HIERARCHICAL DEPTH PROCESSING WITH ADAPTIVE SEARCH RANGE AND FUSION

Zucheul Lee and Truong Q. Nguyen

University of California, San Diego
Department of Electrical and Computer Engineering
9500 Gilman Drive, La Jolla, CA 92092 USA

ABSTRACT

In this paper, we present an effective hierarchical depth processing and fusion for large stereo images. We propose the adaptive disparity search range based on the combined local structure from image and initial disparity. The adaptive search range can propagate the smoothness property at the coarse level to the fine level while preserving details and suppressing undesirable errors. The spatial-multiscale total variation method is investigated to enforce the spatial and scaling consistency of multi-scale depth estimates. The experimental results demonstrate that the proposed hierarchical scheme produces high quality and high resolution depth maps by fusing individual multi-scale depth maps, while reducing complexity.

Index Terms— Depth, Hierarchical, Stereo, Fusion

1. INTRODUCTION

Large stereo images are more favorable to customers since they can show realistic, high resolution imagery with a wide field of view. However, high resolution images pose a challenging problem for many computer vision tasks.

Hierarchical (multi-resolution) depth schemes are efficient in dealing with large stereo images by reducing matching ambiguity and computational complexity. However, it is difficult to achieve high accuracy and reduce complexity at the same time. In stereo matching area, local methods [1, 2] based on window matching, and global methods [3, 4, 5, 6] based on belief propagation, have used the hierarchical scheme. In fact, the hierarchical scheme is useful for avoiding local minima in correspondence matching, but it has a limitation such as error propagation from coarse to fine levels. The limitation cannot guarantee that final matching accuracy will be improved. Therefore, most hierarchical methods focus on reducing computational complexity at the expense of accuracy. Two hierarchical algorithms [7] and [8] have reduced disparity search range to speed up. However, the reduced search value set as constant at each pixel may propagate error. The hierarchical stereo method with thin structure [9] emphasizes the importance of search range shifted by the disparity of the corresponding coarse point. However, there is no discussion on how to find the optimal disparity search range. All of these hierarchical methods process small images such as the Middlebury datasets [10]. We are dealing with stereo images about twenty times larger than that of the Middlebury datasets. As image scale increases, so does the importance of mitigating the limitation of hierarchical disparity methods. Another challenging factor arising in the high resolution image will be discussed in Section 2.

Generally, real images and video frames are susceptible to various noise factors such as camera and illumination distortion. Therefore, a consistency function is required in the disparity estimation process. For example, temporal consistency should be considered in video frame disparity estimation. Similarly, scaling consistency needs to be taken into account in multi-scale disparity processing of large stereo images. To the best of our knowledge, the disparity scaling consistency issue has not been studied.

The contribution of this paper is the adaptive pixel-wise disparity search range, which is based on the local structure of image and initial disparity map. The optimal adaptive search range can propagate smoothness in the homogeneous areas and help to recover the initial disparity error. We investigate the spatial-multiscale total variation (TV) to enforce both spatial and multi-scaling consistency. Finally, the adaptive search range and spatial-multiscale TV play a role in fusing multi-scale disparity maps by guiding estimation and combining the complementary information, respectively. We quantitatively evaluate the effectiveness and advantage of the proposed method and then demonstrate that it achieves high-quality depth map on large real-world panoramic views.

The rest of the paper is organized as follows. The problem that we are solving is described in Section 2. The details of the proposed method are presented in Section 3. Section 4 shows experimental results and discusses their significance. We conclude with some remarks in Section 5.

2. PROBLEM

It is challenging to obtain a high quality and high resolution depth map on large stereo panoramas $(8,192 \times 4,096)$. For large image processing, we take into account a hierarchical framework and a partitioning-stitching approach.

Fig. 1 illustrates the problem addressed in this paper. At the coarsest level in Fig. 1(b), the prominent features and overall information appear in the smooth form, but sharp edges and details are lost as shown in the red boxes. In contrast, most details and edges are preserved at the finest level in Fig. 1(c), but a lot of errors are present in the low and high-textured areas as shown in the green boxes. It can be observed that fine details are too small to detect at the coarse level. As the resolution level increases, size of ambiguous areas tends to increase. For example, the structure information in the low-textured area becomes more ambiguous and the structure in the high-textured area tends to look like a repetitive pattern as image resolution increases. The difficult problem is how to fuse only the beneficial characteristics at the coarse and finer level while suppressing undesirable errors.

This work is supported in part by NSF grant CCF-1065305, by Intel/CISCO under the VAWN program and by the Technology Development Program for Commercializing System Semiconductor funded by the Ministry of Knowledge Economy (MKE, Korea). (No. 10041126, Title: International Collaborative R&BD Project for System Semiconductor).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

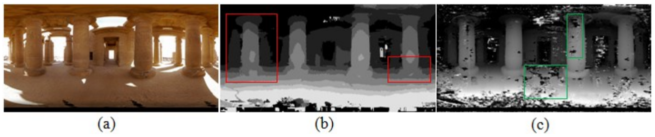

Fig. 1. Depth maps in hierarchical framework. (a) Left panorama. (b) Coarsest depth map. (c) Finest depth map.

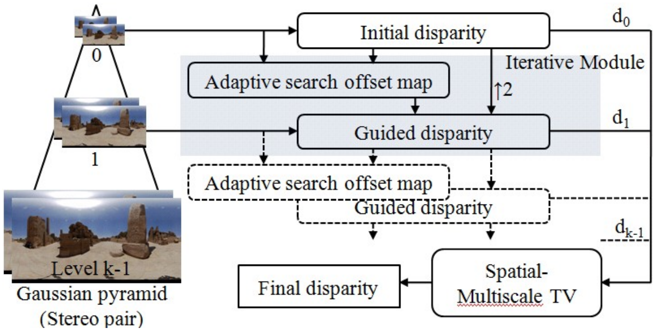

Fig. 2. System diagram.

3. PROPOSED METHOD

3.1. Proposed system framework

Fig. 2 illustrates the proposed multi-stage framework, which consists of three main blocks: adaptive search offset map, guided disparity estimation, and spatial-multiscale TV. First, we build a stereo pair of Gaussian pyramids with k levels, and then estimate an initial disparity, using the local method with 3-moded census (LM3C) [11]. The adaptive search offset map is constructed by using image intensity and the initial depth. The shaded blue box in Fig. 2 indicates the iterative portion of the algorithm. The final disparity map is obtained by applying the spatial-multiscale TV to multi-resolution disparity maps $(d_0, d_1, .., d_{k-1})$.

3.2. Structure tensor-based Adaptive search range

For the next level estimation, we can utilize given initial priors to adaptively minimize the disparity search range without loss of accuracy. The reduced disparity search range (R_k) for the next level k can be expressed as

$$2d_{k-1} - \Delta d \leq R_k < 2d_{k-1} + \Delta d \tag{1}$$

where d_{k-1} is the disparity estimate at level $k-1$, and Δd is the adaptive disparity search offset.

Fig. 3 illustrates the disparity search offset (Δd) in the full search range (L). It is crucial for hierarchical scheme to properly choose Δd since a small Δd at a certain point can increase speed and matching disambiguation, while a larger Δd at a different point can better resolve complicated object edges [9]. Hence, the estimation quality and speed directly depend on d_0 and Δd.

Structure tensor is known as the second-moment matrix, which is calculated by the summation of the outer product components of the local gradient from a neighborhood [12]. We consider the 2D structure tensor for detecting two-dimensional features of an image:

$$\mathbf{S} = \sum \nabla \mathbf{I} \nabla \mathbf{I}^T = \begin{pmatrix} \sum \mathbf{I}_x^2 & \sum \mathbf{I}_x \mathbf{I}_y \\ \sum \mathbf{I}_x \mathbf{I}_y & \sum \mathbf{I}_y^2 \end{pmatrix}. \tag{2}$$

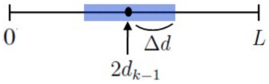

Fig. 3. Illustration of disparity search range.

$\mathbf{S}$ is a symmetric positive semi-definite matrix, which possesses two non-negative eigenvalues λ_{max}, λ_{min}. $\mathbf{I}_x$ is the image gradient along x direction. There are three distinct cases for the relative values of these two eigenvalues [13]:

- $\lambda_{max} \approx \lambda_{min} \approx 0$: low-textured area with almost no structure
- $\lambda_{max} \gg 0, \lambda_{min} \gg 0$: high-textured area with ambiguous orientation
- $\lambda_{max} \gg 0, \lambda_{min} \approx 0$: one dominant orientation.

For the local structure acquisition, both image intensity and the initial disparity map are exploited as a prior since they may reveal different but complementary structure information. The following observations are made:

- Matching ambiguities occur in both low and high-textured areas
- Disparity jumps occur at real disparity edges, which generally match the corresponding image edges

The first observation indicates that Δd should be small enough to reduce the ambiguity at the next high level. Small Δd in turn propagates the desirable smoothness property of the coarse disparity estimates. The second one represents that Δd should be large enough to detect the big disparity changes and in turn recovers the initial disparity errors. We can define a function of eigenvalues of the matrix $\mathbf{S}$ satisfying the aforementioned observations as

$$i(\lambda_{max}, \lambda_{min}) = \frac{\lambda_{min} + \epsilon}{\lambda_{max} + \epsilon} \tag{3}$$

where ϵ is used for the robust function near zero eigenvalue. Too small value of ϵ makes the function sensitive to eigenvalues. A reasonable value of ϵ is 0.1 empirically.

The two eigenvalues represent a scaling term along each orthogonal direction in the ellipsoidal representation of the matrix $\mathbf{S}$. Let $\mathbf{S}_I$ and $\mathbf{S}_D$ be the matrix from the image intensity and initial disparity, respectively. We have four eigenvalues from two matrices: two maximum eigenvalues $(\lambda_{Imax}, \lambda_{Dmax})$ and two minimum eigenvalues $(\lambda_{Imin}, \lambda_{Dmin})$. Typically, disparity map shows the piecewise constant characteristic. Therefore, zero eigenvalues representing no local structure $(\lambda_{Dmax} \approx \lambda_{Dmin} \approx 0)$ are observed in most areas except at disparity edges. At the disparity edges, it shows the same dominant orientation as that of the image. This implies that corresponding eigenvectors from $\mathbf{S}_I$ and $\mathbf{S}_D$ have the same direction. Hence, the eigenvalues can be linearly combined as $\lambda_{max} = \lambda_{Imax} + \lambda_{Dmax}$ and $\lambda_{min} = \lambda_{Imin} + \lambda_{Dmin}$. Note that the eigenvalue computation is performed on the normalized image intensity and disparity map for direct linear combination. An exponential function based on the combined eigenvalues at the pixel p can be defined as

$$s(p) = e^{-i_p(\lambda_{max}, \lambda_{min})}. \tag{4}$$

For simplicity, we define this function as a local edge strength function, which produces a high value along the edges ($\lambda_{max} \gg$

0, $\lambda_{min} \approx 0$) and a low value on the low ($\lambda_{max} \approx \lambda_{min} \approx 0$) and high-textured area ($\lambda_{max} \gg 0, \lambda_{min} \gg 0$).

Finally, the search offset Δd is adaptively determined according to four eigenvalues of two structure tensors ($\mathbf{S}_I$ and $\mathbf{S}_D$) as

$$\Delta d = [s(p) \times \frac{L}{2}] \tag{5}$$

where $[\cdot]$ represents nearest integer operator, and L is the full disparity search range as shown in Fig. 3.

The adaptive search range calculated using (1) fuses the initial information to the high-scale direction by guiding the next level estimation.

3.3. Multi-scale consistency and fusion

In the hierarchical disparity scheme, estimations at different scale show different results, especially for real-world images. It can be called scale dimensional inconsistency due to the frequent upsampling/downsampling process. To reduce the scaling inconsistency as well as the spatial inconsistency, we use spatial-multiscale TV algorithm. It is based on the augmented Lagrangian method for total variation image restoration presented in [14], which enforces the spatial-temporal consistency for video. We adapt it for scale consistency where disparity estimates at several scales are used. We treat a sequence of multi-scale disparity maps as a space-scale volume: a three dimensional function $f(x, y, s)$ with the spatial coordinate (x, y) and the scale dimensional coordinate s. The multi-scale disparity maps are scaled to the same size so that it can be regarded as one volume. To alleviate spatial and scale dimensional noise while preserving sharp boundaries, we solve the following regularized l_1 minimization problem:

$$\underset{\mathbf{f}}{\text{minimize}} \quad \mu||\mathbf{f} - \mathbf{g}||_1 + ||\mathbf{Df}||_2 \tag{6}$$

where the vector **f** is the unknown disparity map, the vector **g** is the multi-scale disparity map, and the vector $\mathbf{D} = [\beta_x \mathbf{D}_x^T, \beta_y \mathbf{D}_y^T, \beta_s \mathbf{D}_s^T]^T$ denotes the forward difference operators along the horizontal, vertical, and scaling direction. The parameter μ is the regularization constant that controls the relative emphasis of the objective and regularization terms. The parameters $(\beta_x, \beta_y, \beta_s)$ also control the relative emphasis of the spatial and scale dimensional terms. The reader can refer to [14] for details of this algorithm.

Through the spatial-multiscale TV, we can fuse different scale resolutions while maintaining spatial-scaling consistency and preserving edges. It can enhance features which are not visible in an individual disparity map.

4. EXPERIMENTS AND RESULTS

4.1. Performance results on the Middlebury datasets

We apply 2-level hierarchical scheme to the Middlebury datasets in order to evaluate quantitatively. All parameters are fixed throughout the experiment. Fig. 4 shows the disparity results. Fig. 4(b) depicts initial disparity maps at the level 0 in the proposed multi-resolution scheme. Fig. 4(c) shows the local edge strength function $s(p)$ defined in (4). Fig. 4(d) depicts the final disparity map guided by the adaptive search offset map. Fig. 4(e) shows the single-resolution disparity maps. As shown in Fig. 4, the smoothness in the coarse disparity map is propagated into the next level by the adaptive search offset that should be small enough, while the single-resolution scheme yields errors due to the matching ambiguities in these repetitive and homogeneous regions (red boxes). In the green boxes, the errors and lost details present in the initial disparity map are recovered without the error propagation, with the aid of the adaptive search offset that should be large enough.

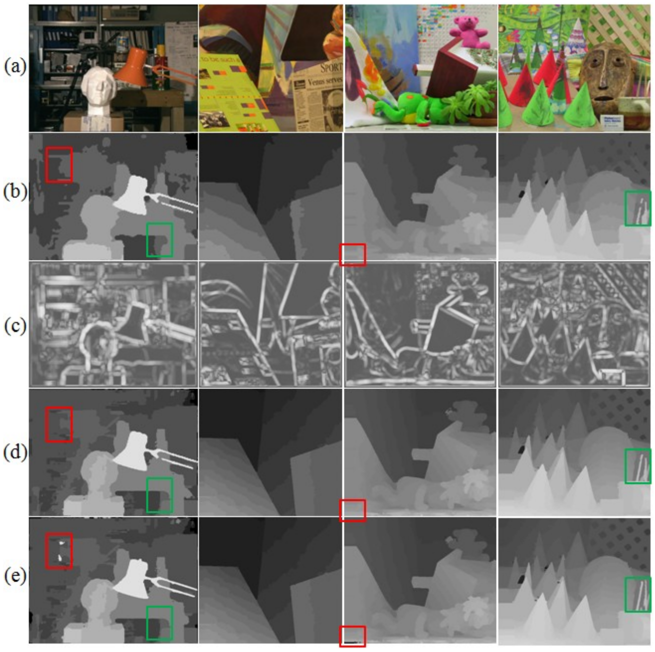

Fig. 4. Middlebury results. (a) Left image. (b) Initial disparity. (c) Local edge strength map $s(p)$. (d) Final guided disparity map. (e) Single-resolution disparity map.

Table 1 shows the performance comparison of three schemes: single-resolution, multi-resolution with fixed search range, and proposed method. The single-resolution method is performed with full search range, while the hierarchical one is conducted with reduced search range (fixed and adaptive). The hierarchical scheme with fixed search range performs worst. It demonstrates that a proper choice for search range is crucial for the quality of hierarchical schemes. Table 2 shows how robust the proposed hierarchical scheme is to other initial disparity algorithms. For experiment, we select two well-known algorithms: CostFilter [15] and local adaptive support weight (LASW) [16]. These experiments are performed without filling process, since the left-right filling is not clearly discussed in [16]. Table 1 and 2 demonstrate that the proposed hierarchical scheme achieves similar or slightly better performance while reducing complexity comparing to the single-resolution scheme. Table 2 also verifies that the proposed hierarchical scheme does not depend on initial results. Note that the proposed hierarchical scheme will be more effective as image resolution increases, as demonstrated in Section 4.2.

The complexity of the single-resolution disparity estimation [11] is $O(MNR)$, where M and N are the size of image and support window, respectively, and R is the disparity search range. The complexity of the 2-level hierarchical disparity estimation decreases to $\frac{5}{8}O(MNR)$ if the search range is reduced by 50%. As a result, the proposed hierarchical scheme is able to have approximately complexity gain of $\frac{3}{8}O(MNR)$. It is possible to further decrease the complexity by applying higher scale pyramid scheme. The hierarchical scheme can also reduce memory complexity. The single-resolution scheme has to build the entire Disparity Space Image

Table 1. Performance comparison (bad pixel rates in non-occlusion area (nonocc) with threshold of 1 and search range percentage denoting the average usage out of the full search range (L)

Dataset	Single-res.		multi-res. (fixed)		Proposed (adapt.)	
	error (%)	search	error(%)	search	error (%)	search
Tsukuba	**2.10**	100%	2.22	50%	2.20	51%
Venus	0.12	100%	0.19	50%	**0.11**	46%
Teddy	5.46	100%	5.41	50%	**5.35**	47%
Cones	**2.12**	100%	2.42	50%	2.30	48%

Table 2. Robustness to other local disparity methods

Dataset	CostFilter [15]			LASW [16]		
	Single-res.	Proposed.		Single-res.	Proposed.	
	error (%)	error (%)	search	error (%)	error (%)	search
Tsukuba	**2.52**	2.76	51%	2.94	**2.89**	60%
Venus	2.04	**1.89**	46%	3.98	**3.46**	53%
Teddy	8.47	**8.42**	48%	14.3	**14.0**	56%
Cones	**3.62**	3.91	49%	**9.43**	9.45	58%

(DSI) [17] that is a function defined over a discretized version of disparity space (x, y, d), while the hierarchical scheme constructs only some part of DSI. There is an additional step for the proposed scheme: the adaptive search offset (Δd) construction, compared to the conventional scheme. The additional computation load is negligible. For Tsukuba image, it takes about $12s$ to complete disparity estimation as in [11] while it takes about $0.1s$ to compute Δd. Currently, the entire algorithm does not operate in real-time. However, the local disparity estimation used in the proposed scheme is suitable for real-time processing using a Graphics Processing Unit (GPU). The parallel computation would decrease the final processing time significantly. Note that the consistency enforcement is not necessary for Middlebury datasets ideally acquired in the laboratory.

4.2. Real-world panoramic results and fusion effects

We use 4-level pyramid for large panoramic views $(8, 192 \times 4, 096)$. The coarsest image size is 1024×512 which becomes a basic size for partitioning. The partitioned disparity results will be stitched. Throughout the experiment, the spatial-multiscale parameters are set to constant values: $\mu = 1$ and $(\beta_x, \beta y, \beta_s) = (1, 1, 2.5)$.

Fig. 5 illustrates the large panoramic disparity results. Fig. 5(b) shows the disparity maps of the RealtimeBP [4] which is a well known hierarchical global method. Unfortunately, the advanced version [5] of the RealtimeBP is not available. Fig. 5(c) depicts the final results of the conventional hierarchical scheme using the same disparity method as in the proposed method. Fig. 5(f) shows the final results of the proposed method. The realtimeBP produces very noisy disparity maps which may result from the propagation failure while globally optimizing. The conventional scheme still produces a lot of errors along object boundaries and staircase errors on the surface. On the other hand, the proposed scheme shows the best quality of disparity map compared to the others. Fig. 5(d) and (e) depict the coarsest and finest maps of the proposed method, respectively. For the real-world image, the spatial-multiscale TV is applied to enforce scaling consistency. As shown in Fig. 5(f), combining the multi-scale maps can enhance features as well as produce the smooth disparity surface. As a result, it is verified that the two proposed algorithms (the adaptive search offset and spatial-multiscale TV) contribute to fusing the advantages taken from both coarse and fine level estimate.

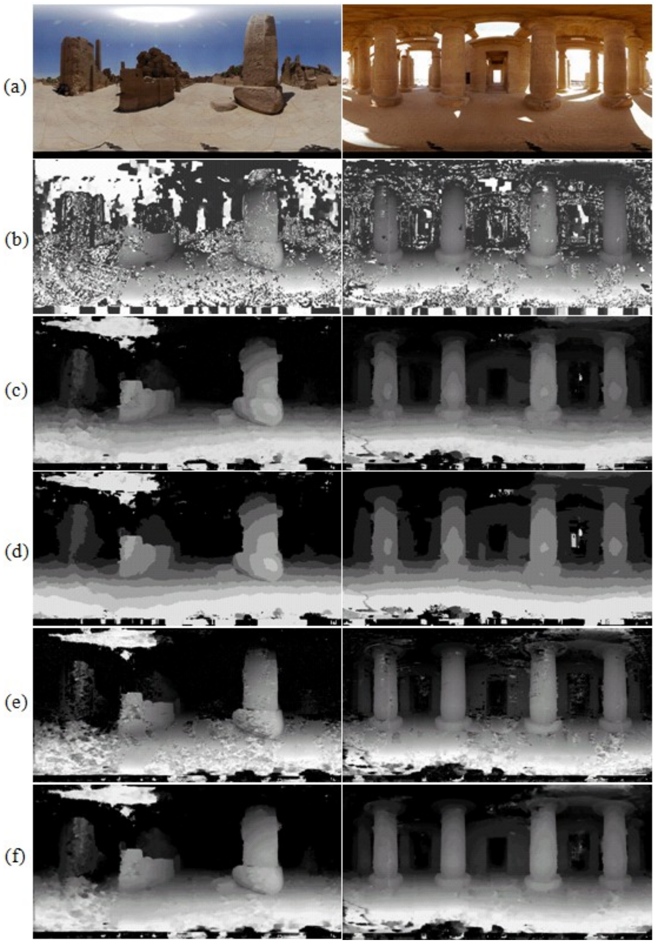

Fig. 5. Disparity results on large real-world panoramic images. (a) Left panorama. (b) RealtimeBP [4]. (c) Conventional hierarchical scheme. (d) Coarsest map in the proposed scheme. (e) Finest map. (f) Final disparity with spatial-multiscale TV.

5. CONCLUSION

It is challenging to process large size images such as stereo panoramas and moreover, achieve accuracy and reduce complexity at the same time. To obtain reliable depth maps from large scale images, we propose an adaptive disparity search range, which is based on the combined eigenvalues of structure tensor matrix of image intensity and initial disparity. To enforce the spatial and scaling consistency, we suggest the spatial-multiscale TV method. Simulation results verify that the proposed hierarchical scheme fuses the multi-scale depth maps effectively and also produces a high-quality depth maps.

Acknowledgment

The authors would like to thank Jurgen Schulze and Tom Defanti for providing the large panoramic images.

6. REFERENCES

[1] M. Accame, F.G.B. De Natale, and D.D. Giusto, “Hierarchical block matching for disparity estimation in stereo sequences,” in *Proc. IEEE ICIP*, 1995, vol. 2, pp. 374–377.

[2] T. Kudo, K. Shirai, and M. Ikehara, “Hierarchical stereo matching via color segmentation,” in *Proc. 12th Digit. Signal Process. Workshop - 4th Signal Process. Education Workshop*, 2006, pp. 522–525.

[3] P. Felzenszwalb and D. Huttenlocher, “Efficient belief propagation for early vision,” in *Proc. IEEE CVPR*, 2004, pp. 261–268.

[4] Q. Yang, L. Wang, R. Yang, S. Wang, M. Liao, and D. Nister, “Real-time global stereo matching using hierarchical belief propagation,” in *Proc. British Machine Vision Conference (BMVC)*, 2006.

[5] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, “Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 492–504, 2009.

[6] S. Grauer-Gray and C. Kambhamettu, “Hierarchical belief propagation to reduce search space using cuda for stereo and motion estimation,” in *Proc. Workshop on Applications of Computer Vision (WACV)*, 2009, pp. 1–8.

[7] M. Sizintsev and R. Wildes, “Efficient stereo with accurate 3-d boundaries,” in *Proc. British Machine Vision Conference (BMVC)*, 2006.

[8] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool, “A hierarchical stereo algorithm using dynamic programming,” in *Proc. IEEE Stereo and Multi-Baseline Vision (SMBV)*, 2001, pp. 166–174.

[9] M. Sizintsev, “Hierarchical stereo with thin structures and transparency,” in *Proc. Computer and Robot Vision (CRV)*, may 2008, pp. 97 –104.

[10] D. Scharstein and R. Szelisk, “Middlebury stereo evaluation version 2,” http://vision.middlebury.edu/stereo/eva, 2010.

[11] Z. Lee, J. Juang, and T.Q. Nguyen., “Local disparity estimation with three-moded cross census and advanced support weight,” *Accepted to IEEE Trans. Multimedia*, 2013.

[12] B. Jahne, “Digital image processing,” Springer, Jan. 2006.

[13] T. Brox, F. Boomgaard, F. Lauze, J. Weijer, F. Weickert, and P. Kornprobst, “Adaptive structure tensors and their applications,” in *Visualization and Processing of Tensor Fields*. Springer Berlin Heidelberg, 2006, pp. 17–47.

[14] S.H. Chan, R. Khoshabeh, K.B. Gibson, P.E. Gill, and T.Q. Nguyen, “An augmented lagrangian method for total variation video restoration,” *IEEE Trans. Image Processing*, vol. 20, no. 11, pp. 3097 –3111, nov. 2011.

[15] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, “Fast cost-volume filtering for visual correspondence and beyond,” in *Proc. IEEE CVPR*, Jun. 2011, pp. 3017–3024.

[16] K. Yoon and I.S. Kweon, “Adaptive support weight approach for correspondence search,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[17] D. Scharstein and R. Szeliski, “Texonomy and evaluation of dense two-frame stereo correspondence algorithms,” *International Journal of Computer Vision 47*, , no. 7-42, 2002.