Accurate dense stereo by constraining local consistency on superpixels

Stefano Mattoccia
DEIS-ARCES, University of Bologna
wwww.vision.deis.unibo.it/smatt

Abstract

Segmentation is a low-level vision cue often deployed by stereo algorithms to assume that disparity within superpixels varies smoothly. In this paper, we show that constraining, on a superpixel basis, the cues provided by a recently proposed technique, which explicitly models local consistency among neighboring points, yields accurate and dense disparity fields. Our proposal, starting from the initial disparity hypotheses of a fast dense stereo algorithm based on scanline optimization, demonstrates its effectiveness by enabling us to obtain results comparable to top-ranked algorithms based on iterative disparity optimization methods.

1. Introduction

The effectiveness of many computer vision applications, such as object recognition, 3D modeling and autonomous robot navigation, rely on accurate 3D dense measurements. Depth from binocular stereo, extensively surveyed by Scharstein and Szeliski's [8, 7], represents a well known methodology to infer accurate dense disparity fields. However, despite the attention it has already received, stereo is still an open and important problem.

Our proposal, starting from the initial disparity hypotheses provided by a dense stereo algorithm, aims at determining locally coherent disparity fields by exploiting the cues provided by the Locally Consistent (LC) technique [5] on superpixels (i.e., connected and photometrically coherent regions) obtained by means of segmentation. Similar to other segmentation-based approaches, we assume that disparity within superpixels varies smoothly. This assumption, although sometimes violated in practice, combined with the plausibilities provided by LC on a superpixel basis, provides a powerful cue to set reliable and locally coherent disparity assignments. Since over-segmentation reduces the risk of grouping pixels with significantly different disparities (e.g., region across depth discontinuities), our proposal relies on a two-phase approach. During the first phase, we *heavily* over-segment the reference image in order to obtain small superpixels that are very likely to satisfy the smoothness assumption. This phase, exploiting the plausibility provided by LC on a superpixel basis, aims at detecting unreliable disparity assignments as well as correcting local disparity inconsistencies. During the second phase, we relax over-segmentation in order to obtain larger superpixels with the purpose of propagating coherent disparity assignments within neighboring points. As the second phase segmentation might fail (e.g., near depth discontinuity), we relax the smoothness assumption on larger superpixels. Our proposal, different from most approaches, does not explicitly model disparity within each superpixel by means of a parametrized surface. Experimental results on the standard Middlebury dataset [8, 7] confirm the effectiveness of our proposal.

This paper is organized as follows. The next section briefly introduces the LC technique and segmentation-based stereo algorithms. Afterwards, we will describe our proposal and provide the experimental results on standard datasets.

2. Related work

Our proposal relies on the cues provided by the LC technique and segmentation. Due to limited space, we will only briefly discuss the LC technique and segmentation-based stereo.

The LC technique is a non-iterative approach, recently proposed [5] to enforce local coherence of disparity fields. Given an initial dense disparity field and by analyzing the behavior of neighboring disparity values, this approach enables us to obtain a measure (referred to as plausibility) that encodes the degree of reliability of each disparity hypothesis. Moreover, the LC technique enables us to obtain two different disparity fields in a single iteration. A detailed description of the LC technique can be found in [5].

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.452

Starting from the work of Tao and Sawhney [9], image segmentation has been systematically used in stereo and most current top-ranked algorithms (e.g., [10, 4, 12, 11]) rely on this cue to obtain the excellent results reported in [7]. In most cases, these methods iteratively model the disparity within each superpixel assuming planar slanted surfaces, perform robust plane fitting (by means of RANSAC or histogram voting) on a superpixel basis, and perform disparity optimization by means of *graph cuts*, *belief propagation* or *cooperative optimization*. Other approaches deploy segmentation for matching cost computation (e.g., [6, 2]) or to detect outliers/refine disparity assignments (e.g., [3, 6]). An extensive survey of the stereo vision literature was given in [8] and an evaluation of most state-of-the-art approaches can be found in [7].

3. Proposed approach

Our method starts by considering the disparity hypotheses made by a dense stereo algorithm. For our experiments, we deploy the disparity fields $\mathcal{D}_S$ provided by the C-SemiGlobal algorithm [3] available in [7]. Processing $\mathcal{D}_S$ by means of the LC approach [5] allows us to obtain two locally coherent disparity fields $\mathcal{D}_R$ and $\mathcal{D}_T$ (w.r.t. the reference image R and the target image T, respectively) and a *plausibility distribution* $\mathcal{P}(p|d)$ that, according to [5], encodes for each point p of the reference image R the accumulated posterior probability for each disparity hypothesis $d \in [d_{min}, d_{max}]$. The two locally consistent disparity fields $\mathcal{D}_R$ and $\mathcal{D}_T$, obtained by means of a simple *winner-takes-all* strategy on $\mathcal{P}(p|d)$, are cross-checked (w.r.t. the reference image R), obtaining the $\mathcal{D}^*$ disparity field, in order to identify inconsistent disparity assignments.

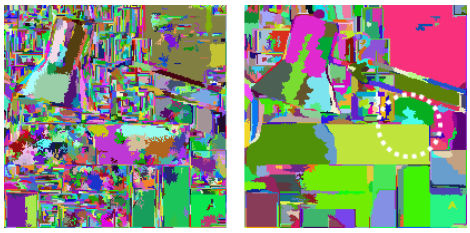

Figure 1. Tsukuba (portion of the reference image R): superpixels obtained by means of (Left) over-segmentation and (Right) relaxing over-segmentation. The circle highlights a superpixel overlapping a depth discontinuity (table/background).

Once $\mathcal{D}^*$ and $\mathcal{P}(p|d)$ are available, we aim at further constraining local consistency within *coherent* neighboring points. For this purpose, as shown in the left of Figure 1, we over-segment the reference image R so as to to cluster photometrically coherent neighboring connected points. This step aims to obtain superpixels that are very likely to satisfy the smoothness assumption.

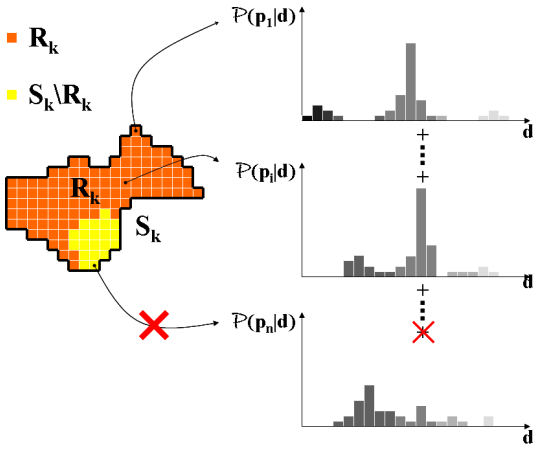

Figure 2. Gathering the plausibility distribution $\mathcal{P}(S_K|d)$ on a superpixel basis. In yellow, points marked as unreliable in $D*$.

At this point we gather the plausibility distribution on each superpixel S_K by summing the plausibility distribution $\mathcal{P}(p|d)$ of each point $p \in S_K$. In order to limit perturbations brought in by unreliable points (e.g., occlusions), we discard those points that do not satisfy cross-checking between $\mathcal{D}_R$ and $\mathcal{D}_T$. That is, as depicted in Figure 2, if R_K is the set of reliable pixels belonging to S_K (i.e., points marked as reliable in $\mathcal{D}^*$), the plausibility distribution within each superpixel S_K is defined as:

$$\mathcal{P}(S_K|d) = \sum_{p \in R_K \subset S_K} \mathcal{P}(p|d) \qquad (1)$$

For each S_K, we analyze the resulting plausibility distribution $\mathcal{P}(S_K|d)$ and the percentage of valid points V (i.e., the ratio between the cardinality of R_K and S_K). For what concerns the plausibility distribution, we search for the *dominant* disparity $\tilde{d}(S_K)$ (i.e., the most plausible disparity) within S_K,

$$\tilde{d}(S_K) = \underset{d \in [d_{min}, d_{max}]}{\operatorname{argmax}} \mathcal{P}(S_K|d) \qquad (2)$$

Gathering this information, for each $p \in S_K$ we aim to regularize the disparity field $\mathcal{D}^*$ as well as to detect inconsistent disparity assignments, adopting the following criteria:

$p \in R_K$ - Point p has a potentially valid disparity assignment $d(p)$; therefore, we validate $d(p)$ by enforcing that this value is consistent with other points within the superpixel. Local consistency within S_K is validated by comparing $d(p)$ to the dominant disparity $\tilde{d}(S_K)$. That is, we compute $\delta(p) = |d(p) - \tilde{d}(S_K)|$ and we set p as unreliable if $d(p)$ *significantly* differs from $\tilde{d}(S_K)$ (i.e., $\delta(p) \geq \delta_t$, δ_t being a parameter set empirically) deferring inference of disparity to a later stage. Conversely, if $d(p)$ is consistent with $\tilde{d}(S_K)$, we keep the $d(p)$ assignment, trusting in the disparity set by LC.

$p \notin R_K$ - In this case, $d(p)$ was retained unreliable by cross-checking; therefore, if p belongs to a superpixel containing a *sufficient* number of valid points (i.e., $V \geq V_t$, V_t being a parameter set empirically), we assume that we are within a superpixel densely populated by disparities and we replace the missing measurement with the dominant disparity $\tilde{d}(S_K)$. Conversely, if the segment is not sufficiently dense (i.e. $V < V_t$), we let $d(p)$, as invalid, deferring inference of disparity to a later stage when larger superpixels, including more neighboring points, will be analyzed.

For segmentation, we use the Mean Shift algorithm[1] [1] with *spatial* bandwidth S_B=6, *range* bandwidth $R_B = 1.0$ and *minimum region area* $MRA = 5$. As parameters δ_t and V_t, we deploy 2 and 40% respectively. In both cases (i.e., $p \in R_K$ and $p \notin R_K$), the key assumption is that over-segmentation enables us to be confident that disparity within S_K varies smoothly. It is noteworthy that our approach, despite not explicitly modeling disparity within superpixels with a parametrized surface model (e.g., a plane), allows the handling ($\delta_t \geq 2$) of slanted surfaces. The output of this stage is the regularized disparity field $\mathcal{D}^{(1)}$. Typically, in $\mathcal{D}^{(1)}$, locally incoherent disparity assignments have been detected and reliable disparity assignments have been propagated within superpixels (e.g., near occlusions). Nevertheless, $\mathcal{D}^{(1)}$ still contains missing disparity assignments (mostly in occluded regions).

Once that disparity coherence on superpixels obtained by means of over-segmentation has been strictly enforced, the successive phase mostly aims at propagating reliable disparity assignments in regions containing missing values. For this purpose, we relax over-segmentation so as to enforce local coherency on larger superpixels. In this second phase, we exploit the evidence that relaxing over-segmentation allows the clustering of (portions of) multiple *previous* superpixels. Constraining $\mathcal{P}(S_K|d)$ on the larger superpixels, by means of the approach described previously, enables consistent and reliable disparity assignments to propagate within neighboring points. However, relaxing over-segmentation may lead to superpixels that lie across depth discontinuity (e.g., the superpixel highlighted with the circle in Figure 1). Therefore, to deal with this problem, we enforce a weaker smoothness constraint. That is, we allow larger disparity variations within superpixels by setting $\delta_t = 10$ (keeping parameter $V_t = 40\%$). Regarding the Mean Shift algorithm, in this phase, we relax over-segmentation by deploying $S_B = 6$, $R_B = 3.0$, $MRA = 10$. Typically, the resulting disparity field $\mathcal{D}^{(2)}$ is almost completely dense except within those superpixels completely located within occluded regions. We render $\mathcal{D}^{(2)}$ completely dense by means of a very simple interpolation procedure, obtaining the final disparity field $\mathcal{D}^{(3)}$.

4. Experimental results

We evaluated[2] our proposal deploying the disparity fields $\mathcal{D}_S$ of the C-SemiGlobal algorithm [3] available on Middlebury [7]. These disparity fields were processed by means of the LC technique [5] so as to obtain $\mathcal{D}_R$, $\mathcal{D}_T$ and $\mathcal{P}(p|d)$. For our experiments, the Mean Shift algorithm was always deployed with the fast processing mode enabled (i.e., switch HIGH_SPEEDUP in the source code). Table 1 reports, according to the methodology defined in the Middlebury evaluation site [8, 7] and with constant parameters for the four stereo pairs, the results of the current five top-ranked approaches [10, 4, 12, 11, 13], our proposal ($\mathcal{D}^{(3)}$), the C-SemiGlobal approach [3] ($\mathcal{D}_S$) and the output of the raw LC technique ($\mathcal{D}_R$). The table indicates that our proposal dramatically improves the overall accuracy of [3]. The improvements are significant on the whole dataset, but are particularly notable on Tsukuba and Venus. According to the current Middlebury evaluation, our proposal is ranked 3rd out of 85 approaches. It is also worthy of note that, for Venus, it is the best overall performing algorithm for NOCC and DISC errors, and the best, overall, with regards to NOCC error in Tsukuba. Comparing the results for $\mathcal{D}_R$ and $\mathcal{D}^{(3)}$, notable overall improvements resulting from our proposal compared to the raw LC technique applied to the original $\mathcal{D}_S$ disparity fields can also be highlighted. However, it can also be seen from the table that, on Teddy, compared to the initial disparity field $\mathcal{D}_S$, our method is less effective, according to NOCC and DISC errors. Although

[1] Source code available at: http://www.caip.rutgers.edu/riul/research/code/EDISON/index.html

[2] Additional experimental results available at: www.vision.deis.unibo.it/smatt/ICPR2010.htm

Algorithm	Rank	Tsukuba NOCC	Tsukuba ALL	Tsukuba DISC	Venus NOCC	Venus ALL	Venus DISC	Teddy NOCC	Teddy ALL	Teddy DISC	Cones NOCC	Cones ALL	Cones DISC
CoopRegion [10]	♯1	0.87_{2}	$\mathbf{1.16_{1}}$	$\mathbf{4.61_{1}}$	0.11_{3}	0.21_{2}	1.54_{5}	5.16_{11}	8.31_{8}	13.0_{8}	2.79_{7}	7.18_{4}	8.01_{10}
AdaptingBP [4]	♯2	1.11_{10}	1.37_{6}	5.79_{12}	0.10_{2}	0.21_{3}	1.44_{3}	4.22_{4}	7.06_{5}	11.8_{5}	2.48_{3}	7.92_{7}	7.32_{4}
Proposed $\mathcal{D}^{(3)}$	♯**3**	$\mathbf{0.87_{1}}$	1.31_{3}	4.69_{2}	$\mathbf{0.09_{1}}$	0.29_{9}	$\mathbf{1.29_{1}}$	5.44_{12}	11.0_{15}	13.6_{11}	2.48_{2}	8.16_{10}	6.97_{2}
DoubleBP [12]	♯4	0.88_{4}	1.29_{2}	4.76_{4}	0.13_{6}	0.45_{14}	1.87_{9}	3.53_{3}	8.30_{7}	9.63_{2}	2.90_{9}	8.78_{18}	7.79_{7}
OutlierConf [11]	♯5	0.88_{3}	1.43_{8}	4.74_{3}	0.18_{12}	0.26_{7}	2.40_{15}	5.01_{7}	9.12_{11}	12.8_{7}	2.78_{6}	8.57_{14}	6.99_{3}
SubPixDoubleBP [13]	♯6	1.24_{17}	1.76_{19}	5.98_{13}	0.12_{5}	0.46_{15}	1.74_{8}	3.45_{2}	8.38_{9}	10.0_{3}	2.93_{11}	8.73_{17}	7.91_{9}
...	...	...	...	...	...	...	...	...	...	...	...	...	...
C-SemiGlobal [3] $\mathcal{D}_S$	♯17	2.61_{53}	3.29_{44}	9.89_{49}	0.25_{21}	0.57_{19}	3.24_{26}	5.14_{10}	11.8_{20}	13.0_{8}	2.77_{5}	8.35_{12}	8.20_{11}
...	...	...	...	...	...	...	...	...	...	...	...	...	...
$\mathcal{D}_R$	n.a.	0.93	2.57	4.97	0.21	1.67	2.69	5.39	14.5	13.4	3.07	13.0	7.84

Table 1. Accuracy (errors NOCC, ALL, DISC) according to the methodology defined on the Middlebury evaluation site [8] for top-ranked algorithms, $\mathcal{D}_S$ [3], $\mathcal{D}_R$ and our proposal $\mathcal{D}^{(3)}$.

we will further analyze this fact, a preliminary analysis highlights that most of these errors occur near the bottom border of the image. Thus, it is plausible that these errors can be ascribed to the very simple interpolation step deployed. Finally, the overall execution time (including C-SemiGlobal[3], LC processing and segmentation) on a standard PC with current unoptimized code is about 30 sec for the Teddy stereo pair. This result is comparable to, or even better than, those of the top-ranked approaches in [7].

5. Conclusions

In this paper, we have shown that constraining local consistency on a superpixel basis allows us to dramatically improve the overall performance of a fast and accurate dense stereo algorithm. Our proposal relies on the cues provided by the LC technique [5] and was evaluated deploying the disparity fields provided by [3]. However, our proposal can be used with any dense stereo algorithm. Moreover, although susceptible to further improvements (e.g., deploying a parametrized surface model for disparity within superpixels), our proposal, in its current form, allows us to obtain top-ranked results, comparable to approaches based on global disparity optimizations.

References

[1] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. PAMI*, 24:603–619, 2002.

[2] M. Gerrits and P. Bekaert. Local stereo matching with segmentation-based outlier rejection. In *Proc. CRV 2006*, pages 66–66, 2006.

[3] H. Hirschmuller. Stereo processing by semi-global matching and mutual information. *IEEE Trans. on PAMI*, 2(30):328–341, 2008.

[4] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *ICPR '06*, pages 15–18, 2006.

[5] S. Mattoccia. A locally global approach to stereo correspondence. In *3DIM2009*, pages 1763–1770, Kyoto, Japan, 2009.

[6] S. Mattoccia, F. Tombari, and L. Di Stefano. Stereo vision enabling precise border localization within a scanline optimization framework. In *Proc. Asian Conf. on Computer Vision (ACCV 2007)*, pages 517–527, 2007.

[7] D. Scharstein and R. Szeliski. Middlebury stereo vision. http://vision.middlebury.edu/stereo/.

[8] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. Jour. Computer Vision*, 47(1/2/3):7–42, 2002.

[9] H. Tao and H. Sawhney. Global matching criterion and color segmentation based stereo. In *WACV00*, pages 246–253, 2000.

[10] Z.-F. Wang and Z.-G. Zheng. A region based stereo matching algorithm using cooperative optimization. In *CVPR*, 2008.

[11] L. Xu and J. Jia. Stereo matching: An outlier confidence approach. In *ECCV 2008*, pages 775–787, 2008.

[12] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Trans. PAMI*, 31(3):492–504, 2009.

[13] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *Proc. of CVPR2007*, pages 1–8, 2007.

[3] Execution time of a few seconds according to [3].