Segmentation-Based Adaptive Support for Accurate Stereo Correspondence

Federico Tombari[1,2], Stefano Mattoccia[1,2], and Luigi Di Stefano[1,2]

[1] Department of Electronics Computer Science and Systems (DEIS)
University of Bologna
Viale Risorgimento 2, 40136 - Bologna, Italy
[2] Advanced Research Center on Electronic Systems (ARCES)
University of Bologna
Via Toffano 2/2, 40135 - Bologna, Italy
{ftombari, smattoccia, ldistefano}@deis.unibo.it

Abstract. Significant achievements have been attained in the field of dense stereo correspondence by local algorithms based on an adaptive support. Given the problem of matching two correspondent pixels within a local stereo process, the basic idea is to consider as support for each pixel only those points which lay on the same disparity plane, rather than those belonging to a fixed support.

This paper proposes a novel support aggregation strategy which includes information obtained from a segmentation process. Experimental results on the Middlebury dataset demonstrate that our approach is effective in improving the state of the art.

Keywords: Stereo vision, stereo matching, variable support, segmentation.

1 Introduction

Given a pair of rectified stereo images I_r, I_t, the problem of *stereo correspondence* is to find for each pixel of the reference image I_r the correspondent pixel in the target image I_t. The correspondence for a pixel at coordinate $(\bar{x}, \bar{y})$ can only be found at the same vertical coordinate $\bar{y}$ and within the range $[\bar{x} + d_m, \bar{x} + d_M]$, where $D = [d_m, d_M]$ denotes the so-called *disparity range*.

The basic *local* approach selects, as the best correspondence for a pixel p on I_r, the pixel of I_t which yields the lowest score of a similarity measure computed on a (typically squared) fixed support (*correlation window*) centered on p and on each of the $d_M - d_m$ candidates defined by the disparity range. The use of a spatial support compared to a pointwise score increases the robustness of the match especially in presence of noise and low-textured areas, but the use of a fixed support is prone to errors due to the fact that it blindly aggregates pixels belonging to different disparities. For this reason, incorrect matches tend to be generated along depth discontinuities.

In order to improve this approach, many techniques have been proposed which try to select for each pixel an adaptive support which best aggregates only those neighbouring pixels at the same disparity [1], [2], [3], [4], [5], [6] (see [7] and [8] for a review). Recently very effective techniques [8], [9] were proposed, which represent state of the art

D. Mery and L. Rueda (Eds.): PSIVT 2007, LNCS 4872, pp. 427–438, 2007.
© Springer-Verlag Berlin Heidelberg 2007

for local stereo algorithms. The former technique weights each pixel of the correlation window on the basis of both its spatial distance and its colour distance in the CIELAB space from the central pixel. Though this technique provides in general excellent results, outperforming [9] on the Middlebury dataset[1], in presence of highly textured regions the support can shrink to a few pixels thus dramatically reducing the reliability of the matches. Unreliable matches can be found also near depth discontinuities, as well as in presence of low textured regions and repetitive patterns.

This paper proposes a novel adaptive support aggregation strategy which deploys segmentation information in order to increase the reliability of the matches. By means of experimental results we demonstrate that this approach is able to improve the quality of the disparity maps compared to the state of the art of local stereo algorithms.

In the next section we review the state of the art of adaptive support methods for stereo matching. For a more comprehensive survey on stereo matching techniques see [10].

2 Previous Work

In [9] Gerrits and Bekaert propose a support aggregation method based on the segmentation of the reference image (I_r) only. When evaluating the correspondence between two points, $p \in I_r$ and $q \in I_t$, both correlation windows are identically partitioned into two disjoint regions, R_1 and R_2. R_1 coincides with the segment of the reference image including p, R_2 with its complement. Points belonging to R_1 gets a high constant weight, those belonging to R_2 a low constant weight. Cost computation relies on an M-estimator. A major weakness of the method is that the support aggregation strategy is not symmetrical (i.e. it relies on I_r only) hence does not deploys useful information which may be derived from the segmentation of the target image (I_t). Experimental results shows that [9] is clearly outperformed by the algorithm from Yoon and Kweon in [8], which is currently the best local stereo algorithm.

The basic idea of [8] is to extract an adaptive support for each possible correspondence by assigning a weight to each pixel which falls into the current correlation window W_r in the reference image and, correspondingly, in the correlation window W_t in the target image. Let p_c and q_c being respectively the central points of W_r and W_t, whose correspondence is being evaluated. Thus, the pointwise score, which is selected as the Truncated Absolute Difference (TAD), for any point $p_i \in W_r$ corresponding to $q_i \in W_t$ is weighted by a coefficient $w_r(p_i, p_c)$ and a coefficient $w_t(q_i, q_c)$, so that the total cost for correspondence (p_c, q_c) is given by summing up all the weighted pointwise scores belonging to the correlation windows and normalized by the weights sum:

$$C(p_c, q_c) = \frac{\sum\limits_{p_i \in W_r, q_i \in W_t} w_r(p_i, p_c) \cdot w_t(q_i, q_c) \cdot TAD(p_i, q_i)}{\sum\limits_{p_i \in W_r, q_i \in W_t} w_r(p_i, p_c) \cdot w_t(q_i, q_c)} \quad (1)$$

[1] The image pairs together with the groundtruth are available at: http://cat.middlebury.edu/stereo/data.html

Each point in the window is weighted on the basis of its spatial distance as well as of its distance in the CIELAB colour space with regards to the central point of the window. Hence, each weight $w_r(p_i, p_c)$ for points in W_r (and similarly each weight $w_t(q_i, q_c)$ for points in W_t) is defined as:

$$w_r(p_i, p_c) = exp\left(-\frac{d_p(p_i, p_c)}{\gamma_p} - \frac{d_c\left(I_r(p_i), I_r(p_c)\right)}{\gamma_c}\right) \tag{2}$$

where d_c and d_p are respectively the euclidean distance between two CIELAB triplets and the euclidean distance between two coordinate pairs, and the constants γ_c, γ_p are two parameters of the algorithm.

This method provides excellent results but has also some drawbacks, which will be highlighted in the following by analysing the results obtained by [8][2] on stereo pairs belonging to the Middlebury dataset and shown in Fig. 1.

Depth discontinuities. The idea of a variable support is mainly motivated by depth discontinuities: in order to detect accurately depth borders, the support should separate "good" pixels, i.e. pixels at the same disparity as the central point, from "bad" pixels, i.e. pixels at a different disparity from the central point. It is easy to understand that within these regions the concept of spatial distance is prone to lead to wrong separations, as due to their definition border points always have close-by pixels belonging to different depths. Therefore "bad" pixels close to the central point might receive higher weights than "good" ones far from the central point, this effect being more significant the more the chromatic similarities between the regions at different disparities increase. Moreover, as for "good" pixels, far ones might receive a significantly smaller weight than close ones while ideally one should try to aggregate as many "good" pixels as possible. Generally speaking, weights based on spatial proximity from the central point are constant for each correlation window, hence drive toward fixed - not anymore variable - supports, with all negatives consequences of such an approach.

Fig. 2 shows a typical case where the use of spatial distance would determine wrongly the correct support. Imagine that the current point (the blue point in figure) is on the border of two planes at different depths and characterized by a slightly different colour or brightness. The central image shows the correlated pixels (circles coloured from red - high correlation - to yellow - low correlation) on the basis of spatial proximity, where it can be seen that many "bad" pixels would receive a high weight because of the close spatial distance from the central point. Right image depicts in red the correct support that should be ideally extracted. This effect leads to mismatches on some depth borders of the *Tsukuba* and *Venus* datasets, as indicated by the blue boxes of Fig. 1 (groundtruth is shown in Fig. 6).

Low textured surfaces. A further drawback of [8] deals with matching ambiguities which apply when trying to match points belonging to low textured areas on constant depths. When considering the correspondence of points on these areas, the support should ideally enlarge itself as much as possible in order to maximize the signal-to-noise ratio. Instead, the combined use of the spatial and colour proximities force the

[2] The results shown in this paper were obtained running the authors' code available at: http://cat.middlebury.edu/stereo/code.html

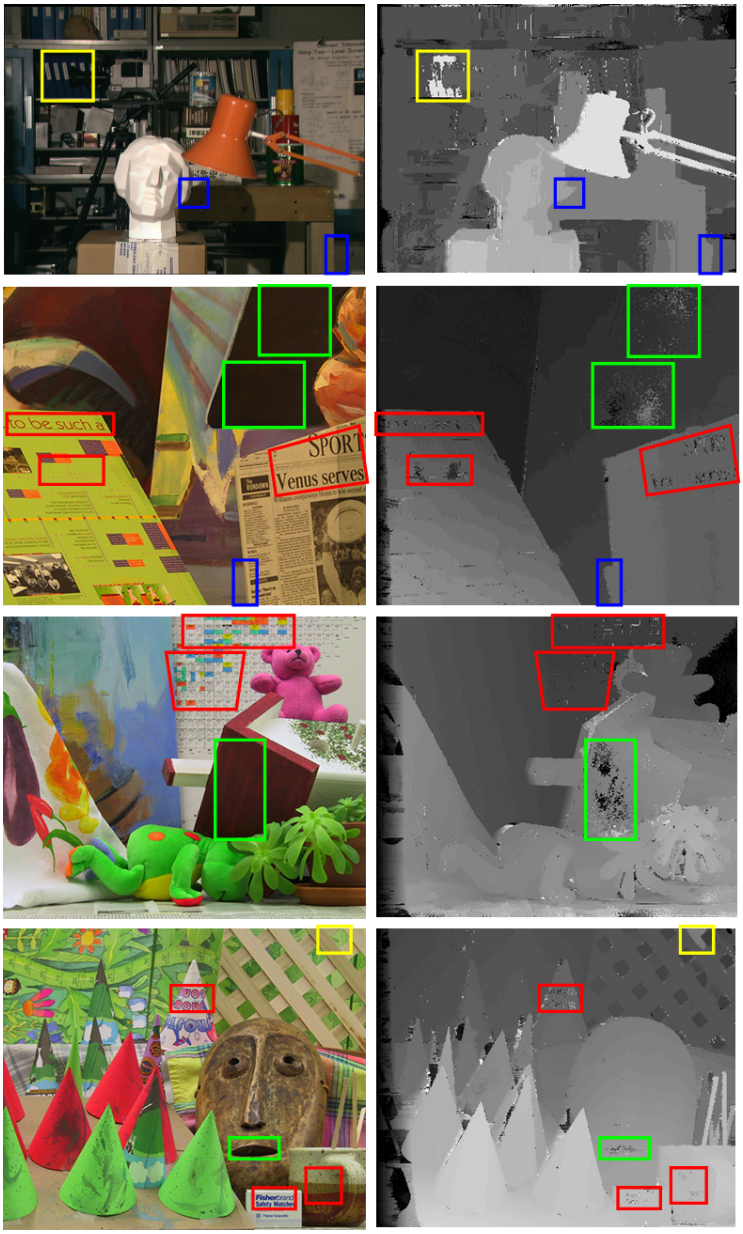

Fig. 1. Some typical artifacts caused by the cost function adopted by [8] on high textured regions (red), depth discontinuities (blue), low textured regions (green), repetitive patterns (yellow). [This image is best viewed with colors].

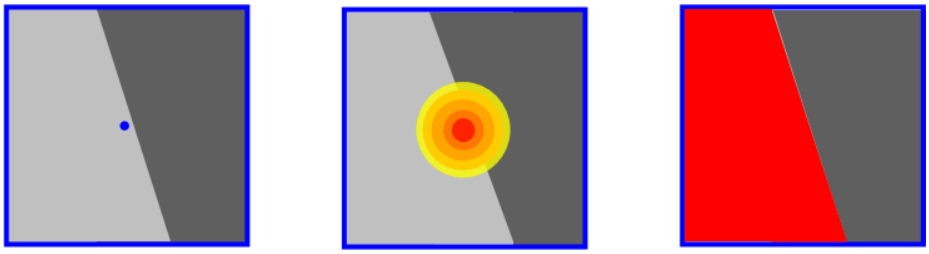

Fig. 2. Example of a correlation window along depth borders (left), correspondent weights assigned by [8] on the basis of spatial proximity (center) and ideal support (right).[This image is best viewed with colors].

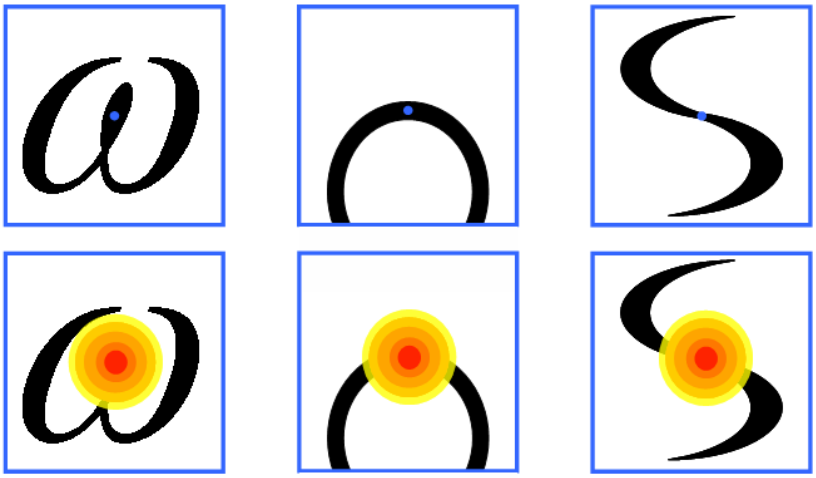

Fig. 3. Examples where the support shrinks to a few elements due to the combined use of spatial and colour proximity. The coloured circles indicate the region correlated to the central pixels on the basis of the spatial proximity.

support to be smaller than the correlation window. This effect is particularly evident in datasets *Venus*, *Cones* and *Teddy*, where the low textured regions denoted by the green boxes of Fig. 1 lead to remarkable artifacts in the correspondent disparity map.

High textured surfaces. Suppose to have a high textured region laying on a constant disparity plane. Then, for all those points having not enough chromatic similarities in their surroundings the aggregated support tends to reduce to a very small number of points. This effect is due to the weights decreasing exponentially with the spatial and colour distances, and it tends to reduce notably the robustness of the matching as the support tends to become pointwise. It is important to note that in these situations the support should ideally enlarge itself and aggregate many elements in the window because of the constant depth.

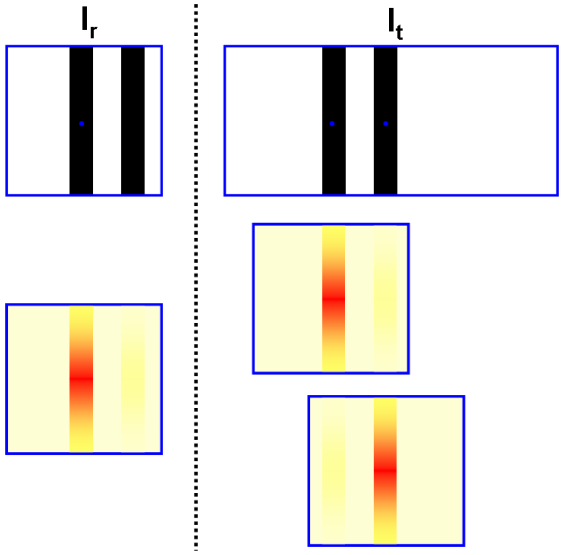

Fig. 4. Typical example of a repetitive pattern along epipolar lines where the aggregation step of [8] would lead to ambiguous match. Red-to-yellow colours are proportional to the weights assigned to the supports.

In order to have an idea of the behaviour of the aggregated support, consider the situation of Fig. 3, where some particular shapes are depicted. In the upper row, the blue point represents the current element for which the support aggregation is computed and the blue square represents the window whose elements concur in the computation of the support. In the lower row the coloured circles denote the points correlated to the central point on the basis of the spatial proximity criterion, where red corresponds to high correlation and yellow to low correlation. As it can be clearly seen the combined use of spatial and colour proximity would lead in these cases to very small aggregated supports compared to the whole area of the shapes as well as to the correlation window area.

Typical artifacts induced by this circumstance are evident in datasets *Venus*, *Cones* and *Teddy* as highlighted by the red boxes in Fig. 1, where it is easy to see that they are often induced by the presence of coloured writings on objects in the scene and that they produce notable mistakes in the correspondent regions of the disparity maps.

Repetitive patterns. Finally, a further problem due to the use of the weight function (1) applies in presence of repetitive patterns along the epipolar lines. As an example consider the situation depicted in Fig. 4. In this case, the blue point in top left image has to be matched with two candidates at different disparities, centered on two similar patterns and shown in top right image. In this situation, the combined use of spatial and colour proximities in the weight function would extract supports similar to the ones shown in the bottom part of the figure, where red corresponds to high weight values and

yellow to low weight values. It is easy to see that the pixels belonging to both candidate supports are similar to the reference support, hence would lead to an ambiguous match. This would not happen, e.g., with the use of the common fixed square support which includes the whole pattern.

In Fig. 1 a typical case of a repetitive pattern along epipolar lines is shown by the yellow box in dataset *Tsukuba*, which lead to mismatches in the disparity map. Also the case depicted by the yellow box in dataset *Cones* seems due to a similar situation.

3 Proposed Approach

The basic idea beyond our approach is to employ information obtained from the application of segmentation within the weight cost function in order to increase the robustness of the matching process. Several methods have been recently proposed based on the hypothesis that disparity varies smoothly on each segment yielded by an (over-)segmentation process applied on the reference image [9], [11], [12]. As the cost function (1) used to determine the aggregated support is symmetrical, i.e. it computes weights based on the same criteria on both images, we propose to apply segmentation on both images and to include in the cost function the resulting information. The use of segmentation allows for including in the aggregation stage also information dealing with the connectiveness of pixels and the shape of the segments, rather than only relying blindly on colour and proximity. Because our initial hypothesis is that each pixel lying on the same segment of the central pixel of the correlation window must have a similar disparity value, then its weight has to be equal to the maximum value of the range(i.e. 1.0). Hence we propose a modified weight function as follows:

$$w'_r(p_i, p_c) = \begin{cases} 1.0 & p_i \in S_c \\ exp\left(-\frac{d_c(I_r(p_i), I_r(p_c))}{\gamma_c}\right) & otherwise \end{cases} \tag{3}$$

with S_c being the segment on which p_c lies. It is important to note that for all pixels outside segment S_c, the proximity term has been eliminated from the overall weight computation and all pixels belonging to the correlation window have the same importance independently from their distance from the central point, because of the negative drawbacks of the use of such a criterion shown in the previous section. Instead, the use of segmentation plays the role of an intelligent proximity criterion.

It is easy to see that this method is less subject to the negative aspects of method [8] outlined in the previous section. The problem of having very small supports in presence of shapes such as the ones depicted in Fig. 3 is improved by segmentation. In fact, as segmentation allows segments to grow as long as chromatic similarity is assessed, the aggregated supports extracted by proposed approach are likely to correctly coincide with the shapes depicted in the figure. Moreover, the use of segmentation in spite of the spatial proximity would allow to extract correctly the support also for border points such as the situation described in Fig. 2, with the extracted support tending to coincide with the one shown on the right of that figure. Improvements are yielded also in presence of low textured areas: as they tend to correspond to a single segment because of the low texture, the support correctly enlarges to include all points of these regions. Finally, in

presence of repetitive patterns such as the ones shown in Fig. 4 the exclusion of the spatial proximity from the weights computation allows only the correct candidate to have a support similar to the one of the reference point.

Moreover, from experimental results it was found that the use of a colour space such as the CIELAB helps the aggregation of pixels which are distant chromatically but which are closer in the sense of the colour space. Unfortunately this renders the colour distance measure less selective, and tends to produce more errors along depth discontinuities.Conversely, the use of the RGB colour space appeared more picky, decreasing the chance that pixels belonging to different depths are aggregated in the same support, but also increasing the number of artifacts along textured regions which lie at the same depth. As the use of segmentation implies adding robustness to the support, we found more convenient to operate in the RGB space in order to enforce smoothness over textured planes as well as to increase the accuracy of depth borders localization.

Finally, it is worth pointing out that there are two main differences between our method and that proposed in [9]: first we apply segmentation on both reference and target images, hence the support aggregation strategy is symmetric. Besides, rather than using two constant weights, we exploit the concept of colour proximity with all benefits of such an approach shown in [8].

4 Experimental Results

In this section we present some experimental results of the proposed method. First we compare our results on the Middlebury dataset with those yielded by [8] using a *Winner-Take-All* (WTA) strategy. The parameter set is kept constant for all image pairs: the set used for the algorithm by Yoon and Kweon is the one proposed in the experimental results in [8], while the set used for the proposed approach is: $\gamma_c = 22.0$, window size $= 51 \times 51$, T (parameter for TAD) $= 80$. For what means the segmentation step in the proposed approach, we use the *Mean-Shift* algorithm [13] with the same constant parameter set, that is: $\sigma_S = 3$ (spatial radius), $\sigma_R = 3$ (range radius), $min_R = 35$ (minimum region size). Figure 5 shows the output of the segmentation stage on both images of each of the 4 stereo pairs used for testing.

Fig. 6 compares the disparity maps obtained by [8] with the proposed approach. Significant improvements can be clearly noticed since the artifacts highlighted in Fig. 1 are less evident or no longer present. In particular, errors within the considered high textured regions on *Venus* and *Teddy* are greatly reduced and almost disappear on *Cones*. Accuracy along depth borders of *Tsukuba* is significantly enhanced while the error along the depth border in *Venus* shrinks to the true occluded area. Moreover, highlighted artifacts present on low textured regions notably decrease on *Venus* and disappear on *Teddy* and *Cones*. Finally, also the artifacts due to the presence of repetitive patterns as shown on *Tsukuba* and *Cones* definitely disappear.

In addition, Table 1 shows the error percentages with regards to the groundtruth, with the error threshold set to 1, computed on the maps of Fig. 6. For each image pair two error measures are proposed: the former is relative to all image area except for occlusions (*N.O.*), the latter only to discontinuities except for occlusions (*DISC*). The error on all image area including occlusions has not been reported because occlusions

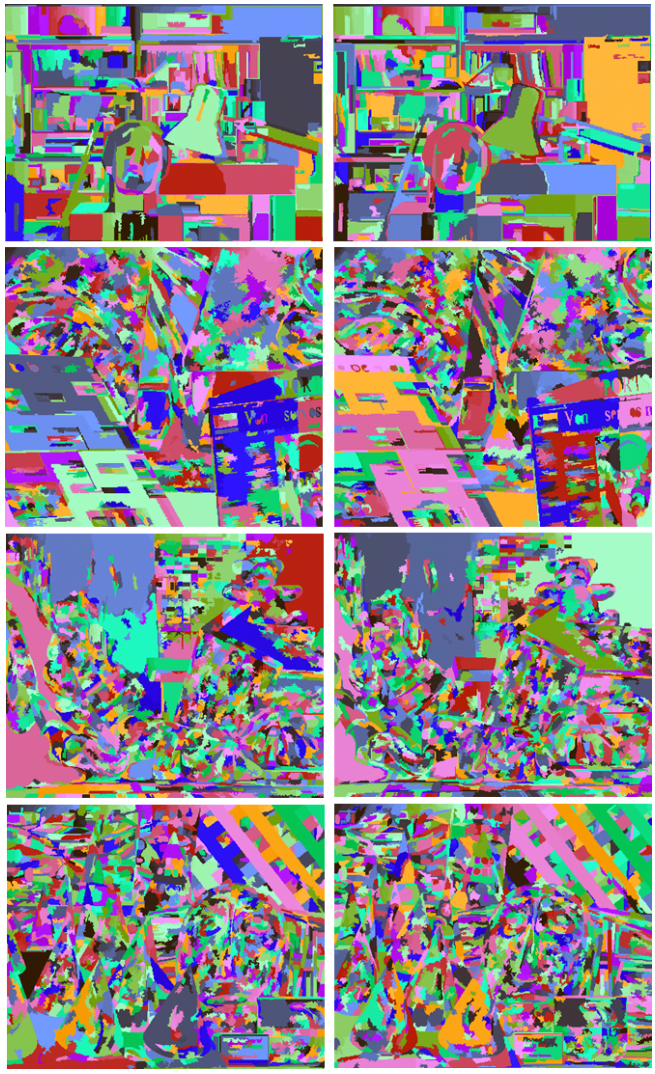

Fig. 5. Output of the segmentation stage on the 4 stereo pairs of the Middlebury dataset

are not handled by WTA strategy. As it can be seen from the table, the use of the proposed approach yields notable improvements for what concerns the error measure on all *N.O.* area. Moreover, by looking only at discontinuities, we can see that generally the proposed approach allows for a reduction of the error rate (all cases except for *Cones*). Benefits are mostly evident on *Venus* and *Tsukuba*.

Finally, we show the results obtained by our method after application of the *Left-Right* consistency check and interpolation of those points which were determined as

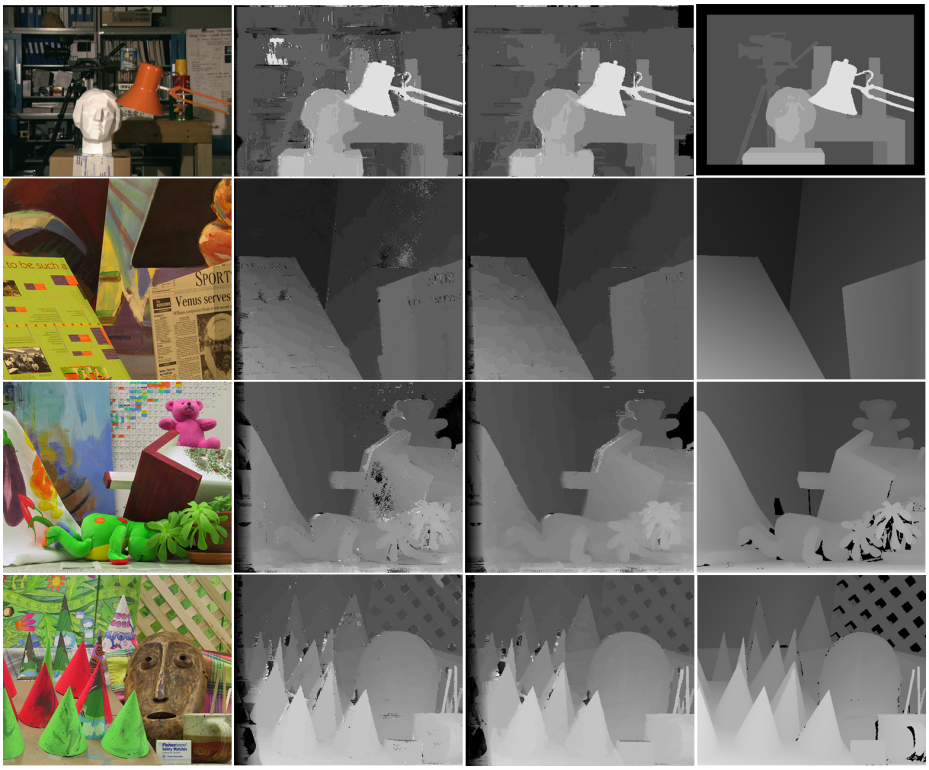

Fig. 6. Reference images (first column), disparity maps computed by [8] (second column) and our approach (third column), ground truth (last column)

Table 1. Comparison between proposed approach and method [8] on the *Middlebury* dataset using a WTA strategy

	Tsukuba	Venus	Teddy	Cones
	N.O. - DISC	N.O. - DISC	N.O. - DISC	N.O. - DISC
Proposed	2,05 - 7,14	1,47 - 10,5	10,8 - 21,7	5,08 - 12,5
[8]	4.66 - 8.25	4.61 - 13.3	12.7 - 22.4	5.50 - 11.9

inconsistent. The obtained disparity maps were submitted and are available at the Middlebury website. We report, in Tab. 2, the quantitative results of our method (referred to as *SegmentSupport*) compared to the submitted results of method [8] (referred to as *AdaptWeight*), together with the overall ranking assigned by Middlebury to the two approaches. The table reports also the results published in [9] which consist only of the error rates on the *ALL* groundtruth maps (all image area including occlusions), since no submission has been done so far on Middlebury. As it is clear from the table and the

Table 2. Disparity error rates and rankings obtained on Middlebury website by the proposed approach (referred to as *SegmentSupport*) compared to method [8] (referred to as *AdaptWeight*) and (where available) [9]

	Rank	Tsukuba	Venus	Teddy	Cones
		N.O. - ALL - DISC	N.O. - ALL - DISC	N.O. - ALL - DISC	N.O. - ALL - DISC
SegmentSupport	♯ 9	1.25 - 1.62 - 6.68	0.25 - 0.64 - 2.59	8.43 - 14.2 - 18.2	3.77 - 9.87 - 9.77
AdaptWeight	♯ 13	1.38 - 1.85 - 6.90	0.71 - 1.19 - 6.13	7.88 - 13.3 - 18.6	3.97 - 9.79 - 8.26
[9]	n.a.	n.a. - 2.27 - n.a.	n.a. - 1.22 - n.a.	n.a. - 19.4 - n.a.	n.a. - 17.4 - n.a.

Middlebury website, currently our approach is the best performing known local method ranking 9th overall (as of July 2007).

5 Conclusions

In this paper a novel support aggregation strategy has been proposed, which embodies the concept of colour proximity as well as segmentation information in order to obtain accurate stereo correspondence. By means of experimental comparisons it was shown that the proposed contribution, deployed within a WTA-based local algorithm, is able to improve the accuracy of disparity maps compared to the state of the art. It is likely that the proposed strategy might be usefully exploited also outside a local framework: this is currently under study.

References

1. Xu, Y., Wang, D., Feng, T., Shum, H.: Stereo computation using radial adaptive windows. In: Proc. Int. Conf. on Pattern Recognition (ICPR 2002), vol. 3, pp. 595–598 (2002)
2. Boykov, Y., Veksler, O., Zabih, R.: A variable window approach to early vision. IEEE Trans. PAMI 20(12), 1283–1294 (1998)
3. Gong, M., Yang, R.: Image-gradient-guided real-time stereo on graphics hardware. In: Proc. 3D Dig. Imaging and modeling (3DIM), Ottawa, Canada, pp. 548–555 (2005)
4. Hirschmuller, H., Innocent, P., Garibaldi, J.: Real-time correlation-based stereo vision with reduced border errors. Int. Jour. Computer Vision (IJCV) 47(1-3) (2002)
5. Kanade, T., Okutomi, M.: Stereo matching algorithm with an adaptive window: theory and experiment. IEEE Trans. PAMI 16(9), 920–932 (1994)
6. Veksler, O.: Fast variable window for stereo correspondence using integral images. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR 2003), pp. 556–561 (2003)
7. Wang, L., Gong, M.W., Gong, M.L., Yang, R.G.: How far can we go with local optimization in real-time stereo matching. In: Proc. Third Int. Symp. on 3D Data Processing, Visualization, and Transmission (3DPVT 2006), pp. 129–136 (2006)
8. Yoon, K.J., Kweon, I.S.: Adaptive support-weight approach for correspondence search. IEEE Trans. PAMI 28(4), 650–656 (2006)
9. Gerrits, M., Bekaert, P.: Local stereo matching with segmentation-based outlier rejection. In: Proc. Canadian Conf. on Computer and Robot Vision (CRV 2006), pp. 66–66 (2006)
10. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. Jour. Computer Vision (IJCV) 47(1/2/3), 7–42 (2002)

11. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: Proc. Int. Conf. on Pattern Recognition (ICPR 2006), vol. 3, pp. 15–18 (2006)
12. Bleyer, M., Gelautz, M.: A layered stereo matching algorithm using image segmentation and global visibility constraints. Jour. Photogrammetry and Remote Sensing 59, 128–150 (2005)
13. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. PAMI 24, 603–619 (2002)