SPARSE WINDOW LOCAL STEREO MATCHING

Sanja Damjanović, Ferdinand van der Heijden and Luuk J. Spreeuwers
Signals and Systems Group, University of Twente, Hallenweg 15, 7522 NH Enschede, The Netherlands

Keywords: Local stereo matching, Sparse window matching, Sum of squared differences, WTA.

Abstract: We propose a new local algorithm for dense stereo matching of gray images. This algorithm is a hybrid of the pixel based and the window based matching approach; it uses a subset of pixels from the large window for matching. Our algorithm does not suffer from the common pitfalls of the window based matching. It successfully recovers disparities of the thin objects and preserves disparity discontinuities. The only criterion for pixel selection is the intensity difference with the central pixel. The subset contains only pixels which lay within a fixed threshold from the central gray value. As a consequence of the fixed threshold, a low-textured windows will use a larger percentage of pixels for matching, while textured windows can use just a few. In such manner, this approach also reduces the memory consumption. The cost is calculated as the sum of squared differences normalized to the number of the used pixels. The algorithm performance is demonstrated on the test images from the Middlebury stereo evaluation framework.

1 INTRODUCTION

Stereo matching has been actual topic of research for almost four decades since one of the first papers appeared in 1979 (Marr and Poggio, 1979). There is *de facto* established evaluation framework for objective comparison of different stereo algorithms (Scharstein and Szeliski, 2002). Stereo algorithms can be classified into two categories: local and global. Although the global algorithms are more sophisticated and achieve high accuracy, the local algorithms are more present in the practical computer vision applications because of its low computational load and efficient hardware implementation (Lu et al., 2007), (Tombari et al., 2008), (Nalpantidis et al., 2008).

In local stereo matching, the cost is aggregated over a support window which is most often rectangular. It is inherently assumed that all pixels within the matching window have the same disparity. The fronto-parallel assumption is not valid for e.g. curved surfaces due to perspective distortion and occlusion. Therefore, the window-based matching produces different artifacts in the final disparity map: the discontinuities are smoothed and the disparity of the texture richer surfaces are propagated into the lower textured areas (Zitnick and Kanade, 2000). Another limitation is the dimension of the objects whose disparity can be successfully recovered; the object's height and width in the image should be at least half the size of the window dimensions in order to be detected in the window matching. The idea that properly shaped support area for cost aggregation can result in better matching result has been long present in the literature (Zhang et al., 2008), (Tombari et al., 2008), (Hosni et al., 2010).

The ideal window for matching would be only one pixel. However, the one-pixel window does not provide sufficiently discriminatory cost for the local stereo matching. In order to combine the support of many pixels for cost aggregation as in the window-based matching but not to be limited by the window dimension like in the pixel-based matching, we introduce the hybrid support: a set of properly chosen pixels within the rectangular window. We use "sparse window" in cost aggregation and the sum of squared differences normalized to the number of pixels (nSSD) for cost aggregation and the *winner takes all* (WTA) in the disparity selection step.

The pixel selection by thresholding is also present in the work (Zhang et al., 2008), which presents the area-based matching technique. The point-based matching within the global framework is considered in (Mattoccia, 2009) by explicit modeling the mutual relationships among neighboring points. In both of these approaches in (Zhang et al., 2008) and (Mattoccia, 2009) the RGB images are used, while we use gray valued images.

Damjanović S., J. Spreeuwers L. and van der Heijden F..
SPARSE WINDOW LOCAL STEREO MATCHING.
DOI: 10.5220/0003369106890693
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2011), pages 689-693
ISBN: 978-989-8425-47-8

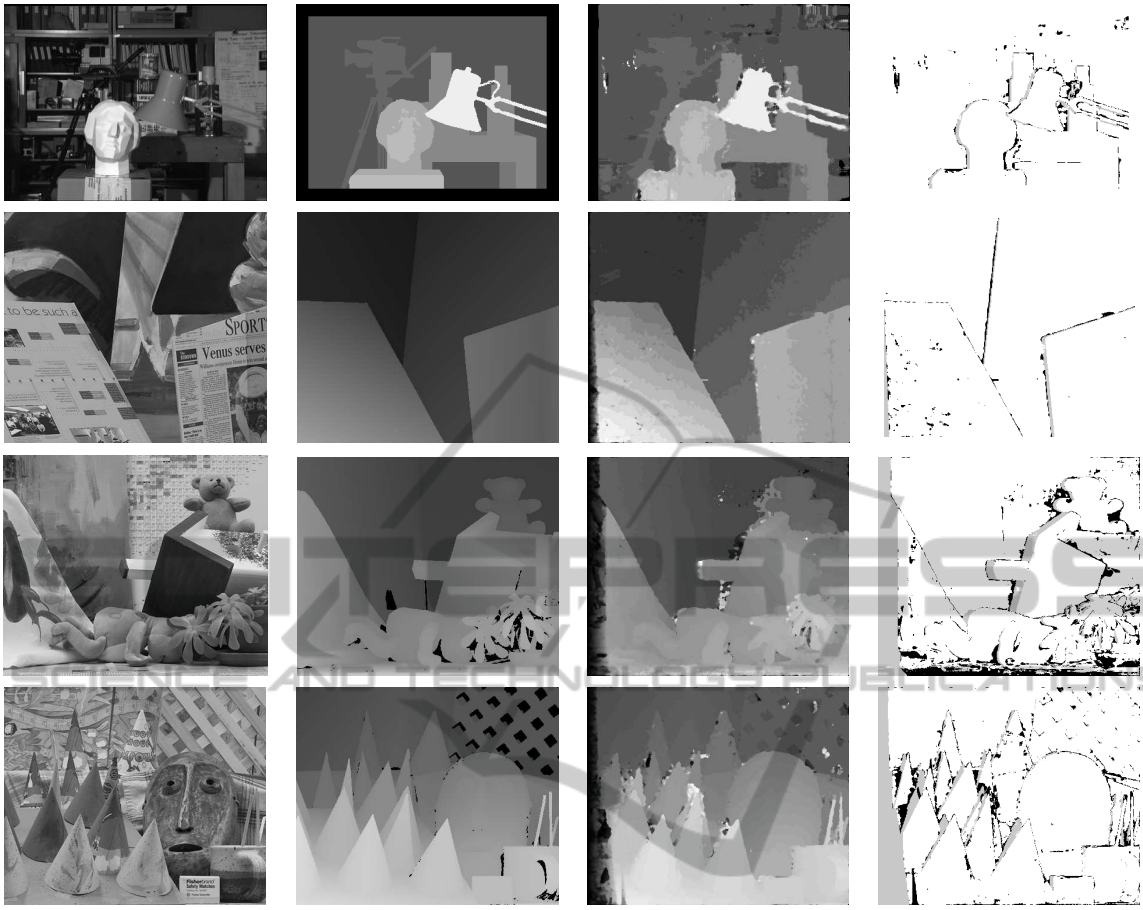

Figure 1: Disparity results for the stereo pairs (1st row: Tsukuba, 2nd row: Venus, 3rd row: Teddy, 4th row: Cones) from the Middlebury database. From left to the right columns show: The left image, Ground truth, Result computed by the sparse window matching technique, Disparity errors larger than 1 pixel. The nonoccluded regions errors with ranking (January 2011) are respectively: *Tsukuba* 2.82% (65), *Venus* 1.20% (67), *Teddy* 9.16% (68), *Cones* 5.91% (75).

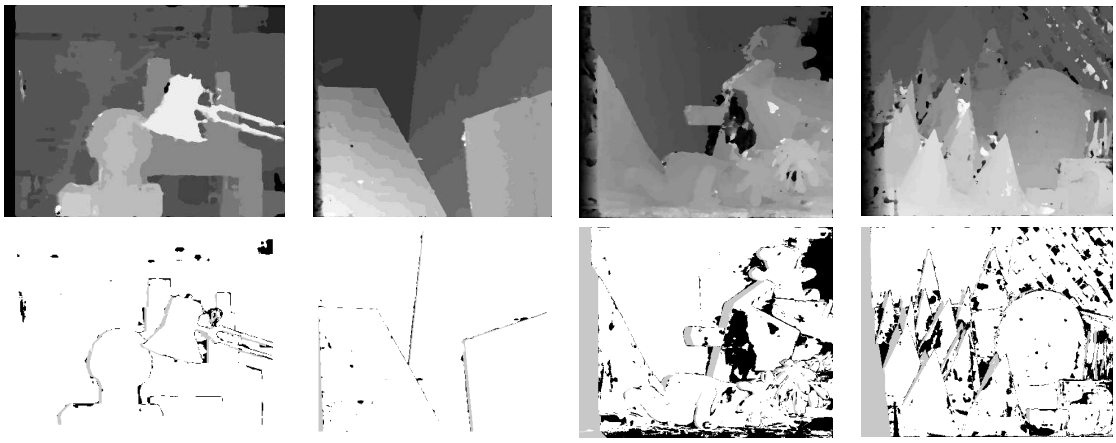

Figure 2: Disparity maps calculated by sparse window technique without the offset compensation [the upper row] and their bad pixels maps [the lower row]. The nonoccluded regions errors with ranking (January 2011) are respectively: *Tsukuba* 2.53% (61), *Venus* 0.63% (47), *Teddy* 17.5% (99), *Cones* 13.8% (101).

2 SPARSE WINDOW MATCHING

2.1 Algorithm Framework

We consider a pair of gray valued rectified stereo images I_L and I_R with disparity range D. We recover the disparity map which corresponds to the reference image I_L. In the matching process, we observe the rectangular $W_x \times W_x, W_x = 2 \cdot w_x + 1$, windows and select some pixels from the left and right matching windows as a suitable for matching. The pixel from the left matching window declared as suitable is selected for the cost aggregation step only if the pixel at the same position from the right window is also declared as suitable for matching. From the N_p selected pixels in each window, we form two $N_p \times 1$ vectors $\mathbf{z_l}$ and $\mathbf{z_r}$. SSD normalized to the number of pixels N_p is used for the cost calculation. Winner-Takes-All (WTA) method is applied to trustworthy disparity candidates. In the postprocessing step, we use the common median filter.

2.2 Pixel Selection

The *continuity constraint* states that disparity varies smoothly everywhere, except on the small fraction of the area on the boundaries of object where discontinuity occurs (Marr and Poggio, 1979). Window based matching methods consider the approximation of the continuity constraint: they assume that all the pixels in the window have the same disparity. This approximation is too rough in many cases e.g. for inclined surface, thin objects, round surfaces. We introduce less restrictive assumption: We assume that the pixels with close gray values have the same disparity i.e. we do not assume that all window pixel have the identical disparity but only some of them. The pixels which are close to the central window pixel in the color space should be used in the cost aggregation step.

We declare the pixel at the position $(i, j), i, j = 1, .., W_x$ in the left window as suitable for matching if its gray value $w_l^{i,j}$ differs from the central pixel's gray value $c_l = w_l^{w_x+1, w_x+1}$ for less than the predefined threshold T_L. The suitable pixels in the right window are chosen in the similar manner. Pixel at the position $(i, j), i, j = 1, .., W_x$ in the right window is declared as suitable for matching if its gray value $w_r^{i,j}$ differs from the central pixel's gray value $c_r = w_r^{w_x+1, w_x+1}$ for less than the predefined threshold T_R. The vectors $\mathbf{z_l}$ and $\mathbf{z_r}$ are formed from the pixels at the position at which pixels in both matching windows are declared as suitable. The pseudo-code of the pixel selection step is given in Algorithm 1.

Algorithm 1: Pixel selection.

```
N_p = 0
for i = 1 to W_x do
  for j = 1 to W_x do
    if |w_l^{i,j} − c_l| < T_L and |w_r^{i,j} − c_r| < T_R then
      N_p = N_p + 1
      add w_l^{i,j} to vector z_l
      add w_r^{i,j} to vector z_r
    end if
  end for
end for
```

With the fixed window size W_x and fixed thresholds T_L and T_R we expect for the low-textured windows to have a high number participating pixels ($N_p \to W_x^2$) and for rich-textured windows sometimes just a few pixels or even one. In these two extreme cases we introduce additional steps in order to prevent errors. In the case of low textured window, we erode the selected pixel mask to prevent that the pixels from the neighboring textureless areas with the similar intensities influence the cost. In the case of rich-textured windows with only several pixels selected for matching, we perform dilation in order to prevent errors due to e.g. aliasing.

2.3 Cost Aggregation

We consider that the constant brightness assumption (CBA) is satisfied in the process of matching. We expect the corresponding pixels to be very close in intensity values, except for the Gaussian noise with the variance σ_n^2. This expectation is justified by the outlier elimination in the process of pixel selection as explained in the previous subsection 2.2. We choose the cost based on the sum of squared differences (SSD) (Belhumeur, 1996), (Cox, 1994). In order to be able to compare the costs with different number of pixels participating in the matching for the same central pixel, we introduce the SSD cost normalized to the number of pixels N_p:

$$C_{nSSD} \propto \frac{1}{N_p} \cdot \frac{\| \mathbf{z_l} - \mathbf{z_r} \|^2}{4 \cdot \sigma_n^2}. \quad (1)$$

The proposed cost eq.(1) is not invariant to unknown pixel offsets which can cause erroneous matching result. We deal with unknown offsets by subtracting a constant from vectors $\mathbf{z_l}$ and $\mathbf{z_r}$, (Damjanović et al., 2009). We choose to subtract the central pixel values c_l and c_r from vectors $\mathbf{z_l}$ and $\mathbf{z_r}$:

$$\mathbf{z_l} = \mathbf{z_l} - c_l \cdot \mathbf{e} \quad (2)$$

$$\mathbf{z_r} = \mathbf{z_r} - c_r \cdot \mathbf{e} \quad (3)$$

where $\mathbf{e}$ is all 1 column vector of the length N_p.

2.4 Adjusted WTA and Postprocessing

The WTA method is used to select the optimal disparity $d^{r,c}$ for the pixel at the position (r,c) in the left image. The WTA method takes into account the number of pixels that support the decision by choosing among the trustworthy disparity candidates. The trustworthy disparity candidates have at least $N_s = K_p \cdot \max\{N_p^{r,c}\}$ pixels participating in the cost aggregation, where $N_p^{r,c}$ is $D \times 1$ vector with number of the participating pixels in the cost aggregation for each possible disparity value. K_p is the ratio coefficient $0 < K_p \leq 1$. The optimal disparity $d^{r,c}$ is found as:

$$d^{r,c} = \arg\min_{d_i}\{C_{nSSD}^{r,c}(d_i) |\, N_p^{r,c}(d_i) > N_s\}, \quad (4)$$

where $r = 1,\ldots,R$ and $c = 1,\ldots,C$, for the image of the dimension $R \times C$ pixels. The postprocessing step performs median $L \times L$ filtering on the disparity map d to eliminate spurious disparities.

3 EXPERIMENT RESULTS AND DISCUSSION

We have used the Middlebury stereo benchmark (Scharstein and Szeliski, 2002) to evaluate the performance of the sparse window technique. The parameters of the algorithm are fixed for all four stereo pairs: $T_L = 10$, $T_R = 10$, $w_x = 15$, $W_x = 31$, $\sigma_n^2 = 0.5$. In the process of pixel selection, we declare the window as textureless if in more than $w_x + 1$ columns and in more than $w_x + 1$ rows, more than half pixels from the left window are selected for matching. The structuring element in erosion step is square $N_E \times N_E$, $N_E = 5$. Dilation is performed with squared $N_D \times N_D$, $N_D = 3$ structuring element, if there are less than N_{min} columns with less than N_{min} pixels or if there are less than N_{min} rows with less than N_{min} pixels, $N_{min} = 5$. WTA parameter is $K_p = 0.5$. Postprocessing step is $L \times L$ median filtering with $L = 5$. These parameters have been found empirically.

The disparity maps obtained by our algorithm (with offset compensation) for the stereo pairs from the Middlebury database are shown in the third column in Figure 1. The leftmost column contains the left images of the four stereo pairs. In the first row are images of the *Tsukuba* stereo pair, followed by *Venus*, *Teddy* and *Cones*. Ground truth (GT) disparity maps are in the second column. The forth column shows the bad disparity maps where the wrong disparities are shown in black. The occlusion regions are in gray and the white regions denote correctly calculated disparity values. The quantitative results in the Middlebury stereo evaluation framework are presented in Table 1. The table shows the ranking of the results together with the error percentages for the nonoccluded region (NONOCC), error for all pixels (ALL), and the error percentage in the discontinuity region (DISC). We consider the ranking of the NONOCC column most important. We do not deal with the occluded and discontinuity regions in our algorithm. The results show that with our hybrid technique edges of the objects are preserved. The disparities of some narrow structures are successfully detected and recovered, although their dimensions are much smaller than the size of the window. Such example of the narrow objects are most noticeable in *Tsukuba* disparity map (the lamp reconstruction) and in *Cones* disparity map (pens in a cup in the lower right corner). On the other hand, the disparities of the large low textured surfaces in stereo pairs *Venus* and *Teddy* are also successfully recovered with the same sparse window technique.

The images in the Middlebury database have different sizes and disparity ranges, as well as different radiometric properties. The stereo pairs *Tsukuba* (384×288 pixels) and *Venus* (434×383) have disparity ranges from 0 to 15 and from 0 to 19. The radiometric properties of the images in these stereo pairs are almost identical, and our algorithm gives even better results without the offset compensation given by eq. (2). The error percentages for the nonoccluded regions for these two pairs without the offset compensation are 2.53% and 0.62% respectively, see Figure 2. Figure 2 shows in the upper row the disparity maps calculated using the sparse window technique without the offset compensation step for all four stereo pair from the evaluation framework and the lower row of figure 2 contains corresponding bad pixel maps with color coding as in the previous figure. The stereo pairs *Teddy* (450×375 pixels) and *Cones* (450×375) have disparity ranges from 0 to 59. The images of these stereo pairs are not radiometrically identical. The sparse window matching without the offset compensation step results in very large errors, see Figure 2. The error percentages for the nonoccluded regions for the stereo pairs *Teddy* and *Cones* without the offset compensation are 17.5% and 13.8%.

4 CONCLUSIONS

We introduced a new sparse window technique for local stereo matching. The algorithm is simple for implementation, as it is based on pixel selection by thresholding, normalized sum of squared differences cost and plain median filtering in the postprocessing step. Our algorithm does not suffer from the common pitfalls of the window-based matching. It does

Table 1: Evaluation results based on the online Middlebury stereo benchmark (Scharstein and Szeliski, 2002): The errors are given in percentages for the nonoccluded (NO) region, the whole image (ALL) and discontinuity (DISC) areas. The numbers in the brackets indicate the ranking in the Middleburry table on January 27th, 2011.

Images	NONOCC	ALL	DISC
Tsukuba	**2.82 (65)**	4.68 (73)	11.7 (67)
Venus	**1.20 (67)**	2.87 (77)	12.4 (73)
Teddy	**9.16 (68)**	18.4 (85)	22.1 (77)
Cones	**5.91 (75)**	16.2 (88)	15.0 (79)

not use color information as many other algorithms and that may improve results in some cases. Yet, the sparse window local stereo matching produces accurate smooth and discontinuity preserving disparity maps. Although, the presented disparity maps are results of only one left to right matching are without parameter optimization, they score well in the comparison with other algorithms, outperforming even some global algorithms and algorithms with much more sophisticated segmentation and postprocessing techniques.

We demonstrated that the sparse window matching is promising technique. Our algorithm can be further improved by introducing disparity map refinement and occlusion treatment.

REFERENCES

Belhumeur, P. N. (1996). A Bayesian approach to binocular stereopsis. *Int. J. Comput. Vision*, 19(3):237–260.

Cox, I. J. (1994). A maximum likelihood n-camera stereo algorithm. In *In IEEE Conf. on Computer Vision and Pattern Recognition*, pages 733–739.

Damjanović, S., van der Heijden, F., and Spreeuwers, L. J. (2009). A new likelihood function for stereo matching: how to achieve invariance to unknown texture, gains and offsets? In *VISIGRAPP 2009, International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Lisboa, Portugal*, pages 603–608, Lisboa. INSTICC Press.

Hosni, A., Bleyer, M., Gelautz, M., and Rhemann, C. (2010). Geodesic adaptive support weight approach for local stereo matching. In *Computer Vision Winter Workshop 2010*, pages 60–65.

Lu, J., Lafruit, G., and Catthoor, F. (2007). Fast variable center-biased windowing for high-speed stereo on programmable graphics hardware. In *ICIP (6)*, pages 568–571.

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London*, B-204:301–328.

Mattoccia, S. (2009). A locally global approach to stereo correspondence. In *3DIM09*, pages 1763–1770.

Nalpantidis, L., Sirakoulis, G. C., and Gasteratos, A. (2008). Review of Stereo Vision Algorithms: from Software to Hardware. *International Journal of Optomechatronics*, 2(4):435–462.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42.

Tombari, F., Mattoccia, S., Di Stefano, L., and Addimanda, E. (2008). Classification and evaluation of cost aggregation methods for stereo correspondence. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Zhang, K., Lu, J., and Lafruit, G. (2008). Scalable stereo matching with locally adaptive polygon approximation. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 313–316.

Zitnick, C. L. and Kanade, T. (2000). A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):675–684.