Fast hierarchical cost volume aggregation for stereo-matching

Sergey Smirnov, Atanas Gotchev
Tampere University of Technology
firstname.secondname@tut.fi

***Abstract*—Some of the best performing local stereo-matching approaches use cross-bilateral filters for proper cost aggregation. The recent attempts have been directed toward efficient approximations of such filter aimed at higher speed. In this paper, we suggest a simple yet efficient coarse-to-fine cost volume aggregation scheme, which employs pyramidal decomposition of the cost volume followed by edge-avoiding reconstruction and aggregation. The scheme substantially reduces the computational complexity while providing fair quality of the estimated disparity maps compared to other approximated bilateral filtering schemes. In fact, the speed of the proposed technique is comparable with the speed of fixed kernel aggregation implemented through integral images.**

I. Introduction

Stereo-matching is one of the most important topics in the computer vision field as it allows estimating scene depth using a passive stereo camera. In order to estimate a dense depth map, four steps are usually performed: computation of a cost volume, local or global cost aggregation, disparity computation and disparity refinement [1]. The resulting disparity is then converted to a depth map. The most computationally expensive step in this scheme is the *cost volume aggregation*, since the empirical cost function values are noisy and should be refined in order to obtain their robust estimates.

Local aggregation is usually expressed by a local filtering operator (e.g. weighted averaging), working in some spatial neighborhood. While generally being faster, some recently proposed local methods match or even outperform global aggregation methods, such as Graph Cuts or Belief Propagation. However, their complexity is still high due to high resolution of camera sensors. Complexity of well-performing local aggregation methods, such as based on bilateral filtering [2], directly depends on the selected kernel size. Large kernels are preferred for handling textureless zones, while smaller kernels are preferred for efficient aggregation.

Recently proposed local stereo-matching techniques use low-complexity bilateral filter approximations [3], [4], [5] or non-local aggregation based on the weighted tree-structures [6] in order to remove the complexity dependence on the kernel size. However their absolute computational time is still relatively high, due to high amount of spatial-domain processing performed at each step of the algorithm.

Fig. 1. Edge-preserving property of the proposed approach: (left) reference image with marked Dirac-impulse positions and (right) their responses encoded with distinctive colors.

Another way to deal with low-confident zones in stereo pairs, is to address them in a hierarchical manner. A number of coarse-to-fine approaches have been proposed for solving the textureless problem as well as for reducing the complexity [7], [8]. Based on the hypothesis that successful match for the low confident area can be found on the *decimated stereo pair*, those approaches try to propagate depth estimates for the problematic areas from the coarser pyramid levels. However, due to modified frequency content in the decimated images as well as modified disparity range, such propagation is tricky and prone to errors.

In contrast to the above techniques, cost volume aggregation performed in a coarse-to-fine manner, can avoid errors related with disparity range shrinkage, and deliver acceptable performance. The method proposed in this paper employs such an approach, where the cost volume aggregation has low complexity in both asymptotic and absolute run-time senses, as it only consists of number of weighted decimations and upsamplings. Figure 1 illustrates the effective support of resulting kernels, which adapt nicely to the image discontinuities. The method makes use of the concept of edge-avoiding wavelets [9], while being specifically tailored for the stereo-matching case.

II. Hierarchical aggregation

Cost volume is usually defined as a 3D structure, where each slice represent dissimilarity between input images at a particular disparity (e.g. shift) [1]:

$$C(x_1, x_2, d) = \|L(x_1, x_2) - R(x_1 - d, x_2)\|_1, \qquad (1)$$

where L and R represent rectified input images as three-component color vectors, and L_1 norm corresponds to dissimilarity measured between color components.

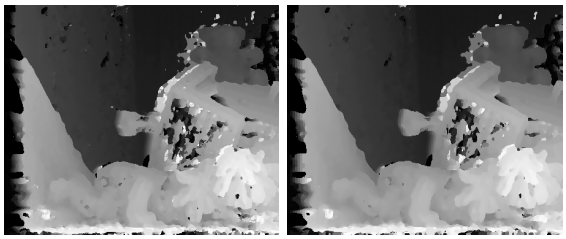

Fig. 2. Disparity estimate for "Teddy" dataset: (left) Gaussian aggregation; (right) simple hierarchical aggregation.

Consider a slice $C(\cdot)$, where local aggregation can be defined on some spatial neighborhood $\Omega_{\mathbf{x}}$ of the processed pixel $\mathbf{x} = (x_1, x_2)$:

$$\tilde{C}(\mathbf{x}) = \Sigma_{\mathbf{p}\in\Omega_{\mathbf{x}}} W(\mathbf{x}, \mathbf{p}) \cdot C(\mathbf{p}), \tag{2}$$

where $\tilde{C}(\cdot)$ is the filtered signal and $W(\mathbf{x}, \mathbf{p})$ is a normalized weighting kernel.

Due to lack of validity in large textureless areas, the local filtering kernel must have adequate support size in order to successfully handle such areas. However, due to the spatially-variant nature of the commonly used filters, the complexity of the direct implementation is dependent on the size of the kernel $\Omega_{\mathbf{x}}$. For instance, this is the case of adaptive support weighting [2], where the filtering kernel corresponds to the cross-bilateral filter:

$$W(\mathbf{x}, \mathbf{p}) = \frac{e^{-\frac{\|I(\mathbf{x}) - I(\mathbf{p})\|}{\sigma_r}} \cdot e^{-\frac{\|\mathbf{x}-\mathbf{p}\|}{\sigma_s}}}{\Sigma_{\mathbf{u}\in\Omega_{\mathbf{x}}} e^{-\frac{\|I(\mathbf{x}) - I(\mathbf{u})\|}{\sigma_r}} \cdot e^{-\frac{\|\mathbf{x}-\mathbf{u}\|}{\sigma_s}}}, \tag{3}$$

where σ_r and σ_s are adjusting parameters of the filter and I is the associated color image.

Decimation of the cost volume along spatial axes by some pre-defined factor f followed by its' backward up-sampling can be regarded as aggregation with some spatial kernel

$$C^{i+1} = C^i \downarrow_f, \tag{4}$$

$$\tilde{C}^i = C^{i+1} \uparrow_f . \tag{5}$$

A. Edge-avoiding aggregation

The resulting filtering, while not shift-invariant across the image, is smooth and approximately isotropic. The strength of the aggregation can be adjusted either by changing the decimation factor f or by performing additional decimation/up-sampling steps in a coarse-to-fine manner. In contrast to the decimation of the input stereo pair, usually applied in coarse-to-fine stereo approaches [7], [8], down-sampling of the original cost volume does not affect the disparity range and thus the coarse cost volume is fully compatible with the finer one.

In terms of quality such method resembles aggregation with Gaussian kernel. Figure 2 shows disparity estimates for "Teddy" dataset [10] aggregated with a large Gaussian kernel and with a three-level pyramidal decomposition/reconstruction procedure.

The basic mechanism of decimation/reconstruction of the cost volume results in over-smoothing of strong depth discontinuities, similarly to the effect of applying spatially-invariant filtering kernels (c.f. Figure 2). A color image pyramid calculated along with the cost volume pyramid can be utilized in order to penalize aggregation across strong color boundaries and to support edge-preserving hierarchical aggregation.

Consider the Gaussian image pyramid, constructed for the color image, associated with the cost volume (hereafter we omit the decimation factor f):

$$I^{i+1} = I^i \downarrow, \tag{6}$$

$$\tilde{I}^i = I^{i+1} \uparrow, \tag{7}$$

where I^i is an input image at i-th pyramid level and $\tilde{I}^i$ is the same image reconstructed from the coarser level by up-sampling. In Laplacian pyramid fashion, residual image can be also constructed for each layer of the pyramid:

$$\Delta^i(\mathbf{x}) = I^i(\mathbf{x}) - \tilde{I}^i(\mathbf{x}). \tag{8}$$

In the textureless zone, the reconstructed image $\tilde{I}^i$ will be very close to the original image I^i and hence the residual image Δ^i will predominantly contain small values. One can construct the aggregation mechanism at level i imposing some edge-preserving properties:

$$\hat{C}^i(\mathbf{x}) = W^i(\mathbf{x}) \cdot \tilde{C}^i(\mathbf{x}) + (1 - W^i(\mathbf{x})) \cdot C^i(\mathbf{x}), \tag{9}$$

where $\hat{C}^i(\mathbf{x})$ is the resulting aggregated cost at level i, $W(\mathbf{x})$ is a range-weighting component $W^i(\mathbf{x}) = G(\Delta^i(\mathbf{x}), \sigma)$, and the weighting function G can be defined in the bilateral style:

$$G(\mathbf{y}, \sigma) = e^{-\frac{\|\mathbf{y}\|}{\sigma}}. \tag{10}$$

For the case of highly textured areas, the decimation will result in substantial smoothing. Subsequently, the details signal $\Delta^i(\mathbf{x})$ will be strongly penalizing the aggregation.

In order to complete the proposed scheme, the backward cost up-sampling from level $i+1$ to level i should use already aggregated cost at $i+1$:

$$\tilde{C}^i = \hat{C}^{i+1} \uparrow . \tag{11}$$

B. Separable wavelets

Separable wavelets [9] are favoured for cost volume aggregation, described in the Section II-A because of possibilities they offer to parallelize the decomposition.

Considering only a 1-D image I and 1-D signal C and for the case of $f = 2$, the decimation can be formulated in an edge-avoiding manner [9]:

$$I^{i+1}(x) = \frac{I^i(2x) + w_1 \cdot I^i(2x-1) + w_2 \cdot I^i(2x+1)}{1 + w_1 + w_2}, \tag{12}$$

$$C^{i+1}(x) = \frac{C^i(2x) + w_1 \cdot C^i(2x-1) + w_2 \cdot C^i(2x+1)}{1 + w_1 + w_2}, \tag{13}$$

where x is a pixel coordinate at the decimated image, and w_1 and w_2 are weights calculated based on the difference between the corresponding color values:

$$w_1 = G(I^i(x) - I^i(x-1), \sigma_d), \tag{14}$$

$$w_2 = G(I^i(x) - I^i(x+1), \sigma_d), \tag{15}$$

where σ_d controls the amount of kept details.

Note, that since we do not need to keep details at each decomposition step, the three-steps lifting scheme *split-predict-update* used in the [9] can be replaced by a single operation.

C. 1-D hierarchical aggregation

When considering pixels near strong color edges, one can see a drawback of the initial scheme (Eq. 4 - 11). Due to penalization on the finer levels (Eq. 9), those pixels may remain not aggregated, even if they are associated with large textureless areas. In [5], a direct bilateral filter has been used in order to smooth out cost values within small support window and thus avoid penalization of the boundary pixels. The proposed 1-D interpolation scheme resolves the problem using small overlaps during the decimation and the up-sampling steps. 1-D decimation (Eq. 12) uses three pixels at the finer resolution to obtain value of one pixel at the coarser level, for decimation factor equal to two. Backward step, which combines aggregation with up-sampling can also be re-defined with some overlap:

$$\tilde{C}^i(2x) = \frac{w \cdot \tilde{C}^{i+1}(x) + w_* \cdot C^i(2x)}{w + w_*} \tag{16}$$

$$\tilde{C}^i(2x+1) = \tag{17}$$

$$= \frac{w_L \cdot \tilde{C}^{i+1}(x) + w_R \cdot \tilde{C}^{i+1}(x+1) + w_+ C^i(2x+1)}{w_L + w_R + w_+},$$

w, w_L and w_R are defined using color difference between image at the current level i and the coarser one at the level $i+1$:

$$w = G(I^i(2x) - I^{i+1}(x), \sigma_c^l), \tag{18}$$

$$w_L = G(I^i(2x+1) - I^{i+1}(x), \sigma_c^l), \tag{19}$$

$$w_R = G(I^i(2x+1) - I^{i+1}(x+1), \sigma_c^l). \tag{20}$$

The weights w_* and w_+ can be defined similarly to Eq. 9, however such weighting can be prone to errors caused by noise. We introduce an additional parameter γ, $\gamma \in (0; 1]$ to regulate the shape of the impulse responses:

$$w_* = min(\gamma, 1 - w_1), \tag{21}$$

$$w_+ = min(\gamma, 1 - max(w_1, w_2)), \tag{22}$$

where $\gamma = 1$ corresponds to the weighting approach with significant spikes in the actually generated filtering kernels. A decreased γ value results in smoother aggregation, making kernels similar to the bilateral ones.

Another approach to overcome noise is to apply level-dependent aggregation, controlled by the parameter σ_c^l, where l is the number of the current level and N is the total number of levels:

$$\sigma_c^l = \sigma_c \cdot \left(\frac{l}{N}\right)^{-1}. \tag{23}$$

Figure 1 shows impulse responses for number of Dirac-impulses placed at interesting points in the cost volume and aggregated with the proposed coarse-to-fine scheme. Each impulse response is encoded with distinctive color. The figure illustrates the adaptive and edge-preserving performance of the algorithm.

D. Implementation and complexity

To correctly estimate the actual complexity of the proposed approach one should carefully consider the implementation details. As the color-weighting is intensively used during both decimation and aggregation steps, their direct calculation can significantly degrade the computational speed. Pre-calculated lookup tables can be used in order to avoid unnecessary weights calculations. In such a way, single decimation/up-sampling step has to perform $2 \cdot X$ weighted averaging operations, where X is the original resolution of the color image. The overall computational time Y can be estimated considering number of levels N in the decomposition:

$$Y = 2 \cdot \sum_{l=1}^{N} \frac{X}{(2)^l}. \tag{24}$$

The overall computational time tends to $4 \cdot X$ when $N \to \infty$, which is still linear.

We have implemented the proposed approach on MS Visual C++ using Matlab/MEX interface to obtain compiled MEX functions, directly accessible from the Matlab environment. We used number of optimizations, such as OpenMP multi-processing and look-up tables for the weight maps.

III. Results

The proposed approach addresses the problem of the cost volume aggregation, which is one of the steps in the general stereo-matching method. It, however, can be extended to the complete stereo-matching solution by introducing additional post-processing step. Such post-processing can be done in the same manner as proposed in [4]. Initial aggregation of the left and the right cost volumes, followed by left-to-right correspondence check are used to obtain an initial disparity estimate along with the confidence map. Another cost volume, synthesized in the probabilistic manner from the initial disparity is aggregated once again to obtain the final disparity estimate. The same technique can be adjusted to the proposed approach using the same algorithm parameters during post-processing as those, used during the initial aggregation step.

Table I presents the performance of the proposed local stereo approach tested on the Middlebury stereo datasets [1]. Disparity error in three different zones: non-occluded areas (nonocc), whole image (all) and near discontinuities (disc) is estimated for each of four stereo pairs *Tsukuba*, *Venus*, *Teddy* and *Cones*. Disparity results of the first aggregation step

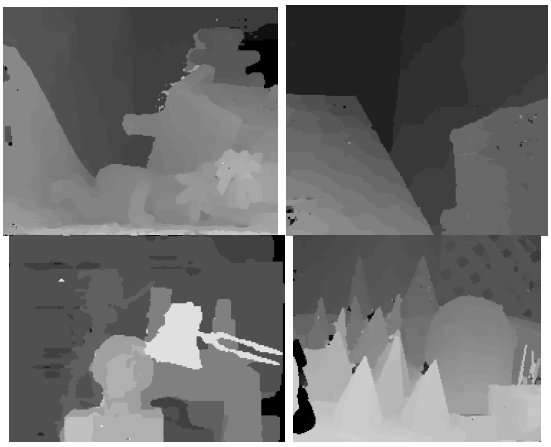

Fig. 3. Raw aggregation results for Midddlebury stereo evaluation [1] datasets.

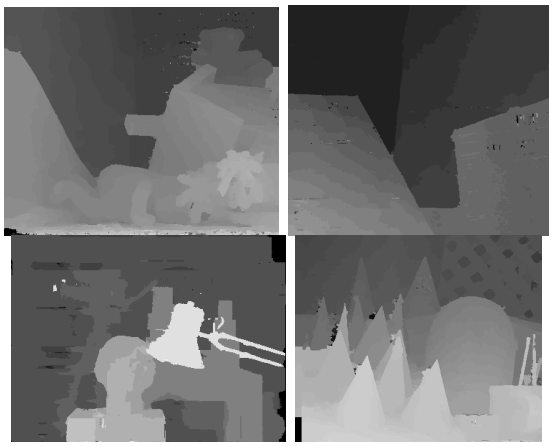

Fig. 4. Post-filtered results for Midddlebury stereo evaluation [1] datasets.

(without post-processing) (marked **(A)** and visualized in Figure 3) have fair quality acceptable for number of applications. However, results obtained with the post-processing (marked **(B)** and visualized in Figure 4 illustrate significantly improved quality in both numerical and visual senses.

We compare our approach with the adaptive coarse-to-fine (ACTF) approach, proposed by Sizintsev [7]. The latter has inferior performance, as seen in Fig. 8 in [7], while having comparable running time (approaches A1-A4 from the [7]). The running time provided in [7] cannot be directly used to compare our approaches due to different hardware configurations, while still may imply approximate equivalence of the computational complexity.

The actual running time of our proposed approach averaged across several runs is 2.61 seconds for all four Middlebury datasets for the single aggregation step and 6.12 seconds for obtaining fully post-processed results (including all other required steps). Average run-time of the fixed window aggregation, implemented with integral images on the same PC and implemented in similar manner as a compiled MEX is 1.02 sec. However, since our MEX implementation re-calculates lookup tables at each run and uses intensive memory allocations, the complexity difference can be further minimized by using standalone implementation written as pure C/C++ code.

	Tsukuba			**Venus**		
	nonocc	all	disc	nonocc	all	disc
(A)	2.84	4.20	12.6	2.48	3.28	10.0
(B)	1.41	2.02	5.85	2.58	3.10	7.72
[ACTF]	6.5	8.0	16.5	4.0	5.0	15.0
	Teddy			**Cones**		
	nonocc	all	disc	nonocc	all	disc
(A)	11.7	16.7	22.8	6.15	13.0	14.0
(B)	8.52	13.6	18.2	3.75	10.1	9.41
[ACTF]	8.0	15.5	20.5	5.5	14.0	14.0

TABLE I
RESULTS OF THE PROPOSED APPROACH ON THE MIDDLEBURY DATASETS; **(A)** SINGLE AGGREGATION, **(B)** POST-PROCESSED. LINE MARKED **[ACTF]** APPROXIMATE ADAPTIVE COARSE-TO-FINE RESULTS FROM [7].

IV. CONCLUSIONS

We have proposed new hierarchical cost volume aggregation scheme, based on the edge-avoiding wavelet decomposition. While it shares same ideas with basic coarse-to-fine stereo-matching methods, it successfully solves their main drawbacks by constructing a pyramid out of the 3D cost volume rather than out of the image pairs. Disparity propagation from coarser levels goes naturally by up-sampling of the coarse cost volume, while in classical coarse-to-fine methods some complicated non-linear fusion approaches are required. The proposed method have demonstrated a fair performance while having low computational complexity. It also bears potential for further improvements.

V. ACKNOWLEDGEMENT

This work has been supported by a collaborative project with Nokia Research Center co-funded by the Finnish Funding Agency for Technology and Innovation (TEKES).

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 7, pp. 7–42, April-June 2002.

[2] K.-J. Yoon and I.-S. Kweon, "Adaptive support-weight approach for correspondence search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 650–656, 2006.

[3] J. S. Kaiming He and X. Tang, "Guided image filtering," in *ECCV*, 2010, p. 114.

[4] Q. Yang, "Recursive bilateral filtering," in *ECCV*, 2012, pp. 399–413.

[5] ——, "Hardware-efficient bilateral filtering for stereo matching," *PAMI*, no. 99, 2013.

[6] ——, "A non-local cost aggregation method for stereo matching," in *CVPR*, 2012, pp. 1402–1409.

[7] M. Sizintsev and R. P. Wildes, "Coarse-to-fine stereo vision with accurate 3d boundaries," *Image and Vision Computing*, vol. 28, no. 3, pp. 352 – 366, 2010.

[8] P. F.-G. Yi-Hung Jen, Enrique Dunn and J.-M. Frahm, "Adaptive scale selection for hierarchical stereo," in *Proc. BMVC*, 2011, pp. 95.1–95.10.

[9] R. Fattal, "Edge-avoiding wavelets and their applications," *ACM Transactions on Graphics*, 2009.

[10] D. Szeliski and R. Scharstein, "High-accuracy stereo depth maps using structured light," in *Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE Computer Society, 2003, pp. 195–202.