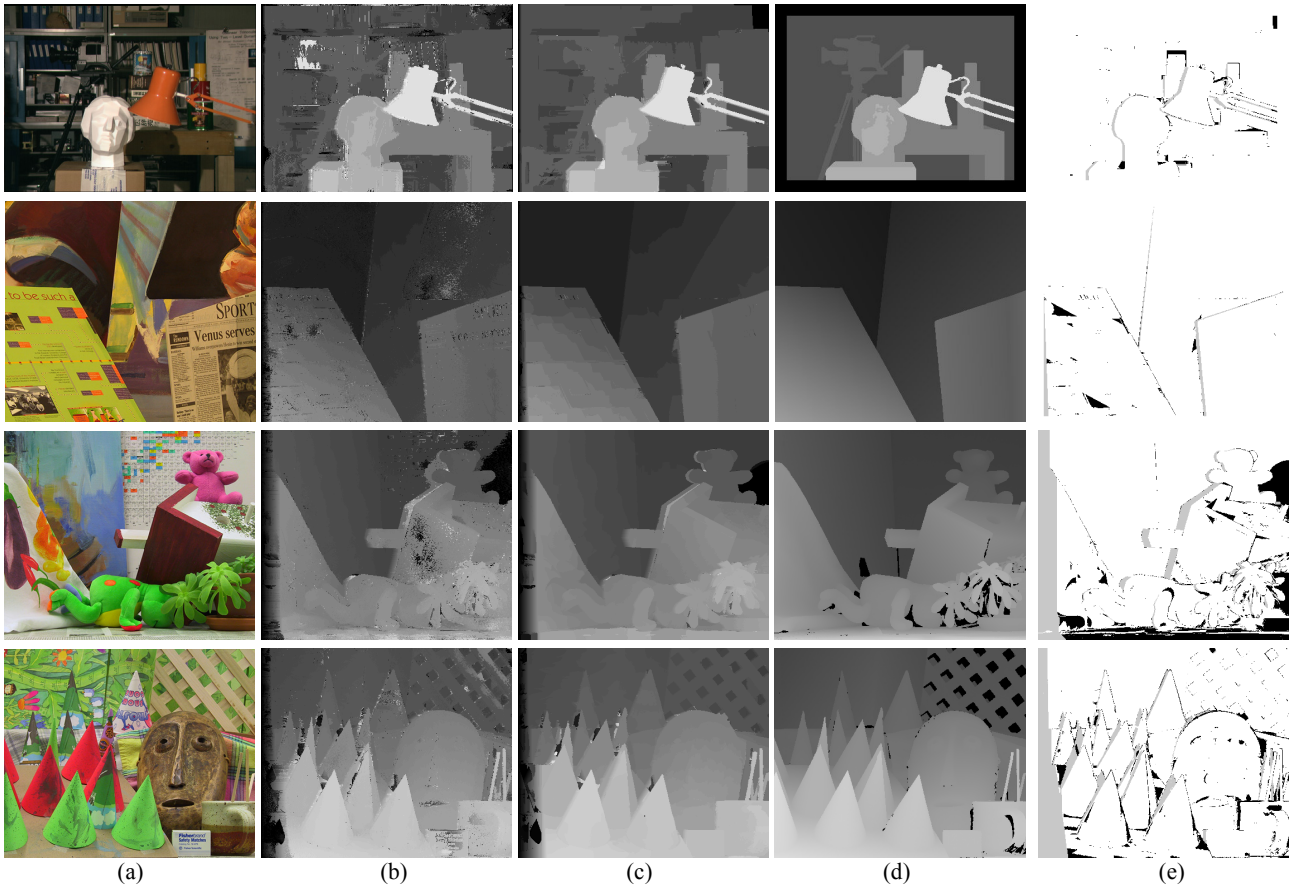
Fig. 1 shows the initial matching and aggregated cost plane when disparity is 0. Since we process stereo matching in the feature space, the proposed method localizes the same depth levels only, i.e., it diffuses pixels inside same depth levels while preventing pixels from being diffused across different depth levels. Therefore, distinct discrimination is observed across the different depth levels. As shown in Fig. 1, the results of the proposed method are superior to these of [5], although relatively small neighborhood is used. Note that the original 'Adaptive weight' results in [5] were not mentioned in the paper, since they used the additional handling for improving the accuracy of the stereo matching. Fig. 2 shows the estimated disparity maps with the proposed method. The occlusion handling methods or post-processing are not used for fair evaluation of the cost aggregation only. The disparity maps are sharp at depth discontinuities, and smooth well enough at homogeneous regions. We use the results of [6] for quantitative comparison with other cost aggregation methods as shown in Table. 1. The symbol '*' indicates the results of Fig. 2. (b). We present the results of NonOcc (all points except for occlude areas) and Disc (only points along depth discontinuities, not including occluded areas) only. We could find that the proposed method obtained the comparable performance with several cost aggregation methods. Especially, the performances of the 'Tsukuba' and 'Venus' are the best among all cost aggregation methods.

5. CONCLUSION

This paper has proposed new cost aggregation method for stereo matching. We have studied how to design a correspondence algorithm without modeling depth cues but disparity. In order to model disparity cues, we have defined a feature space and its similarity. The proposed approach has been formulated via anisotropic diffusion in terms of intra- and inter-coherence. The proposed anisotropic diffusion on feature space can be referred to as a dense feature matching in the viewpoint of cost aggregation,

Table 1 OBJECT EVALUATION FOR THE PROPOSED METHOD

Algorithm	Tsukuba		Venus		Teddy		Cone	
	NonOcc	*Disc*	*NonOcc*	*Disc*	*NonOcc*	*Disc*	*NonOcc*	*Disc*
Segment support	2.28	7.5	1.21	5.88	10.99	22.01	5.42	11.83
Proposed method	1.8	7.27	1.13	4.92	11.2	23.2	5.6	12.4
Adaptive weight **[5]**	4.66	8.25	4.61	13.3	12.7	22.4	5.5	11.9
Adaptive weight* **[5]**	5.4	8.78	6.62	13.2	15.5	25.1	10.8	18.3
VariableWindows	4.1	10.79	10.66	9.94	13.93	25.53	7.24	13.86
Reliability	5.14	18.31	3.86	11.51	16.96	30.62	13.52	21.55
ShiftableWindows	6.53	21.8	6.6	13.54	16.16	30.19	9.55	22.99
Sgementat.based	8.18	18.77	8.06	20.85	15.78	29.66	13.22	24.55



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

Fig. 2. Results for (from top to bottom) 'Tsukuba', 'Venus', 'Teddy' and 'Cone'. (a) reference images, (b) adaptive weight [5], (c) proposed method, (d) ground truth maps, (e) error maps.

so that it can be thought of as the hybrid approach which utilizes the advantages of feature- and area-based approaches together. We also have introduced the feature confidence into the proposed method so that the reliability is adaptively imposed into features. We have verified the performance of the proposed method qualitatively and quantitatively. We will extend this algorithm into occlusion handling. Furthermore, other applications will be also investigated with the proposed diffusion equation.

6. REFERENCES

[1] D. Scharstein and R. Szeliskim, "A Taxonomy and Evaluation of Dense Two-Fame Stereo Correspondence Algorithm," *Int. J. Comput. Vis.*, vol. 47, no. 1-3, pp. 7-42, Apr. 2002.

[2] O. Veksler, "Dense Features for Semi-Dense Stereo Correspondence," *Int. J. Comput. Vis.*, vol. 47, no. 1-3, pp. 247-260, Apr. 2002.

[3] C. V. Jawahar and P. J. Narayananm "Generalised Correlation for Multi-feature Correspondence," *Pattern Recognition*, vol. 35, no. 6, pp. 1303-1313, Jun. 2002.

[4] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 9, pp. 920–932, Sep. 1994.

[5] K. Yoon and I. Kweon, "Adaptive Support-Weight Approach for Correspondence Search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, Apr. 2006.

[6] F. Tombari, S. Mattoccia and L. Stefano, "Classification and Evaluation of Cost Aggregation Methods for Stereo Correspondence," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, Jun, 2008.

[7] B. Julesz, "Depth Perception without Familiarity Cues," *Science*, vol. 145, no. 3630, pp. 356-362, Jul. 1964.

[8] R. v. d. Boomgaard and J. v. d. Weijer, "On the Equivalence of Local-Mode Finding, Robust Estimation and Mean-Shift Analysis as used in Early Vision Tasks," in *Proc. IEEE Int. Conf. on Pattern Recognition*, pp. 927-930. 2002.

[9] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, May. 2002.

[10] K. Prazdny, "Detection of Binocular Disparities," Biol. Cybern., vol. 52, no. 2, pp. 93-99, Jun. 1985.

[11] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629-639, Jul. 1990.

[12] F. Durand and J. Dorsey, "Fast Bilateral Filtering for the Display of High-Dynamic-Range Images," *ACM Trans. Graphics*, vol. 21, no. 3, pp. 257-266, July, 2002.

[13] http://vision.middlebury.edu/stereo

[14] D, Min and K. Sohn, "Cost Aggregation and Occlusion Handling with WLS in Stereo Matching," *IEEE Trans. Image Processing*, vol. 17, no. 8, Aug. 2008.