Adaptive Scale Selection for Hierarchical Stereo

Yi-Hung Jen
dedanann@cs.unc.edu
Enrique Dunn
dunn@cs.unc.edu
Pierre Fite-Georgel
pierre.georgel@gmail.com
Jan-Michael Frahm
jmf@cs.unc.edu

3D Computer Vision Group
Computer Science Department
University of North Carolina
Chapel Hill, NC, USA

Abstract

Hierarchical stereo provides an efficient coarse-to-fine mechanism for disparity map estimation. However, common drawbacks of such an approach include the loss of high frequency structures not observable at coarse scale levels, as well as the unrecoverable propagation of erroneous disparity estimates through the scale space. This paper presents an adaptive scale selection mechanism to determine a suitable resolution level from which to begin the hierarchical depth estimation process for each pixel. The proposed scale selection mechanism allows us to robustly implement variable cost aggregation in order to reduce the variability of the photo-consistency measure across scale space. We also incorporate a weighted shiftable window mechanism to enable error correction during coarse-to-fine depth refinement. Experiments illustrate the effectiveness of our approach in terms of disparity accuracy, while attaining a computational efficiency compromise between full resolution and hierarchical disparity map estimation.

1 Introduction

Stereo disparity map estimation entails determining the set of image-wide pixel-level correspondences among a pair of input images. Canonical hierarchical stereo approaches sequentially process different levels of the input imagery scale space representation, using previously computed disparity estimates to constrain the search range at the current level. Accordingly, hierarchical stereo offers a computationally efficient framework for coarse-to-fine depthmap estimation and refinement. However, by virtue of the sampling theorem, disparity estimates for fine-grain scene structures (i.e. originally imaged at spatial frequencies greater than half the sampling frequency of the coarsest level of the scale space) may not be properly estimated. Also, the image smoothing and sub-sampling inherent to discrete scale space generation may induce erroneous disparity estimates propagated along the scale space during the coarse-to-fine refinement process. Accordingly, the two main factors hindering the correctness of hierarchical stereo approaches are *1)* the variability at different

BMVC 2011 http://dx.doi.org/10.5244/C.25.95

scales of the photo-consistency criterion function used for template matching, and *2)* systematic attempts to perform template matching in regions of the scale space where the image signal is insufficiently sampled. As a result, block matching hierarchical approaches provide poor boundary localization for coarse scene structure, while systematically suppressing (i.e. smoothing) fine structure details.

In this work we address photo-consistency measure variability and signal undersampling by developing an adaptive hierarchical stereo approach where the coarsest scale at which to evaluate a given pixel's disparity is bounded by the effective resolution of its neighboring texture. To further reduce matching variability across scales, our template photo-consistency measure is based on variable cost aggregation. As our experiments will show, scale selection is critical to robustly implement variable cost aggregation in hierarchical stereo. Finally, local disparity estimates are propagated through a *weighted shiftable window* mechanism to mitigate error propagation in the coarse-to-fine refinement process. The integration of these algorithmic elements yields a hierarchical stereo framework which outperforms either of these components independently in terms of accuracy, while offering an efficiency compromise between hierarchical stereo and full resolution processing.

2 Background and Related work

We restrict our literature review to only encompass highly related stereo works, and refer the reader seeking further comparisons and classification to [10]. In its most basic form, stereo photo-consistency template matching can be straightforwardly implemented through image block comparisons. Such an approach assumes the image block under analysis corresponds to a single frontoparallel surface. The degree to which the observed scene violates these assumptions will cause systematic abberations in the obtained disparity map. Examples of works addressing the analysis and mitigation of arbitrarily oriented surfaces in stereo disparity estimation can be found in [2, 9]. Violations of the single surface assumption typically lead to poor disparity boundary localization due to surface overextension. To overcome this issue, different researchers have proposed the use of variable cost aggregation within the photo-consistency function [3, 14]. Among these approaches, adaptive weighing of each pixel contribution enables texture driven pixel segregation within a single image block [7, 15]. The underlying assumption is that within an image block there exists a strong correlation between a pixel's intensity (color) and the 3D surface to which it belongs. While this assumption may not hold in general (i.e. highly textured smooth surfaces), empirical results indicate the effectiveness of this approach at estimating fine structures. However, larger comparison windows are preferable in order to achieve statistically robust performance. Accordingly, the computational overhead involved in variable cost aggregation may be significant, requiring an effective search strategy in order to balance the improved qualitative results with overall efficiency.

Hierarchical stereo offers an efficient disparity search paradigm aimed at reducing the number of template matching comparisons required for local stereo approaches. These *Coarse-to-Fine* (CTF) approaches perform full disparity range search at reduced resolutions in the scale space and refine such estimates across scale space through local search. Accordingly, they provide significant computational savings when compared with exhaustive disparity search at full resolution. However, the limitations of block-based matching are exacerbated while operating in the discrete scale space, leading to both heavy distortions near surface boundaries as well as loss of fine detail structures. The work of Sizintsev [12] imple-

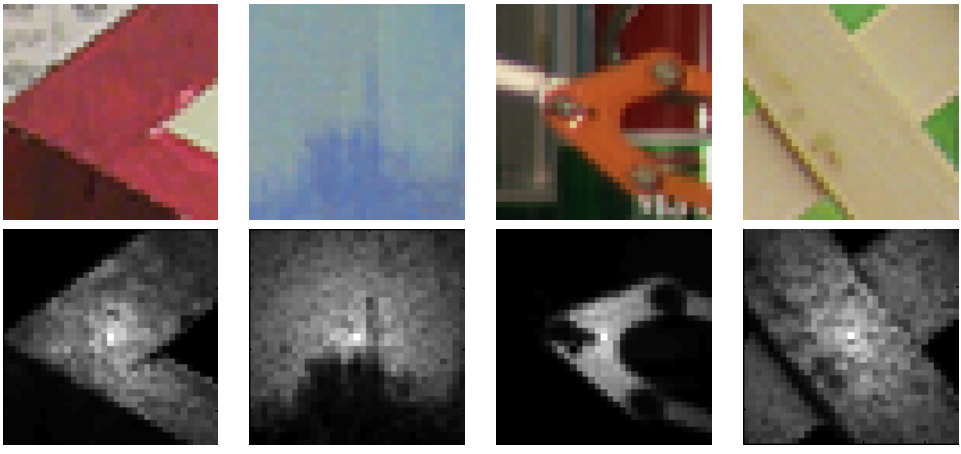

Figure 1: Variable cost aggregation. Top row: Color image patches. Bottom row: Corresponding pixel weights as defined by Equation 2, brighter color indicates higher values.

ments a hierarchical stereo approach, which incorporates adaptive matching window support across scales, in a manner similar to shiftable windows [1]. While such an approach improves border localization, it does not provide a mechanism for fine structure recovery. This issue was later addressed in [8] by matching with a generalization of the Laplacian pyramid that explicitly encodes the energy magnitude component of the band-passed images. Such representation encodes fine scale details in low resolution images, which allows for accurate recovery of thin structures during CTF processing.

Our work integrates variable cost aggregation into the CTF framework as a means to simultaneously improve both boundary localization and structure granularity under a single framework. As our results illustrate, in order for adaptive pixel weighing to be effective, the compared image blocks must contain sufficient texture to enable pixel segregation. Moreover, it will be this texture variability which will enable reliable template matching. Kanade and Okutomo [4] developed an iterative approach for adaptive window selection based on uncertainty reduction. In that approach, both the size and shape of the window were refined successively as a function of both intensity and depth variation within the matching window. Our work addresses a similar issue, but does so within the CTF framework by making pixel-wise decisions regarding in which level of the scale space to perform disparity estimate initialization (i.e. full disparity range evaluation).

3 Approach

We present a scale adaptive hierarchical stereo approach aimed at *1)* recovering detailed structure lost during successive convolution and sub-sampling in the generation of the discrete scale space representation, and *2)* resolving ambiguity in the initialization of succeeding levels from the upsampled disparity image.

3.1 Variable cost aggregation

On observing the progressive loss of high frequency content in a Gaussian image pyramid, edges that are part of surface contours become less distinct. Accordingly, these edges become indiscriminate from the surrounding surface and photo-consistency based matches obtained from coarse resolution levels may not reflect the true disparity of the surface. If one wishes to retain the computational efficiency of CTF stereo approaches, the challenge is to find the right level in the image pyramid at which the surface is salient. Without prior knowledge of the underlying surface shape, it is reasonable to infer surface shape from color segmentation. The work presented in [15, 16] suggests that surface priors may be effectively modeled locally by color similarity and spatial proximity. The goal of computing this surface membership prior is to segregate and prioritize pixels within a given pairwise template matching function. In practice, different relative weights are assigned to each pixel in the template as follows:

$$\begin{aligned} W_c(x,y,t) &= exp^{-\frac{\sqrt{\sum_{c=1}^{ch}(I_t(xc,yc,c)-I_t(x,y,c))^2}}{k_c}} \\ W_d(x,y) &= exp^{-\frac{\sqrt{(x-xc)^2+(y-yc)^2}}{k_d}} \\ W_f(x,y,t) &= W_c(x,y,t) \times W_d(x,y) \end{aligned} \tag{1}$$

k_c and k_d are parameters weighing between color and proximity respectively and c is the channel index in the image. The cost aggregation then is rewritten as:

$$C(xc,yc,d) = \sum_{(x,y)\in w(xc,yc)} W_f(x,y,ref) \times W_f(x+d,y+d,target) \times \|I_{ref} - I_{target}\| \tag{2}$$

Under this formulation, weighting the contribution by color distance conforms to the assumption that objects lying on the same surface are similar in color; whereas weighting by proximity models the correlation of disparity from neighboring pixels to the center pixel on the matching template. Figure 1 shows some weight images and illustrates how pixels belonging to the same surface are given higher correlation. The resulting weight transformation performs a similarity-based scalarization of the color space in the neighborhood of a given reference pixel. Moreover, given that this transformation has a very close resemblance to Bilateral filtering used in edge preserving image noise reduction [13], we use its output not only to evaluate photo-consistency but also to perform our scale space analysis.

3.2 Adaptive scale selection

We note that our method of extracting surface priors through the use of color segmentation presents the disadvantage that surface priors become less consistent with the true surface as resolution decreases. This is tightly coupled to the loss of signal variation inherent in scale space traversal. Accordingly, to find the most favorable scale to perform disparity search, we quantify the saliency of the surface prior at each level and search for an extremum in the scale space. We quantify this saliency by convolving the surface prior image with a Laplacian of Gaussian (LoG) kernel. The shape of the template here fits the purpose well, as we want to gauge when a structure is no longer distinguishable at a given resolution. Moreover, saliency is maximal when the contour of the surface prior fits the template. Figure 2 shows an example using scale-space representation of an surface prior image (i.e. variable cost weights computed as described in section 3.1) and illustrates how the shape of the LoG

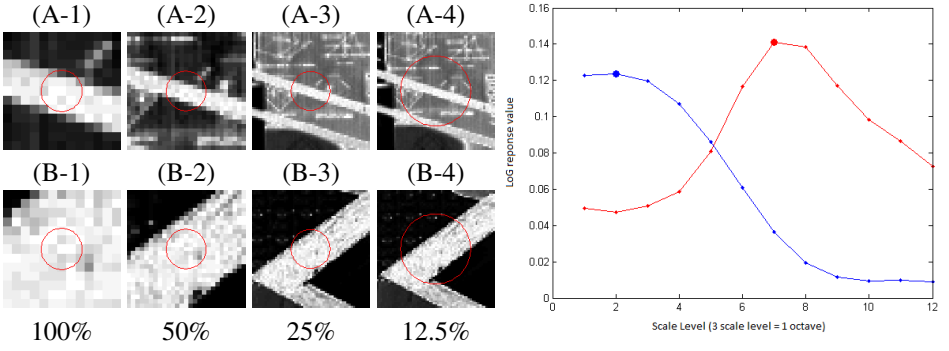

Figure 2: Scale selection examples. At left: A pair of compact LoG kernels (shown in red) are applied to weight image regions surrounding a pixel of interest. For illustration purposes, the image is fixed but the LoG kernel is increasing. At right: response curve from all levels in the pyramid at the sampled locations. (A-1 to 4) response values are drawn in red, while (B-1 to 4) in blue. Markers indicate the selected scale corresponding to the maximum response value.

kernel that reported the maximum in scale space is consistent with the shape of the sampled surface prior. Figure 2 also shows the response curve for the two sampled locations. Notice that the marked level's octave and the shape of the LoG is in accordance. The motivation for performing LoG filtering on the pixel weight image instead of directly filtering the input color images is to achieve a strong coupling between our photo-consistency measure and our scale selection mechanism. Our experiments will illustrate how the two alternative scale maps (color-based vs. weight-based) provide different scale distributions and performance benchmarks.

The surface priors computed through adaptive weighting are locally defined since the weight of a given pixel depends on the image template in which it is contained. Moreover, while DoG pyramids [6] require that the pixel value at (x,y) be homogeneous to each instance of $(x,y) \in w(xc,yc)$, our measure of surface saliency has to be computed by convolving with a LoG filter separately. Surface saliency is computed by fixing a small LoG filter applied to a scale space representation with 3 octaves at each pyramid level, in order to avoid losing structure and reduce computation. The maximum in the scale space is identified for each pixel and the corresponding octave is stored in a scale map. The entire process mimics that of a blob detection shown in [5]. An example of a scale map is presented in Figure 3. Notice we were unable to select a coarser resolution on the back wall as the depth prior provided by Equation 2 did not fit the observed scene, e.g. a large surface containing small patches of solid color, black writings and lines. Accordingly, in these cases we will pay the computational penalty of evaluating all such pixels at full resolution, but not suffer any loss in accuracy.

3.3 Variable cost shiftable window

For each pixel, we perform correspondence search at the resolution octave designated by the scale map. The matching disparity value is then upsampled and scaled to the next octave. In order to reduce boundary localization errors, we adopt a shiftable window method in the spirit of the one presented in [12]. We first upsample coarse depth estimates and for each

Figure 3: At Left: Source image. At right: Computed scale map. The brighter color in the scale corresponds to higher pyramid levels (i.e. coarser resolution).

pixel (xc, yc), we define a set $w(xc, yc)$ of neighboring pixels. We then weigh the photo-consistency value for each pixel in $w(xc, yc)$ according to Equation 2 and select the disparity offset corresponding to the minimum *weighted* photo-consistency cost from this set. Our approach differs from [12] as we don't consider each upsampled disparity value equally. Instead, we further restrict the propagation of erroneous corrections by enforcing texture coherency (i.e. surface priors) into our shiftable window edge localization mechanism. The computational cost of this procedure is negligible as the shiftable window weight values already need to be computed for photo-consistency estimation at the current level.

4 Experiments and Implementation

We compare our proposed scale selective CTF (SCTF) method with other state-of-the-art cost aggregation approaches such as Fast Bilateral Stereo (FBS) [7], Adaptive Weighting (AW) [17] and Accurate Boundary CTF (ACTF) [12]. Our evaluation considers three variants of our proposed method: SCTF performs scale selection analysis on the template weight image, SCFT* performs scale selection on the input color images, while SCFT† bypasses scale selection by starting each pixel at the coarsest level and only incorporates our modified shiftable window for disparity refinement. Experiments were carried out on the Tsukuba, Venus, Teddy and Cones datasets from the Middlebury stereo benchmark [11] and illustrated the advantages of the proposed approach in terms of pixel level accuracy as well as over all surface recovery.

The presented results were generated with our own implementation of the respective methods' published description. Moreover, for performance evaluations we considered the direct output of the algorithms without any post processing enhancement (e.g. median filtering, left-right consistency validation, occlusion mitigation). Accordingly, the results presented here may differ from the original results presented by their respective authors'.

The operational parameters were set as follows. For FBS and AW, we followed the same parameters provided by the authors (FBS: $w_c = 39$, $w_b = 3$, $k_d = 14$, $k_c = 23$ and TAD threshold 53. AW: $w_c = 35$, $k_d = 31$, $k_c = 13$ and TAD threshold 40). For ACTF, we selected the best result from our own experimentation ($w_c = 5$, $w_r = 5$, disparity refine range [-1 .. 1], and Normalized Cross Correlation (NCC) using 4 level Gaussian pyramid). We note that our ACTF implementation incorporates only the interscale shiftable window and does not

	Tsukuba			Venus			Cones			Teddy		
	NOCC	ALL	DISC	NOCC	ALL	DISC	NOCC	ALL	DISC	NOCC	ALL	DISC
SCTF	2.37	4.05	9.91	1.50	2.49	10.02	**4.44**	**12.29**	10.32	**10.19**	**16.96**	21.52
SCTF†	3.68	4.97	17.19	**1.27**	**2.18**	10.41	5.73	12.97	14.51	10.98	17.39	24.21
SCTF*	2.59	4.13	11.30	1.52	2.50	10.20	4.86	12.40	11.60	10.60	17.30	22.60
FBS	2.95	4.75	8.69	1.29	2.87	**7.62**	5.23	15.3	11.34	10.71	19.8	**20.82**
AW	**1.86**	**3.67**	**8.26**	2.09	3.20	11.89	4.49	12.59	**9.98**	10.54	17.43	21.61
ACTF	8.19	10.22	25.59	7.82	9.02	33.18	7.76	17.36	20.91	10.84	19.53	27.00

Table 1: Benchmark data for the considered cost aggregation methods. We represent the percentage of incorrect pixels in disparity map under different scenarios: NOCC concerns all pixels that are not occluded, ALL takes all pixels into consideration, and DISC considers only pixels around discontinuities.

perform half-occlusion handling. As our results will indicate, the omission of this mechanism significantly diminishes performance for ACTF and highlights the effectiveness of our proposed modified shiftable window mechanism. For SCTF, the cost aggregation considered was AW (w_c = 35, w_r = 5, k_d = 17.5, k_c = 13, disparity refine range [-1 .. 1] and TAD threshold 40). The weight image used for scale selection is computed according to our AW parameters. The scale selection pyramid is a LoG pyramid with σ = 1.6, 4 octaves each with 3 levels. As was with ACTF, we used 4 level Gaussian pyramid for CTF stereo. Table 1 shows the percentage of incorrect pixels for the considered benchmarks.

The results presented in Table 1 demonstrate the advantage of using scale selection while integrating adaptive weighting into the shiftable window disparity refinement process. First, we note that performing scale selection (SCTF and SCTF*) is an overall improvement over directly applying adaptive weighing in hierarchical stereo (SCTF†), which can be attributed to the reduction of erroneous disparity initialization in the CTF process. Nevertheless, we also observe instances where the scale agnostic version of our hierarchical approach (SCTF†) outperforms variable cost aggregation methods working at full resolution. We attribute this to the more robust shiftable window mechanism used for disparity refinement. We also note a clear advantage of performing scale selection analysis in the weight image (SCTF), as opposed to analyzing the input color images (SCTF*). This is due to performing scale selection in a similar domain as the one used for our photo-consistency function (i.e. adaptive weighting). However, the accuracy gains come at the expense of performing additional image processing during the scale selection process. Finally, the improvements of our different variations over hierarchical stereo using equally weighted shiftable windows (ACTF).

Figure 4 shows the disparity maps generated from the benchmarks on the Middleburry sequences[10]. As the Figure shows, SCTF was able to better recover textureless surface regions and produce less noise compared to FBS and AW, e.g. on the top right corner in Teddy bear map. This is due to our modified shifting window update that uses estimates from neighboring pixels to repair regions with erroneous disparities. Moreover, in contrast to ACTF and SCTF†, the use of scale selection enabled us to capture detailed structure, e.g. the lamp arm in Tsukuba and the paint brushes in Cones. However, we observe errors where our photo-consistency measure was unable to correctly recover the chimney region in the Teddy bear case at lower resolutions, producing a hole in that region.

Photo-consistency evaluation through adaptive weight template matching is considerably

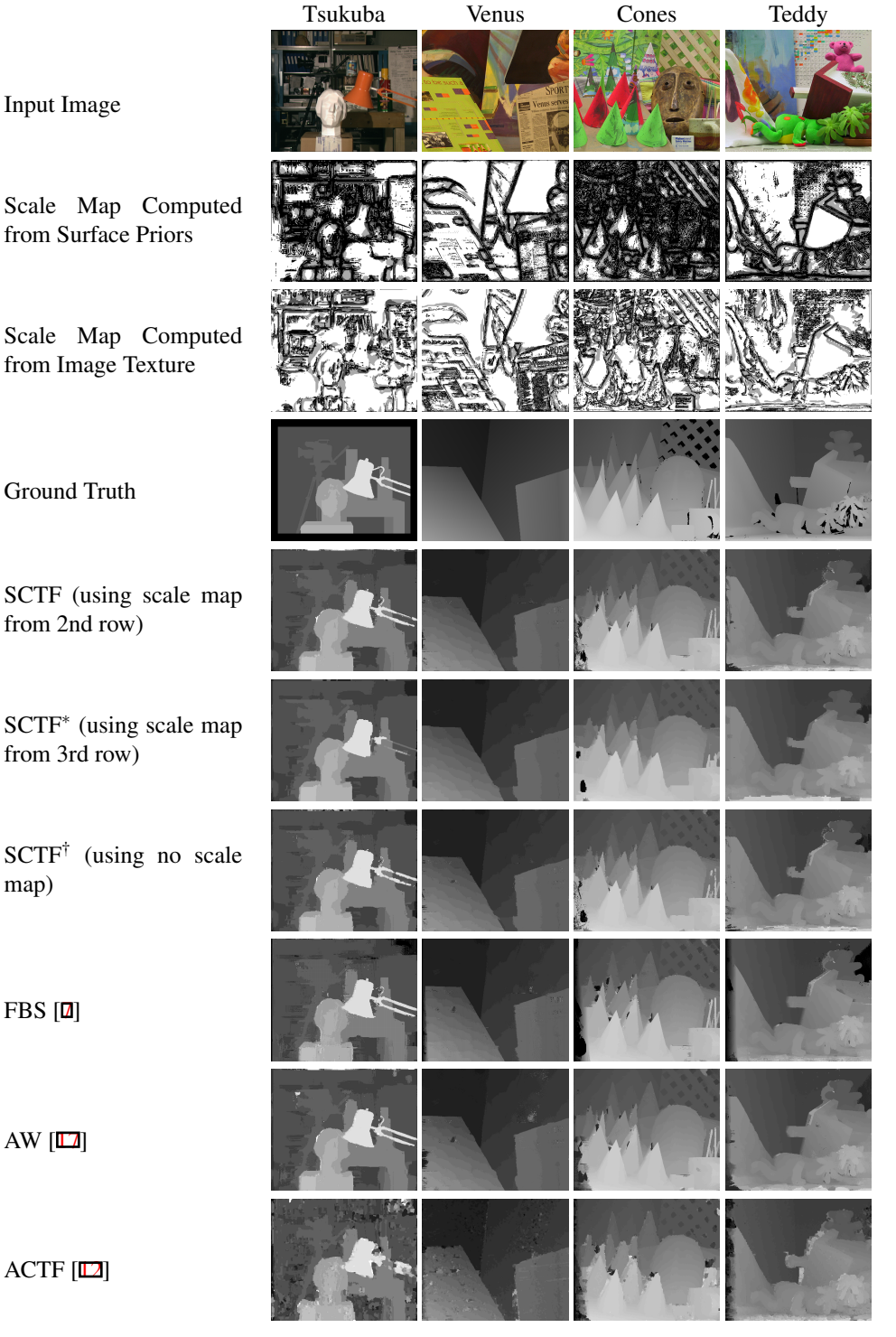

Figure 4: Disparity maps and scale maps for the approaches considered.

	Tsukuba		Venus		Cones		Teddy	
	SCTF	SCTF*	SCTF	SCTF*	SCTF	SCTF*	SCTF	SCTF*
Full-Res Pixel %	44.2	17.1	28.1	17.3	63.7	24.6	34.9	17.2
1st Octave Pixel %	14.2	14.7	10.6	10.5	14.6	14.2	13.6	10.4
2nd Octave Pixel %	8.6	18.9	9.0	17.3	11.1	21.8	12.0	19.8
3rd Octave Pixel %	32.8	49.2	52.1	54.8	10.4	39.2	39.3	52.4
Total Comparisons	1.7M	1.4M	2.3M	2.1M	6.1M	4.1M	4.6M	3.6M
AW Comparisons	2.2M		3.3M		7.7M		7.7M	
ACTF Comparisons	1.2M		1.9M		2.4M		2.4M	

Table 2: Computational costs of scale selective CTF disparity search. While performance is correlated to the distribution of scales within the computed scalemap, full-resolution processing and canonical CTF approaches provide lower and upper bounds on performance.

more robust and computationally expensive in comparison to sum of squared (or absolute) differences. Our motivation for combining CTF and adaptive weighting is to reduce the number of required template comparisons, as they represent the most computationally demanding phase of adaptive image block-based disparity estimation. It is straightforward to determine the number of template comparisons required for full resolution disparity search and for CTF stereo. However, for our proposed scenario of selective scale initialization the total number of comparisons is determined by the scalemap computed from the input imagery. As Table 2 shows, there are significant variations in the pixel scale distribution depending on whether the color image (SCTF*) or the adaptive weight image (SCTF) was utilized to generate the scalemap. Moreover, as Tables 1 and 2 illustrate, the performance/cost tradeoff is dependent on the distribution of pixel scales in the scale map.

5 Discussion and Conclusion

We have addressed the use of adaptive weight matching templates within the context of hierarchical stereo. In order to achieve accurate boundary localization in a CTF framework, we have identified the need for discriminative disparity search within the scale space. To this end, we propose a scale selection mechanism based on spatial and appearance consistency. As a result, our approach is either competitive or outperforms full resolution variable cost aggregation approaches, while gaining the computational efficiency advantages of CTF approaches. On the other hand, by also incorporating conditional weighing in the inter-scale boundary correction process we clearly outperform fine structure estimates of current state of the art CTF stereo methods using variable cost aggregation.

It is reasonable to assume that scale selection based on more accurate structure priors would lead to improved accuracy. Accordingly, an iterative approach where the scale selection mechanism considers both the current depth prior as well as the image texture would be favorable. Initial tests on a straightforward closed loop scale selection were unsuccessful as the improvement among successive iterations was marginal, leading to an unreasonable number of iterations. Accordingly, adopting a statistical modeling approach such as the one used in [4] is a promising future research path.

Acknowledgements. We like to thank SPAWAR, DOE grant DEFG52-08NA28778, and Nvidia for their support.

References

[1] Andrea Fusiello and Vito Roberto. Efficient stereo with multiple windowing. In *CVPR*, pages 858–863, 1997.

[2] D. Gallup, J.-M. Frahm, P. Mordohai, Qingxiong Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *CVPR*, 2007.

[3] Minglun Gong, Ruigang Yang, Liang Wang, and Mingwei Gong. A performance study on different cost aggregation approaches used in real-time stereo matching. *International Journal of Computer Vision*, 75:283–296, 2007.

[4] Takeo Kanade and Masatoshi Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 920–932, 1994.

[5] Tony Lindeberg. Feature detection with automatic scale selection. In *International Journal of Computer Vision*, 1998.

[6] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[7] Stefano Mattoccia, Simone Giardino, and Andrea Gambini. Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering. In *ACCV*, 2009.

[8] Sizintsev Mikhail. Hierarchical stereo with thin structures and transparency. In *Proceedings of the 2008 Canadian Conference on Computer and Robot Vision*, pages 97–104, 2008.

[9] Abhijit S. Ogale and Yiannis Aloimonos. Shape and the stereo correspondence problem. *International Journal of Computer Vision*, 65:147–162, 2005.

[10] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002.

[11] Daniel Scharstein and Richard Szeliski, 2007. http://vision.middlebury.edu/stereo/.

[12] Mikhail Sizintsev and Richard P. Wildes. Efficient stereo with accurate 3d boundaries. In *BMVC*, 2006.

[13] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998.

[14] F. Tombari, S. Mattoccia, L. Di Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. In *CVPR*, 2008.

[15] Federico Tombari, Stefano Mattoccia, and Luigi Di Stefano. Segmentation-based adaptive support for accurate stereo correspondence. In *PSIVT*, 2007.

[16] Yibing Yang and Alan L. Yuille. Multi-level enhancement and detection of stereo disparity surfaces. In *AI*, 1995.

[17] Kuk-jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006.