Leveraging confident points for accurate depth refinement on embedded systems

Fabio Tosi, Matteo Poggi, Stefano Mattoccia
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{fabio.tosi5, m.poggi, stefano.mattoccia}@unibo.it

Abstract

Despite the notable progress in stereo disparity estimation, algorithms are still prone to errors in challenging conditions. Thus, heuristic disparity refinement techniques are usually deployed to improve accuracy. Moreover, state-of-the-art methods rely on complex CNNs requiring power hungry GPUs not suited for many practical applications constrained by limited computing resources. In this paper, we propose a novel technique for disparity refinement leveraging on confidence measures and a novel, automatic learning-based selection method to discard outliers. Then, a non-local strategy infers missing disparities by analyzing the closest reliable points. This framework is very fast and does not require any hand-tuned thresholding. We assess the performance of our Non-Local Anchoring (NLA), standalone refinement techniques and methods leveraging on confidence measures inside the stereo algorithm. Our evaluation with two popular stereo algorithms shows that our proposal significantly ameliorates their accuracy on Middlebury v3 and KITTI 2015 datasets. Moreover, since our method relies only on cues computed in the disparity domain, it is suited even for COTS stereo cameras coupled with embedded systems, e.g., nVidia Jetson TX2.

1. Introduction

Stereo is one of the most popular technique to infer depth from two or more images and challenging datasets, such as KITTI [5, 17] and Middlebury [28], clearly emphasized that it is still an open problem. State-of-the-art algorithms [3, 12] require expensive and power-hungry GPUs to run in a reasonable amount of time, making them unsuited for many practical applications constrained by hardware resources or energy consumption. Conventional (i.e., pre-deep learning) algorithms still achieve accurate results leveraging on multi-step pipelines, each one contributing to increasing the overall effectiveness with different degrees of reliability. A notable example is the Semi-Global Matching algorithm (SGM) [7], implemented in many variants thanks to its trade-off between accuracy and complexity, that usually deploy interpolation and refinement steps on estimated disparity maps. In their seminal work, Zbontar and LeCun [43] showed how plugging deep learning into a conventional stereo SGM pipeline yielded very accurate results on KITTI and Middlebury datasets not far from end-to-end networks [3, 12].

One of the steps involved, referred to as disparity refinement, attempts to recover errors from the disparity map. While some refinement procedures rely on simple filters (e.g., median or bilateral filters) others exploit cues from the disparity map and the input stereo pair. Confidence measures allow to detect unreliable matches produced by stereo algorithms and, recently, strategies based on machine-learning achieved state-of-the-art results [26]. Confidence measures have been deployed in different steps of stereo pipelines, with the aim to further improve the overall accuracy. In this paper, we propose Non-Local Anchoring (NLA), a novel disparity refinement method relying on confidence measures, outlined in Figure 1. Given a disparity map generated by any stereo algorithm, the confidence
Figure 2. NLA in action on KITTI 2015 dataset [17]. (a) Left frame from stereo pair 000027, (b) raw disparity map computed by the Block-Matching algorithm (BM) algorithm, (c) sparse disparity map containing assumed reliable points RP, (d) refined disparity map with our proposal.

measure allows us to detect and removing erroneous pixels. Then, for each discarded pixel, among the remaining reliable points (RP) a subset of anchors, not necessarily in a close neighborhood of the examined pixel, is chosen to infer, according to both spatial and color information from the reference image, a new disparity value. Moreover, a CPU-friendly machine-learning framework based on a random forest classifier is proposed to deal with automatic identification of unreliable disparity assignments by analyzing local and global properties of the confidence on the whole image. This novel strategy allows us to remove the need for a heuristic selection of a confidence threshold often carried out in this field [31, 29].

To assess the effectiveness of our proposal, we report an extensive evaluation on the Middlebury v3 dataset comparing our framework to conventional disparity refinement methodologies as well as with recently proposed confidence-based approaches, acting on the Disparity Space Image (DSI) [27] also referred to as the cost-volume. Differently, NLA acts in the disparity domain hence does not require at all the cost volume that might be not available in some circumstances, e.g., when dealing with a commercial off-the-shelf (COTS) stereo camera. Factors like the number of anchors deployed and a further local aggregation strategy included in our proposal will be discussed and compared. Moreover, we evaluate our framework also on KITTI 2015 dataset [17] to further confirm the effectiveness of our method on indoor and outdoor data. Figure 2 shows the outcome of our proposal on the frame 166 of the KITTI 2015 dataset deploying the disparity map generated by the popular Block-Matching (BM) algorithm.

Finally, we point out how the proposed strategy works by acting in the disparity domain only with reduced computational complexity, fitting very well with COTS stereo cameras and in general with embedded devices, such as nVidia Jetson TX2 used to measure runtime in our experiments.

2. Related work

Confidence measures The confidence measure literature is relevant to our work. Initially Hu and Mordohai [8] carried out an extensive taxonomy and evaluation of confidence measures, categorizing them according to the processed cues. They also proposed a common metric to compare the effectiveness of different confidence measures and assess their performance to detect uncertain disparity assignments. More recently, machine learning techniques have been used to infer more effective confidence measures. Hausler et al. [6] deployed, for the first time, a random forest combining different (as much as possible) orthogonal measures and features aimed at classifying each pixel as a correct or wrong match. Improved methods based deploying a random forest were proposed in [31, 20, 23]. Recent works deployed CNNs to infer confidence as well. In particular in [24] by only processing the left disparity map, in [29] using as input cue the left and right maps and in [36] exploiting local and global cues. Finally, confidence measures have been effectively deployed as cues to improve the overall accuracy of stereo within conventional pipelines as recently shown in [29, 23, 20, 31]. Poggi et al. [26] evaluated conventional and learning-based confidence measures highlighting how the learned-ones are much more effective. Finally, we point out that unsupervised learning of confidence measures has been tackled in [18] and [37].

Pre-deep learning stereo. The extensive literature concerning stereo has been enriched, in the years, by several contributions and novel methodologies. Despite that, the majority of them (with rare exceptions such as [16]) can be still categorized according to the taxonomy proposed
by Scharstein and Szeliski [27], which also lists common steps: cost computation, cost aggregation, disparity computation and disparity refinement. Algorithms focusing on the first two steps are referred to as local while global methods mainly rely on the optimizing steps. The latter ones are usually more effectively but computationally more expensive. Semi-Global Matching (SGM) proposed by Hirschmuller [7] represents a good trade-off between performance and accuracy making it one of the most deployed solutions to infer dense disparity maps in practical applications. Disparity refinement techniques are usually deployed on top of traditional stereo pipelines to push accuracy toward optimality further. Common approaches consist of image filtering operations, like median filter [21], weighted median filter [44], guided filter [40], bilateral filter [33] or the recent fast bilateral solver [2]. One of the most effective and used refinement technique consists in a left-right consistency check to detect both occlusions and mismatches, interpolating the former points with the background and the latter ones with a median from nearby disparity. Such strategy [7, 43] is often coupled with median and bilateral filters.

Deep learning stereo. The recent spread of deep learning was applied to stereo as well, with [32, 43] being the first ones to tackle matching cost computation employing a Convolutional Neural Network (CNN) working on image patches. Other authors followed this path addressing efficiency [4, 13]. In [22] a CNN is deployed to combine multiple out-of-the-box stereo matchers to obtain more accurate results, inspired by the work of Spyropolous and Morohosi [32] which carried it out by using a random forest. Conversely, Mayer et al. [16] proposed DispNet, the first deep architecture for end-to-end stereo computation, completely departing from conventional stereo methodologies. Although this method is not the most accurate one on KITTI and Middlebury datasets, it represents a ground-breaking approach to tackle stereo. Since then, the study of end-to-end architectures for dense stereo matching became dominant. Kendall et al. [10] introduced 3D CNNs to regularize a volume obtained through concatenation between left and right features on the disparity axis. Further works improve the results of this approach [3, 39, 11] or 2D networks [19, 12], optionally learning stereo jointly with other tasks [38, 30, 9]. Some works addressed domain shift issues affecting the aforementioned frameworks, by either adapting to new environments offline [34] or online [35], as well as leveraging the guide sourced by external sensors [25].

Although CNNs represent the preferred choice in terms of accuracy for both stereo and confidence estimation, often their hardware requirements overwhelm the resources available in many embedded systems.

3. Non-Local Anchoring

In this section, we introduce the proposed NLA framework, that given a disparity map \( D \) and a confidence map \( C \) encoding the uncertainty of each pixel (the higher the confidence, the better the assumed reliability), infers a completely dense and more accurate map. It starts by classifying each disparity point belonging to \( D \) in two categories: reliable and unreliable points, for short RP and UP respectively. In literature [31, 29], this task is accomplished by setting a threshold value \( \xi \) and considering as RP the points with a confidence value higher than \( \xi \). That is,

\[
RP = \{ p \in D, C(p) \geq \xi \} 
\]

consequently, the remaining ones are considered UP,

\[
UP = \{ p \in D, C(p) < \xi \}
\]

A new disparity map \( D' \) is then obtained by removing from \( D \) the UP set. The resulting \( D' \) map is characterized by a lower error rate, ideally 0, at the cost of a sparser distribution of pixels compared to \( D \).

Afterward, the full density of \( D' \) is restored by looking at reliable information within the RP set. To do so, given a pixel \( p \) and a 2D vector \( d \), we first define a subset of pixels \( P(p,d) \) as the path on which \( p \) lays according to the direction of \( d \):

\[
P(p,d) = \{ q \in D, \alpha \in N, q = p + \alpha d \}
\]

For a pixel \( u \in UP \), we define its anchor along direction \( d \), as the closest pixel to \( u \) laying on path \( P(u,d) \):

\[
a(u,d) = \{ v \in RP, \min_{v} |u - v| \}
\]

Given a set of paths on which \( u \) lays, a set \( A(u) \) of anchors will contribute to computing the new disparity value for such pixel. In particular, each anchor \( a \in A(u) \) spreads its disparity to \( u \), weighting it according to a similarity function between features \( I(u) \) and \( I(a) \) as follows:

\[
w(u,a) = \mathcal{G}(|I(u) - I(a)|) \cdot \mathcal{G}(|u - a|)
\]
The cues collected by each anchor $\mathcal{A}(u)$ are used to build a weighted histogram, on which each $w(u, a)$ increases the index corresponding to disparity hypothesis of pixel $a$. Finally, the weighted median is computed among the collected contributions:

$$D(u) = \min_k \sum_{i=0}^k w(u, a_i) \geq \frac{1}{2} \sum_{i=0}^n w(u, a_i) \quad (6)$$

We rely on a Gaussian function $G$ to encode the similarity between the unreliable pixel and one of its anchor points and on color intensity $\mathcal{I}(u)$ in the reference image. This strategy, coupled with the weighted median, enables edge-preserving disparity propagation. Figure 3 shows an example of anchoring for an unreliable pixel (red), receiving the contribution from a set of anchors (yellows).

Computational complexity for NLA is extremely low, as all the anchors of each unreliable pixel and their corresponding weights can be processed on a single image scan for each path in constant time. It only depends on the size of the image and the number of paths deployed for anchoring. It is worth observing that our proposal, conversely from other methods, is not constrained to a restricted area (i.e., local patches). Moreover, differently from recent methodologies exploiting confidence to improve stereo accuracy [31, 20, 23, 29], our framework acts on the disparity domain hence not requiring any information from the DSI thus enabling, for instance, its deployment with COTS devices.

Optionally, before replacing the unreliable pixel $u$ according to the outlined strategy, a further local aggregation step can improve the effectiveness of the information gathered from nearby points. This optional phase can be carried out by building a DSI with the $w(u, a_k)$ weights and filtering it according to the same similarity function $G$. This step enables the collection of additional contributions from nearby UP pixels $q_k$, which set of anchors $\mathcal{A}(q_k)$ is different from $\mathcal{A}(u)$.

### 4. Threshold-free RP selection

According to the description reported in Section 3, classifying the disparity values in UP and RP plays a key-role for NLA to achieve optimal performance. Thus, choosing the confidence threshold $\xi$ is of paramount importance. This strategy is common to other successful attempts to exploit confidence measures inside stereo algorithms [31, 29] or, in general, when we want to remove erroneous matches from the disparity map. For such tasks, proper tuning of the threshold $\xi$ is required to achieve the best results.

To address this issue, we propose a second level framework to effectively distinguish pixels into RP and UP according to features extracted from the confidence map without any manual tuning. To this aim, we fed to a random forest, trained in classification mode, the following local and global features computed from the confidence map:

- $C_p$, the confidence value for pixel $p$
- $\mu_{\mathcal{N}}(C_p)$, the average confidence computed on a local window $\mathcal{N}$, centered in $p$ and made of $\bar{N}$ pixels
  $$\mu_{\mathcal{N}}(C_p) = \frac{1}{\bar{N}} \sum_{q \in \mathcal{N}} C_q \quad (7)$$
- $\sigma_{\mathcal{N}}(C_p)$, the variance of confidence on a local window $\mathcal{N}$, centered in $p$ and made of $\bar{N}$ pixels
  $$\sigma_{\mathcal{N}}(C_p) = \frac{1}{\bar{N}} \sum_{q \in \mathcal{N}} [C_q - \mu_{\mathcal{N}}(C_p)]^2 \quad (8)$$
Table 1. Experimental results averaged on Middlebury v3 with BM algorithm. Best results are in bold.

<table>
<thead>
<tr>
<th>Stereo algorithm</th>
<th>bad 1(%)</th>
<th>bad 2(%)</th>
<th>RMSE</th>
<th>MAE</th>
<th>bad 1(%)</th>
<th>bad 2(%)</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>35.13</td>
<td>32.32</td>
<td>14.32</td>
<td>6.85</td>
<td>26.37</td>
<td>23.54</td>
<td>10.94</td>
<td>4.49</td>
</tr>
<tr>
<td>+ WMF [44]</td>
<td>26.22</td>
<td>22.92</td>
<td>10.08</td>
<td>4.18</td>
<td>17.22</td>
<td>13.91</td>
<td>5.56</td>
<td>2.00</td>
</tr>
<tr>
<td>+ WMF + GiF [40]</td>
<td>26.33</td>
<td>22.92</td>
<td>11.27</td>
<td>4.75</td>
<td>17.41</td>
<td>14.04</td>
<td>7.59</td>
<td>2.86</td>
</tr>
<tr>
<td>+ WMF + JBF [40]</td>
<td>28.03</td>
<td>24.93</td>
<td>10.95</td>
<td>4.68</td>
<td>18.86</td>
<td>15.75</td>
<td>6.87</td>
<td>2.44</td>
</tr>
<tr>
<td>+ LC [14]</td>
<td>24.23</td>
<td>20.00</td>
<td>11.68</td>
<td>4.74</td>
<td>16.30</td>
<td>12.23</td>
<td>7.93</td>
<td>2.65</td>
</tr>
<tr>
<td>+ NLA + O1</td>
<td>22.90</td>
<td>19.86</td>
<td>9.36</td>
<td>3.63</td>
<td>14.08</td>
<td>11.20</td>
<td>5.33</td>
<td>1.71</td>
</tr>
<tr>
<td>+ NLA + opt.</td>
<td>6.23</td>
<td>4.07</td>
<td>3.06</td>
<td>0.85</td>
<td>2.20</td>
<td>1.21</td>
<td>1.77</td>
<td>0.44</td>
</tr>
</tbody>
</table>

5. Experimental results

In this section, we evaluate the effectiveness of the proposed NLA framework with disparity maps obtained, on challenging datasets, by two stereo algorithms:

- Block Matching (BM), a local method computing matching costs on a $5 \times 5$ census transformed image [41], locally aggregated by a $5 \times 5$ box filter
- SGM [7], using as data term the normalized aggregated costs as BM algorithm and penalty parameters $P_1$ and $P_2$ tuned to 0.2 and 0.5

The choice was driven by the fast inference enabled by the two algorithms. Embedded stereo cameras with onboard processing (e.g., [1] or [15]) can run both BM and SGM at more than 30 FPS, sourcing disparity estimates in real-time with limited power consumption. In such a scenario, NLA can further improve the overall accuracy with low complexity, making it suited for embedded systems.

To exhaustively assess the effectiveness of our proposal, we compare it to state-of-the-art disparity refinement methods acting in the disparity domain. Moreover, since NLA relies on a confidence measure, we also compare it with recent methodologies exploiting confidence prediction to improve stereo accuracy [20, 23, 29] acting in the DSIF domain. We also evaluate for NLA the effect yielded by a different number of anchors and by the optional aggregation step outlined. Moreover, we validate the effectiveness of UP/RP selection module by reporting comparison with the manual optimal choice of the $\xi$ value by cross-validation. We evaluate all these aspects on the Middlebury v3 [28] training dataset and then we evaluate the effectiveness of the overall NLA framework also on KITTI 2015 [17].

5.1. Evaluation on Middlebury v3

In this section, we provide exhaustive experimental results concerning the full NLA framework (i.e., deploying the random forest for threshold-free RP selection and local aggregation step) and other refinement methods on the Middlebury v3 training dataset.

Comparison with other refinement strategies. In tables 1 and 2 we report results achieved by the following disparity refinement methods: fast bilateral solver (FBS [2]), median filter (MF [21]), weighted median filter (WMF [40]) and census transformed image (CTI [5]). In this paper, we evaluate the effect yielded by a different number of anchors and by the optional aggregation step outlined. Moreover, we validate the effectiveness of UP/RP selection module by comparing with the manual optimal choice of the $\xi$ value by cross-validation. We evaluate all these aspects on the Middlebury v3 training dataset and then we evaluate the effectiveness of the overall NLA framework also on KITTI 2015 [17].
Table 2. Experimental results averaged on Middlebury v3 with SGM [7] algorithm. Best results are in bold. Algorithms marked with * use the original confidence measure proposed in the paper.

<table>
<thead>
<tr>
<th>Stereo algorithm</th>
<th>All</th>
<th>Non-occ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM [7]</td>
<td>bad 1%</td>
<td>bad 2%</td>
</tr>
<tr>
<td>+ FBS [2]</td>
<td>25.06</td>
<td>21.55</td>
</tr>
<tr>
<td>+ WMF + GF [40]</td>
<td>22.22</td>
<td>19.56</td>
</tr>
<tr>
<td>+ WMF + JBF [40]</td>
<td>22.25</td>
<td>19.80</td>
</tr>
<tr>
<td>+ LRI + MF + BF [43]</td>
<td>22.01</td>
<td>17.68</td>
</tr>
<tr>
<td>+ LC [14]</td>
<td>20.39</td>
<td>16.60</td>
</tr>
<tr>
<td>+ Lev.stereo* [20]</td>
<td>22.22</td>
<td>19.52</td>
</tr>
<tr>
<td>+ PBCP* [29]</td>
<td>23.72</td>
<td>21.31</td>
</tr>
<tr>
<td>+ PBCP [29]</td>
<td>23.72</td>
<td>21.31</td>
</tr>
<tr>
<td>+ NLA + O1</td>
<td>18.68</td>
<td>15.44</td>
</tr>
<tr>
<td>+ NLA + opt.</td>
<td>7.72</td>
<td>5.18</td>
</tr>
</tbody>
</table>

Table 3. Experimental results on Middlebury v3 with SGM, comparing results obtained by NLA when using a threshold or the random forest classification of the RP. Best results in bold.

<table>
<thead>
<tr>
<th>Stereo algorithm</th>
<th>All</th>
<th>Non-occ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM [7]</td>
<td>bad 1%</td>
<td>bad 2%</td>
</tr>
<tr>
<td>+ NLA + O1 (ξ = 0.4)</td>
<td>18.90</td>
<td>15.86</td>
</tr>
<tr>
<td>+ NLA + O1 (ξ-less)</td>
<td>18.68</td>
<td>15.44</td>
</tr>
</tbody>
</table>

We followed implementation notes, hyper-parameters tuning and code provided by the authors [23], training on a subset of images from KITTI 2012 dataset (the first 20 images) as in [26]. Since the effectiveness of the confidence measure is crucial for our method, we also report in the final row the results achieved by NLA processing an optimal confidence measure, capable of ideally distinguishing between RP and UP. This represents the lower bound for the error rate with NLA. The automatic selection method proposed was trained on the 13 additional images available in Middlebury v3 dataset [28] for each of the two considered algorithms. Table 1 report the effectiveness of disparity refinement methods with the BM algorithm. We can notice how the proposed NLA outperforms all of the considered refinement methods. In particular, compared to the second best method LC, NLA is more effective by nearly 2% on both all pixels and non-occluded. The last row highlights how, if an ideal confidence measure is deployed, our framework is capable of reducing the error rate from over 35% of wrong pixels in the image to almost 6%. Table 2 shows the results with SGM [7]. Since our SGM implementation is based on BM algorithm to obtain the data term, we first highlight how the results obtained by processing maps by NLA are very similar (even better in this case) to those obtained by running SGM optimization on the entire DSI (without applying any
Figure 5. Qualitative results on Motorcycle stereo pair. First row: reference image and ground-truth disparity. Then, from top to bottom, disparity maps with overimposed bad1 rate and error maps for, respectively, SGM [7], SGM+Lev.stereo [20], Smart-SGM [23], SGM+PBCP [29] and SGM+NLA. All methods use O1 as confidence measure.

additional post-processing step, not deployed on our baseline SGM). This proves the effectiveness of our proposal when compared to more complex approaches such as SGM. Moreover, the DSI of the filtering map is not required with NLA, while SGM necessarily needs such information. In these experiments, we also deploy three additional methodologies relying on confidence measures to improve the results of SGM. The first one is a confidence-based modulation of the DSI carried-out before the SGM optimization, referred to as Lev.stereo [20]. The second one is a weighted sum of the contribution of the different scanlines, according to confidence, referred to as Smart-SGM [23]. The last one consists of a dynamic setting of the smoothness terms P1 and P2 according to confidence, referred to as PBCP [29]. We included them as representative state-of-the-art methodologies relying on confidence measures to improve the accuracy of stereo and we report results obtained when processing the confidence measures they were proposed with (marked with * in the table) as well as with the same one deployed by NLA for a fair comparison. We can observe how the NLA framework outperforms all of them, obtaining its best accuracy deploying the O1 measure. Moreover, our proposal works in the disparity domain, not requiring intermediate results from the SGM pipeline and it is thus a general-purpose technique suited for any stereo algorithm. Figure 5 shows a qualitative comparison between considered approaches and NLA.

Evaluation of RP selection. Once confirmed the superiority of the full NLA framework, in this section we inquire about the effectiveness of the threshold-free RP selection enabled by the random forest classifier. Table 3 shows comparison between the results achieved by the manually selected threshold through k-fold cross-validation, highlighting how the random forest selection strategy increases, on average, the accuracy of the refined disparity maps when considering all pixels, while it performs slightly worse on non-occluded pixels, thus mainly improving selection and refinement occluded regions.

Ablation studies on NLA and runtime. To better understand the key factors enabling for such improvements, we report results concerning the use of a different amount of anchors as well as without the optional local aggregation step, deployed during the previous evaluations. Figure 6 plots the error rate as a function of the number of anchors.
Table 4. Experimental results averaged on KITTI 2015 with BM and SGM algorithms. Best results are in bold.

<table>
<thead>
<tr>
<th>Stereo algorithm</th>
<th>bad 3% - BM</th>
<th>bad 3% - SGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>37.30</td>
<td>10.78</td>
</tr>
<tr>
<td>MF [21]</td>
<td>19.95</td>
<td>8.73</td>
</tr>
<tr>
<td>WMF [44]</td>
<td>21.03</td>
<td>8.81</td>
</tr>
<tr>
<td>LRI [43]</td>
<td>25.29</td>
<td>10.12</td>
</tr>
<tr>
<td>LRI +MF + BF [43]</td>
<td>18.90</td>
<td>9.11</td>
</tr>
<tr>
<td>+ Lev.stereo* [20]</td>
<td>-</td>
<td>10.10</td>
</tr>
<tr>
<td>+ Lev.stereo [20]</td>
<td>-</td>
<td>9.52</td>
</tr>
<tr>
<td>+ Smart-SGM [23]</td>
<td>-</td>
<td>8.47</td>
</tr>
<tr>
<td>+ PBCP* [29]</td>
<td>-</td>
<td>10.63</td>
</tr>
<tr>
<td>+ PBCP [29]</td>
<td>-</td>
<td>10.62</td>
</tr>
<tr>
<td>NLA + O1</td>
<td>11.42</td>
<td>7.68</td>
</tr>
</tbody>
</table>

Figure 7. Qualitative results on KITTI 2015 dataset [17]. From top to bottom, stereo pairs 085, 186 and 197. From left to right, reference frame and disparity maps from SGM [7] or refined with NLA.

6. Conclusions

In this paper, we proposed a fast, yet accurate, non-local disparity refinement technique based on confidence measures. It jointly enables the benefits of techniques acting in the disparity domain and the power of confidence measures extracted from the same domain. Conversely from other similar techniques, leveraging on confidence measures and designed for specific algorithms, our proposal acts outside the stereo pipeline, making it a general purpose alternative, hence totally agnostic to the stereo algorithm generating disparity maps. Experimental results on popular datasets confirmed the superiority of NLA compared to known techniques when dealing with disparity maps obtained from algorithms suited for deployment on embedded devices.

We compare our results with best methods MF, WMF and LC approaches. We can observe how, even on this very different dataset, the NLA framework can reduce the error rate of the raw disparity maps by nearly 26% (BM) and by more than 3% (SGM), notably outperforming the other refinement techniques. Since the scene contents depicted by KITTI 2015 are more smooth compared to indoor scenes considered before (e.g., large road planes), the smoothing constraint enforced by SGM is stronger than the non-local refinement processed by NLA, being nonetheless capable of reaching with BM a comparable degree of accuracy with significantly lower computational efforts. Focusing entirely on SGM results, we report, as for the Middlebury v3 evaluation, the improvements yielded by state-of-the-art confidence-based cost modulations proposed in [20, 23, 29]. Similarly to Middlebury v3, we evaluated the three previous strategies with their originally proposed confidence measures as well as with the same plugged into NLA for a fair comparison. The trend previously highlighted is confirmed on KITTI 2015 as well. Figure 7 shows additional qualitative results on KITTI 2015, comparing raw disparity maps by BM and SGM with those refined through NLA.
References


[29] A. Seki and M. Pollefeys. Patch based confidence prediction for dense disparity map. In British Machine Vision Conference (BMVC), 2016. 2, 3, 4, 5, 6, 7, 8


