Fusion of Range and Stereo Data for High-Resolution Scene-Modeling

Georgios Evangelidis, Miles Hansard, Radu Horaud

Fusion of Range and Stereo Data for High-Resolution Scene-Modeling

Georgios D. Evangelidis, Miles Hansard, and Radu Horaud

***Abstract*—This paper addresses the problem of range-stereo fusion, for the construction of high-resolution depth maps. In particular, we combine low-resolution depth data with high-resolution stereo data, in a maximum a posteriori (MAP) formulation. Unlike existing schemes that build on MRF optimizers, we infer the disparity map from a series of local energy minimization problems that are solved hierarchically, by growing sparse initial disparities obtained from the depth data. The accuracy of the method is not compromised, owing to three properties of the data-term in the energy function. Firstly, it incorporates a new correlation function that is capable of providing refined correlations and disparities, via subpixel correction. Secondly, the correlation scores rely on an adaptive cost aggregation step, based on the depth data. Thirdly, the stereo and depth likelihoods are adaptively fused, based on the scene texture and camera geometry. These properties lead to a more selective growing process which, unlike previous seed-growing methods, avoids the tendency to propagate incorrect disparities. The proposed method gives rise to an intrinsically efficient algorithm, which runs at 3FPS on 2.0MP images on a standard desktop computer. The strong performance of the new method is established both by quantitative comparisons with state-of-the-art methods, and by qualitative comparisons using real depth-stereo data-sets.**

***Index Terms*—Stereo, range data, time-of-flight camera, sensor fusion, maximum a posteriori, seed-growing.**

I. Introduction

Many computer vision methodologies, including dense 3D reconstruction [1], [2], gesture recognition [3], [4], and object detection [5] have benefited from recently-developed depth sensors. These sensors rely on active-light principles, including modulated-light and pulsed-light cameras, commonly denoted time-of-flight (TOF) [6], [7], or projected-pattern triangulation cameras [8]. Regardless of the working principle, however, these sensors provide low-resolution (LR) or mid-resolution depth maps that are inadequate for a number of applications such as 3DTV and film production. For example, many tasks in the film production industry greatly benefit from a high-resolution (HR) and high-quality depth map [9].

While HR depth maps can be obtained from multiple-view matching and reconstruction using standard color cameras, it is well known that stereo matching is problematic when the scene contains weakly textured areas, repetitive patterns, or occlusions; these situations are very common in both indoor and outdoor environments. Active-light sensors do not suffer from these limitations, although their own depth data are quite noisy in the presence of scattering, non-Lambertian materials, and slanted surfaces. The complementary nature of HR stereo and LR depth sensors leads to the design of *mixed* camera systems [10]–[18], which seem to be the most promising approach, at present, for high-quality 3D depth maps.

In this context, this paper addresses the problem of HR 3D reconstruction from the combination of a photometric camera pair and an active-light camera, provided that the multiple-camera setup is calibrated [19]. The combination of a stereo matching algorithm and of an active-light sensor raises the central question of devising a matching algorithm with the following features: (i) it considerably increases the resolution of the depth data, e.g., by a factor of ten, (ii) it eliminates depth-sensor errors wherever possible, (iii) it overcomes the limitations of stereo algorithms in textureless areas, and (iv) it is able to compete with a depth sensor in terms of speed. Hence, the availability of an efficient and robust stereo algorithm that takes advantage of LR depth sensors and that provides dense and accurate HR depth maps, possibly with subpixel accuracy, is particularly desirable.

To this end, we propose a 3D reconstruction method that merges depth-sensor measurements with photo-consistency stereo matching. We address the problem from the perspective of seed-growing, starting from a small number of *control points* whose disparities are then propagated to yield a dense disparity map. We show that this can be cast into maximum a posteriori (MAP) formulation (Sec. III), which leads, in turn, to a series of local optimization problems that are solved hierarchically by a novel region-growing process (Sec. V). While the proposed method may not reach the global optimum, it allows us to devise an intrinsically efficient methodology that bridges the gap between global optimizers based on Markov random fields (MRF) and locally-optimal winner-take-all (WTA) strategies (Sec V-E).

Efficient stereo-only or stereo-depth fusion methods rely on control points, by exploiting either feature correspondences in stereo [20], [21] or depth data in fusion [14], [15], and they assume that these points are of very good quality. A key contribution of this paper is that this requirement is relaxed in order to devise a method tolerant to bad control points. We propose to truly combine LR depth-sensor data with HR rich photometric information, whenever and wherever possible, showing that fusion is helpful, even in the early stage of depth initialization (Sec. IV-A & IV-B). The data term of the proposed MAP formulation benefits from a new cross-

This work has received funding from Agence Nationale de la Recherche under the MIXCAM project number ANR-13-BS02-0010-01.

G. D. Evangelidis and R. Horaud are with Perception Team, INRIA Grenoble Rhône-Alpes, 655, avenue de l'Europe, 38330 Montbonnot Saint-Martin, France, email:georgios.evangelidis@inria.fr, radu.horaud@inria.fr

M. Hansard is with the Vision Group, School of Electronic Engineering and Computer Science, Queen Mary, University of London, Mile End Road, London E1 4NS, UK, e-mail:miles.hansard@qmul.ac.uk

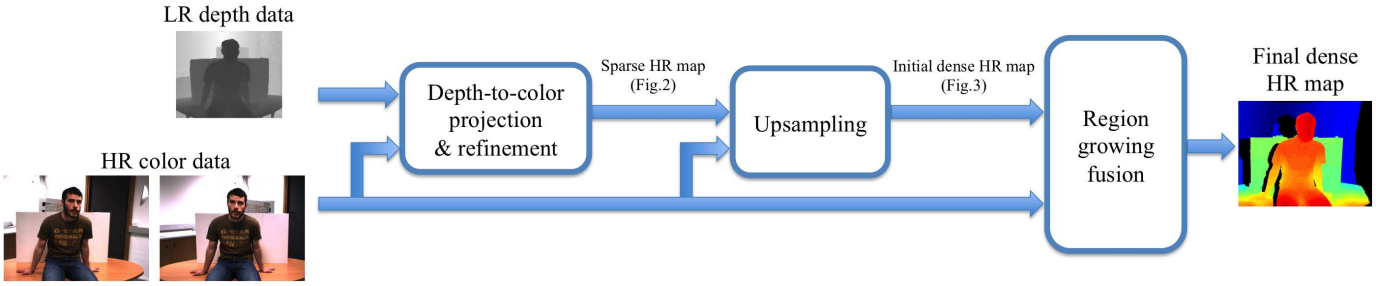


Fig. 1: The pipeline of the proposed depth-stereo fusion method. The low-resolution (LR) depth data are projected onto the color data and refined to yield a high-resolution (HR) sparse disparity map. Starting from these disparity *seeds*, an upsampling process provides an initial HR dense disparity map. Both the HR seeds and the initial dense disparity map are then used by the region-growing depth-stereo fusion to produce the final HR depth map. A prominent feature of our method is that fusion takes place at several data processing stages.

correlation function, which provides real-valued disparities via subpixel corrections computed in closed-form (Sec.V-B1), and takes advantage of a depth-guided cross-product aggregation (Sec. V-B2). Moreover, the data-term adaptively merges the stereo- and depth-consistency terms guided by the scene texture and the camera geometry (Sec. V-C). These advantages lead to a more selective growing-of-disparities process, thus preventing the algorithm from propagating erroneous depth-sensor data – a phenomenon that is often associated with propagation techniques – as experimentally verified (Sec. VI). It is important to note that the proposed method can be used 'as is' with any depth sensor, as it requires neither sensor-dependent confidence maps nor a sensor-dependent model. The proposed fusion pipeline is illustrated in Fig. 1.

Supplementary materials, in particular image datasets and Matlab code are available online.[1]

II. Related Work

We review pure stereo matching methods with emphasis on local algorithms, owing to their computational suitability for HR images. We also review upsampling methods and depth-stereo fusion methods. More detailed surveys of stereo matching algorithms and depth-stereo fusion methods can be found in [22] and [9], respectively.

A. Pure Stereo Matching

Stereo matching methods can be broadly classified into global and local [22]. Global algorithms [23] typically adopt an MRF formulation and solve a single optimization problem based on a MAP criterion. Despite their superiority over local methods, global algorithms are extremely time-consuming, and hence unattractive for fast fusing of depth-sensor data with high-resolution images. Local algorithms solve per-pixel optimization problems and the state-of-the-art methods build on adaptive cost aggregation [24]–[27]. Most methods, however, must visit the entire cost volume to find an optimal disparity value at each pixel. This volume grows rapidly with respect to the input, as the width of the image typically multiplies the number of disparities. Therefore, although they are able to provide LR disparity maps in real-time, they remain slow and subject to memory issues in HR stereo. Note that global algorithms need several approximations to obtain LR disparity maps in real-time [28].

More interestingly, algorithms that rely on control-point correspondences [29] are drastically more efficient since they avoid visiting the whole cost volume. Region-growing approaches start from reliable but sparse correspondences (seeds) and propagate them in textured areas [20]. [30] suggests a similar propagation scheme where orientation-consistent disparities are propagated to neighbors at the cost of finding a plane equation per pixel. [21] proposed a generative model, where the prior disparity comes from a 2D triangulated mesh whose vertices (control points) are obtained from matches between low-level features. As with [20], textureless areas remain intractable and the final map is reliable only when the matches are dense and uniformly distributed over the images. Notice that the proposed method relies on the idea of control points that are transferred from a depth sensor, thus avoiding the limitation owing to untextured areas.

To obtain continuous disparities that are required in many scenarios, e.g., 3D reconstruction [31], local stereo algorithms typically employ two strategies: (i) fitting a curve around the correlation peak [14], [22] or (ii) integrating an intensity interpolation kernel into the (dis)similarity function [32] whose optimization leads to subpixel correction. The latter is also the case in the fusion framework of [16], [17] that inherently takes advantage of inter-pixel depth estimations.

B. Depth Upsampling and Depth-Stereo Fusion

Any prior depth information, even at low resolution, is likely to help dense disparity estimation. Apart from a naive interpolation, the bilateral filter [33] can post-process an interpolated map using color HR images for guidance [31]. Alternatively, a joint bilateral filter applies spatial and range kernels to the LR depth map and HR color image respectively, so that upsampling is a by-product of filtering [34]. Upsampling methods, however, are limited to clean depth LR data and cannot reconstruct accurate HR maps when LR data are delivered by depth sensors.

[1] https://team.inria.fr/perception/research/dsfusion/

The above limitation of the upsampling methods gives rise to fusion approaches that can merge depth and stereo data in either early or late stages, once the sensors are calibrated. Late fusion suggests merging two depth maps, one obtained with stereo and one from the range sensor, possibly upsampled [18], [35]. The majority of fusion methods, however, merge the stereo and range-sensor data at an earlier processing level. [10] estimates a LR TOF-based disparity map, which initializes a coarse-to-fine stereo scheme [36]. A semi-global dynamic programming approach is followed in [14] with the TOF-based disparities, wherever available, being considered as error-free matches. In a global framework, [11] produces mid-resolution depth maps by merging TOF and stereo data within a graph-cut stereo-matching paradigm; each energy term exploits both modalities. Likewise, MRF-based formulations have been proposed [12], [13], [37]. In [37], ground control points reflect an extra regularization term in the MRF energy function. The work described in [12] uses an MRF scheme to merge depth distributions of each sensor alone, but the goal is a LR depth map. [13] extends [12] by means of weighted fusion. A balanced fusion (based on several confidence maps) within a total variation model is also proposed in [17]. A similar variational model that infers the HR map in a coarse-to-fine manner is adopted by [16]. Note that, unlike most fusion methods and similar to the proposed one, [16] is not tuned to a specific depth sensor.

More closely to the present work, [15] fuses the data within the seed-growing method of [20]. In particular, TOF-based disparities constitute seeds while a triangulation-based interpolated TOF map regularizes the seed-growing process. When the TOF data are noisy, however, this approach tends to produce incorrect disparities, and to propagate false positives during the growing process. The proposed method differs considerably from [15] in terms of depth initialization, cost function, and fusion strategy. The proposed initial map is robust to depth discontinuities while it also guides the cost aggregation inside a window. Moreover, our likelihood term integrates functions that are capable of providing sub-pixel disparity corrections [32]. This turns out to be very beneficial, not only for the continuous nature of the final map, but also for the growing process itself, thus propagating more reliable messages (disparities). Note that the subpixel disparity correction is obtained from a closed-form solution – an interesting feature for efficiency. Finally, our algorithm benefits from an adaptive fusion scheme that better balances the contribution of each modality (depth or color), and that results in fewer unmatched pixels.

III. Problem Formulation

The main mathematical notations that are used throughout the paper are summarized in Table I. As discussed, the direct upsampling of LR depth data suffers from limitations, in particular when the upsampling factor is high.[2] Therefore, our goal is to build D by jointly taking advantage of both sensing modalities. Given a proper calibration, (e.g., [19], [38]), the mapping of the LR depth image onto the rectified HR color images will typically yield a *sparse* disparity map $D\downarrow$, which is almost evenly distributed across the HR grid. Note that in [20], [21], the initial matches between the stereo images do not correspond to a uniform sparse version of D, as they are unpredictably distributed, due to the reliability of properly detecting interest points in images. Instead, the sparse map obtained with a depth sensor can be used to guide a stereo algorithm, *regardless of the presence or absence of scene texture*.

[2]In our experiments, the upsampling factor is $10\times$ in each dimension, that is $100\times$ in the number of pixels, e.g., from 0.02Mp to 2MP.

TABLE I: The main mathematical notations used in the paper.

p, q:	Pixel locations of the high-resolution grid
$p\downarrow$, $q\downarrow$:	Pixel locations of the sparse grid
d_p:	Unknown disparity of pixel p, initialized by d_p^0
$d_{p\downarrow}$:	Known disparity of pixel $p\downarrow$ (observed)
D:	Unknown HR disparity map, initialized by D^0
$D\downarrow$:	Known sparse version of D (observed)
$\mathcal{D}, \mathcal{D}^0, \mathcal{D}\downarrow$:	Sets of all random variables (disparities) associated with D, D^0 and $D\downarrow$ respectively, with $d_p \in \mathcal{D}$, $d_p^0 \in \mathcal{D}^0$ and $d_{p\downarrow} \in \mathcal{D}\downarrow$
t_p:	Subpixel disparity correction of p
$\mu_p = (d_p; t_p)$:	Disparity-correction pair referred here to as *meta-disparity* with $\lvert t_p\rvert < 1$
$\mathcal{M} = \{\mu_p\}$:	Set of meta-disparities
$\mathcal{S} = \{s_p\}$:	Set of observed stereo pixel intensities
I_p:	Intensity of pixel p
$\mathbf{u}(x, y)$:	a vectorized (zero-mean) form of an intensity window centered at the 2D position (x, y)
$\mathcal{N}_p$:	Neighborhood of pixel p
med:	2-d median operator
I_R, D_R:	Intensity and disparity maps defined on a sub-region R
$E_S(d_p)$, $E_D(d_p)$:	Stereo-based and depth-based energy of p for given disparity
$g(\xi;\gamma) = e^{-\lvert\xi\rvert/\gamma}$:	Exponential mapping of ξ with scale γ

We propose to model the estimation of $\mathcal{D}$, and therefore the map D, as a *maximum a posteriori* (MAP) problem, based on the available depth and stereo observations. However, instead of immediately using $\mathcal{D}\downarrow$, we first estimate a dense initial map D^0 and its associated set $\mathcal{D}^0$. Then, we obtain the final disparity map by solving the following optimization problem:

$$\mathcal{D}^* = \arg\max_{\mathcal{D}} P(\mathcal{D}|\mathcal{S}, \mathcal{D}^0), \tag{1}$$

where $P(\mathcal{D}|\mathcal{S}, \mathcal{D}^0)$ is the posterior distribution of disparities given the observations $\mathcal{S}$ and $\mathcal{D}^0$. The proposed depth initialization method is described in Sec. IV and the proposed solution to the MAP formulation (1) is described in Sec. V. Note that a reliable estimation of D^0 is quite important, since it guides several components of the fusion methodology.

IV. Depth Initialization

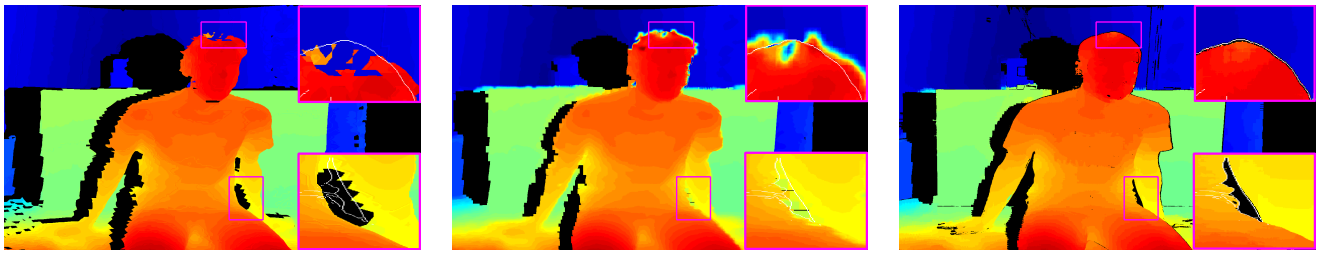
A two-step approach is proposed in order to obtain the initial disparity map D^0. First, we refine $D\downarrow$ to deal with mapping errors. Second, we upsample the refined sparse map in a novel way using color information to obtain the initial dense map D^0. As shown below, this leads to an initialization robust to depth discontinuities, which in turn helps the growing.

A. Sparse-Depth Refinement

We assume a camera setup with a depth camera mounted in between the two color cameras; other depth-stereo setups are

(a) (b) (c)

Fig. 2: (*Best viewed on screen*) (a) The mapping of depth data onto the left image causes artifacts in the presence of depth discontinuities. A cascade of (b) geometry-consistency and (c) color-consistency filters refines the sparse disparity map. Depth values are color-coded from red (close) to blue (far).



(a) (b) (c)

Fig. 3: (*Best viewed on screen*) Depth upsampling results using (a) triangulation-based interpolation [15] after cutting big triangles, (b) joint bilateral filter [34] and (c) our method. The depth values are color-coded from red (close) to blue (far), while black areas correspond to non-available values. The white edges in close-ups show the color edges of the image.

discussed in [9]. Regardless of the depth sensor technology and type, the projection of the depth map onto the left and right images implies a parallax effect, and hence occlusions. Moreover, this causes gaps as well as areas with overlapping depth data close to depth boundaries [15]. In the case of TOF cameras, these areas are further contaminated from *jump-edge errors* [39], or 'flying pixels' [40], [3] while a structured-light camera, e.g., Kinect, leaves more gaps due to the offset between the position of the light projector and the position of the infrared sensor. Fig. 2(a) illustrates the artifacts that we briefly discussed: flying pixels and depth-data overlap in the top and bottom closeups, respectively. In order to eliminate these artifacts we apply a geometry-consistency cascade of two filters: the first one removes isolated pixels (mostly flying pixels) and the second one keeps the foremost pixel inside a window to compensate for the above mentioned overlap. An example of applying this filtering is shown in Fig. 2(b). In practice this does not fully refine the sparse depth map. We still observe mismatches near depth discontinuities, because of depth bias and calibration errors. Therefore, a second filter that imposes color consistency is applied, as described below.

We consider a window centered at $p\downarrow$ and split into four equally sized sub-windows W_i, $i = \{1 \dots 4\}$, such that their intersection is only the pixel $p\downarrow$ (see Fig. 4). The output, $d_{p\downarrow}$, of the filter is:

$$d_{p\downarrow} = \text{med}(D\downarrow_{W_{i^*}}) \tag{2}$$

with

$$i^* = \arg\min_i \big(|I_{p\downarrow} - \text{med}(I_{W_i})|\big). \tag{3}$$

[3]Although not considered here, flying pixels towards the camera can be also observed.

The output $d_{p\downarrow}$ is the median disparity of the adjacent sub-window whose median intensity is closest to that at $p\downarrow$. For color images, the term $|I_{p\downarrow} - \text{med}(I_{W_i})|$ can be replaced by the average deviation from the median, over the color channels. This filter leads to a further refinement near depth discontinuities which are pathological areas for stereo algorithms. The result of this kind of filtering is shown in Fig. 2(c). Note that both refinement filters apply to sparse locations only so that their complexity is negligible.

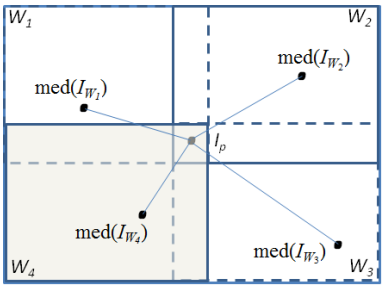


Fig. 4: The window split for the color-consistency filter. The pixel $p\downarrow$ is linked with the closest (shaded) sub-window in terms of the color consistency (links represent color distances from $I_{p\downarrow}$ to med(I_{W_i})).

B. Sparse-Depth Upsampling

While a naive upsampling of the sparse disparity map could be performed, e.g., [15], thereby producing an initial map, strong depth discontinuities are likely to contaminate such an interpolation. Alternatively, cross-bilateral filtering [41] or joint bilateral upsampling [34] may upsample the map using

the HR color image as a guide. The filter support of these methods jointly counts on spatial and range (color) kernels. However, both methods apply a linear smoothing once the filter support per pixel is computed. Instead, we propose a non-linear upsampling strategy that builds on the median filter.

Specifically, the depth (or disparity) at a *dense* pixel location p is initialized by

$$d_p^0 = \text{med}(D{\downarrow}_{\mathcal{N}_p^c}) \tag{4}$$

where $\mathcal{N}_p^c$ is a constrained neighborhood of p, that is $\mathcal{N}_p^c \subset \mathcal{N}_p$, which contains only sparse depth measurements whose color is consistent with I_p:

$$\mathcal{N}_p^c = \left\{ q{\downarrow} \;\middle|\; g(I_p - I_{q{\downarrow}}; \gamma_c) > \epsilon_c \right\} \tag{5}$$

with $g(\xi;\gamma)$ being an exponential mapping (see Table I).

Unlike common bilateral filters, our upsampling process makes a more definitive selection of pixels, thus preserving the depth discontinuities of the scene, while filtering out some of the noise in the depth data. Once the $\mathcal{N}_p^c$ is defined, one can optionally consider a spatial kernel and compute a weighted average instead, in order to better deal with slanted surfaces. Fig. 3 compares our initialization with the upsampling results obtained by [15] and [34]; in the detailed views, the intensity edges are also shown. The proposed method provides more discriminative depth boundaries that coincide with intensity edges. Missing values may be observed in highly textured areas, since $\mathcal{N}_p^c$ may be an empty set. In this example, the radius of $\mathcal{N}_p$ is 20, $\gamma_c = 10$, and $\epsilon_c = 0.2$. The same radius is used for the method of [34] while the scales for spatial and color kernel are 10 and 20 pixels, respectively. The geometry-consistency filter reasonably applies in all cases while our method benefits from our color-consistency refinement as well.

Since the median operator is chosen to account for outliers within a window, the mean operator can be invoked instead when the depth variance almost vanishes (homogeneous areas), thus drastically reducing the computational burden of the upsampling process. Since the vast majority of pixels belong to such areas, the complexity of our filter approaches that of joint upsampling filter [34]. Note that the latter has been extended in [18] by integrating color segmentation results.

V. Depth-Stereo Fusion

Let $d_p \in \mathcal{D}$ have N possible discrete states; the goal is to estimate the disparity (state) of each HR image location through the MAP formulation (1). Once $\mathcal{D}^0$ has been initialized, one can assume that $\mathcal{S}$ and $\mathcal{D}^0$ are conditionally independent, so that the posterior distribution of (1) can be decomposed as

$$P(\mathcal{D}|\mathcal{S},\mathcal{D}^0) \propto P(\mathcal{S}|\mathcal{D})\, P(\mathcal{D}^0|\mathcal{D})\, P(\mathcal{D}). \tag{6}$$

As mentioned, global solutions are prohibitively expensive for high-resolution disparity maps. Therefore, we focus on approximate solutions that allow for the decomposition of the global optimization problem into many local (per-pixel) optimization problems.

The proposed method is based on the seeded region-growing framework [20], [42], where the known message of a location

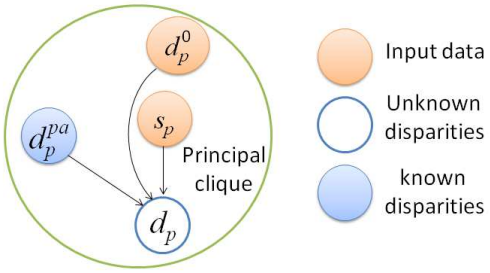


Fig. 5: The principal graph that is iteratively considered in our region-growing fusion method.

(the parent) is propagated to its neighbor (the child). This implies that the estimation of d_p is also conditioned by a parent *known* disparity d_p^{pa}, hence dealing with the principal graph of Fig. 5 for every pixel with unknown disparity (the visiting order of pixels is made clear later). As a result, if $P(d_p)$ is considered uniform, the posterior probability of d_p can be written as[4]

$$\begin{aligned} P(d_p|s_p, d_p^0, d_p^{pa}) &= P(d_p|d_p^{pa})P(d_p|s_p, d_p^0) \\ &\propto P(d_p|d_p^{pa})P(s_p|d_p)P(d_p^0|d_p) \end{aligned} \tag{7}$$

where d_p^{pa} is the parent of d_p and the probability

$$P(d_p|d_p^{pa}) = \begin{cases} \frac{1}{2r+1} & |d_p^{pa} - d_p| \le r \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

has a uniform distribution. In other words, the disparity range for each node is a function of the assigned disparity of his parent. We consider a narrow support area, i.e., a constant low value for r, e.g., 1 or 2; node-dependent parametrization of r is left to future work.

A. MAP as Energy Minimization

As is customary, likelihoods are chosen from the exponential family which leads to an energy minimization framework. Our model assumes an energy-dependent distribution (Boltzmann) for $P(s_p|d_p)$ and a Laplacian one for $P(d_p^0|d_p)$:

$$P(s_p|d_p) \propto \exp\big(-E_S(d_p)/\lambda_S\big) \tag{9}$$

$$P(d_p^0|d_p) \propto \exp\big(-|d_p - d_p^0|/\lambda_D\big). \tag{10}$$

Based on (7-10), Fig. 6 shows an example with the distribution of d_p being constrained by single or joint observations. Because of the exponential terms, the pixel-wise maximization of the posterior distribution reduces to the minimization of the *local* energy

$$E(d_p) = E_S(d_p) + E_D(d_p) \tag{11}$$

where $E_D(d_p) = \lambda|d_p - d_p^0|$ is the regularization term, $\lambda = \lambda_S/\lambda_D$ and $E_S(d_p)$ is the (stereo) data-term which is defined below. The term $E_D(d_p)$ guides the inference in textureless areas, while it penalizes mismatches due to depth discontinuities when the latter are well preserved in D^0. Notice

[4]Strictly speaking, it is an approximation since d_p^{pa} and d_p^0 may not be fully independent.

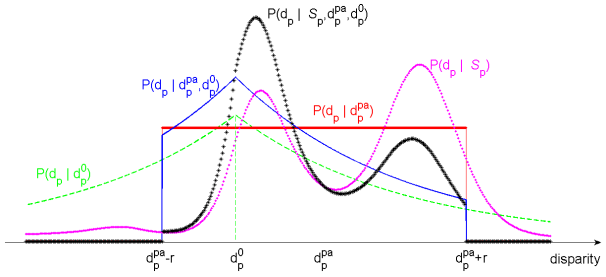


Fig. 6: An example of the probability distribution of d_p constrained by different observations.

that a smoothness constraint is implicitly enforced because of the prior term $P(d_p|d_p^{pa})$ owing to the low value of r. In other words, the support area of $E(d_p)$ is truncated as shown in Fig. 6, and the pixel-wise optimization problem becomes

$$\begin{aligned} &\min_{d_p} && E(d_p) \\ &\text{subject to} && |d_p - d_p^{pa}| \leq r. \end{aligned} \tag{12}$$

The order of visiting pixels and solving (12) obeys a most-confident first-solved rule, as discussed in Sec V-D. Below, once the data-term is defined, we modify (11) to adaptively combine the data-term with the regularizer.

B. The Data Term

The proposed data-term benefits from two properties: its ability to provide energy for subpixel disparities and to adaptively consider pixel-wise similarities within a window. Apart from the advantage of a locally continuous map [31], it is more important to note that these properties lead to a more selective growing in areas with texture (owing to the former) and varying depth (owing to the latter) because of more accurate and reliable local energies.

1) Subpixel disparity: To obtain energy at interpixel locations, we make use of meta-disparities $(d_p; t_p)$ (see Table I). The estimation of t_p builds on [32], which has been shown to be superior to the parabola-fitting around the peak correlation: the pixel $p = (x, y)$ of the left image is matched with the subpixel $p' = (x + d_p + t_p, y)$ of the right image by defining a bounded correlation function $C_{d_p}(t_p)$ ($|C_{d_p}(t_p)| < 1$) which is maximized with respect to t_p for a given d_p. This allows us to define our data-term as

$$E_S(d_p; t_p^*) = 1 - C_{d_p}(t_p^*), \tag{13}$$

where

$$t_p^* = \arg\max_t C_{d_p}(t). \tag{14}$$

Consequently, the total energy (11) is parameterized by t_p^* as well. Note that if t_p^* can be estimated via an analytic solution, the extra computational cost of subpixel disparity estimation is negligible. It becomes clear that, since $C_{d_p}(t_p^*) \geq C_{d_p}(0)$ in (13), more accurate energy values are estimated and more reliable messages are propagated.

Since invariance to photometric distortions is important in stereo matching, we propose to adopt the normalized correlation coefficient, and one of its variants [43], for $C_{d_p}(t_p)$. While the former has been already extended to deal with subpixel disparities [32], the latter has never been extended before. Both use zero-mean vectorized forms of the windows around p and p', let $\mathbf{u}_L(x, y)$ and $\mathbf{u}_R(x + d + t, y)$, with the latter being written via a first-order Taylor approximation as $\mathbf{u}_R(x+d+t, y) \simeq \mathbf{u}_R(x+d, y)+t\Delta\mathbf{u}_R(x+d, y)$ where Δ is a difference operator along the $x-$axis. This is also the case in variational framework [16], [17] where subpixel accuracy is obtained by the early interpolation of the intensity, rather than the late interpolation of correlation around the peak [22].

Enhanced Correlation Coefficient (ECC): The ECC function [32] results from the integration of the above linear interpolation kernel into Pearson coefficient:

$$C_d^P(t) = \frac{\mathbf{u}_L^\top(\mathbf{u}_R + t\Delta\mathbf{u}_R)}{\|\mathbf{u}_L\| \, \|\mathbf{u}_R + t\Delta\mathbf{u}_R\|}. \tag{15}$$

If the denominator of (15) is non-degenerate, then $C_d^P(t)$ is a quasi-concave function of t and its maximization results in a closed-form solution [32].

Enhanced Moravec Correlation Coefficient (EMCC): The Moravec coefficient [43] replaces the denominator of (15) with the mean of the variances. This also allows us to introduce a left-right symmetry, thereby estimating a left disparity $-t/2$ and a right disparity $t/2$ instead of t (such a modification with ECC leads to a complex optimization problem). The *enhanced Moravec correlation coefficient* (EMCC) is defined by

$$C_d^M(t) = \frac{2(\mathbf{i}_L - t/2\Delta\mathbf{i}_L)^\top(\mathbf{i}_R + t/2\Delta\mathbf{i}_R)}{\|\mathbf{i}_L - t/2\Delta\mathbf{i}_L\|^2 + \|\mathbf{i}_R + t/2\Delta\mathbf{i}_R\|^2}. \tag{16}$$

Note that one can easily show that the integration of an interpolation kernel into the cost function of [15] is equivalent with the EMCC scheme. Although (16) is a rational function of t, the next proposition guarantees that the maximizer has an analytic form. We refer the reader to the appendix for the proof and the exact maximizer.

Proposition I: *A rational function of two second-degree polynomials, as in (16), attains at most one global maximum if the denominator is non-degenerate; its maximizer is given by a closed-form solution.*

Note that, the estimation of t could be unreliable for strictly homogeneous areas. The value of the window variance or entropy is a good criterion to assess the reliability of subpixel correction, and to enable it accordingly.

2) Adaptive similarity aggregation: The best-performing local stereo algorithms benefit from an adaptive cost aggregation strategy [24]. This strategy is based on the assumption that depth discontinuities are most likely to coincide with color discontinuities, so that each pixel within a window contributes differently to the (dis)similarity cost based on its spatial and color distance from the central pixel. However, only a few color edges correspond to depth edges and the above assumption should be followed only in the absence of any prior information about the depth. Since in our scenario the prior depth information is available, the spatial and color consistency can be replaced by a depth consistency term.

To be specific, we adopt here the exponential $g(\cdot)$ (see Table I) to compute pixel-wise weights w_q:

$$w_q = g\big(d_p^0 - d_q^0; \gamma_d\big), \tag{17}$$

with $q \in \mathcal{N}_p$. The weights apply element-wise to $\mathbf{u}_L$, $\mathbf{u}_R$, $\Delta\mathbf{u}_R$ and $\Delta\mathbf{u}_L$ in (15) and (16). In other words, we compute the subpixel correction and the optimum local energy after down-weighting pixels in the window that belong to another surface compared to the one of the central pixel. It becomes clear now that not only the term $E_D(d_p)$ but also the stereo term $E_S(d_p)$ in (11) benefit from an upsampling method that is robust to depth discontinuities. Moreover, even if the initial depth map is biased, it is sufficient enough to guide the aggregation step within the window.

C. Adaptive Fusion

While a constant fusion may be reasonable for specific types of scenes (e.g. highly-textured scenes), an adaptive balance of the two terms in (11) is usually preferred, i.e., the less we count on $E_S(d_p)$, the more we should count on $E_D(d_p)$ during the inference.[5] This suggests a convex combination when the scene point of p is viewed by all cameras.

A summary of methods that perform weighted fusion is discussed in [17]. However, most of them consider TOF-based weights for the regularizer (e.g., [13]) which contradicts our goal of a sensor independent fusion. Moreover, directly using the confidence map of a TOF image is not a good strategy [40]. Therefore, we only rely on stereo data to obtain the mixing coefficients. It is well known in stereo or optical flow that the matching of a point is reliable when its associated image patch contains sufficient texture [44], [45]. Since a good indicator for the texture presence is the image entropy [44], an entropy filter provides us with an adequate reliability factor e_p for each window centered at p.

Let us now consider the left image as reference and compute the initial left-to-right disparities. Likewise, we can build a right-to-left disparity map based on the right image, and a cross-checking of these maps can provide an estimation of the major occlusions due to strong depth discontinuities, with respect to the reference image. We refer to these areas as stereo-occlusions and we denote them as Ω_{SO}. Recall now that some points in the left image are not seen by the depth camera, and that this gives rise to gaps in the initial disparity map which can be easily detected. We refer to these areas as depth-occlusions and we denote them as Ω_{DO}. It becomes obvious that the evaluation of $E_S(d_p)$ and $E_D(d_p)$, in Ω_{SO} and Ω_{DO} respectively, should be avoided. Hence, we propose the following adaptive fusion

$$E(d_p) = \eta_p^S E_S(d_p) + \eta_p^D E_D(d_p) \tag{18}$$

with the pair (η_p^S, η_p^D) being defined as

$$(\eta_p^S, \eta_p^D) = \begin{cases} (0,1) & \text{if } p \in \Omega_{SO} \setminus \Omega_{DO} \\ (1,0) & \text{if } p \in \Omega_{DO} \setminus \Omega_{SO} \\ (e_p, 1-e_p) & \text{if } p \in \Omega \setminus \Omega_{SO} \bigcup \Omega_{DO} \\ (inf, inf) & \text{if } p \in \Omega_{SO} \bigcap \Omega_{DO} \end{cases} \tag{19}$$

where Ω defines the whole image area and e_p is the normalized output of the entropy filter. We intentionally add the last case in (19) which shows that the fusion in $\Omega_{SO} \bigcap \Omega_{DO}$ is meaningless and a post-filling method should be followed.

[5]Here we omit the disparity correction t.

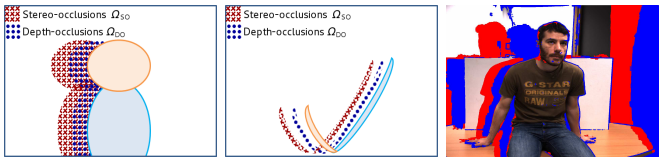
It is important to note that the so-called *ordering constraint* in stereo is valid when large foreground objects appear in the scene, while it is violated when very thin objects are close to the camera (see Fig. 7). The former implies $\Omega_{DO} \bigcap \Omega_{SO} = \Omega_{DO}$ and the latter implies $\Omega_{SO} \bigcap \Omega_{DO} = \emptyset$, provided that the depth sensor is mounted between the color sensors. While the area $\Omega_{SO} \setminus \Omega_{DO}$ can always be predetected, the area $\Omega_{DO} \setminus \Omega_{SO}$ is safely predetected only when the ordering constraint is not valid.[6] This is because Ω_{SO} is detected from the cross-checking of disparity maps that already suffer from depth-occlusions, since they are computed from the depth-to-stereo mapping. Ideally, if the complexity is not an issue, a stereo-occlusion detection scheme (e.g. [46]) based on stereo image pair could be enabled beforehand. Therefore, we prefer to not grow Ω_{DO} that can be optionally filled in a post-processing step. Fig. 7(c) shows the areas for the example of Fig 2.

D. The region-growing algorithm

As has been explained, our method solves pixel-based optimization problems in a region-growing manner, based on the seeds contained in $\mathcal{D}\downarrow$. Since the initial disparities may be noisy and biased, they do not reflect true matches as opposed to [15]. This means that we exploit $d_{p\downarrow}$ to restrict the disparity range of its neighbor, but once a disparity value is assigned to the latter, $p\downarrow$ is reset to a pixel p with unknown disparity. For our convenience, the set $\mathcal{D}\downarrow$ is augmented to an initial set $\mathcal{M}$ of meta-disparities $(d_p; 0)$ with the same cardinality. We also denote with $\mathcal{N}(\mu_p) = \{\mu_p^j\}_{j=1}^4$ the image-based neighbors of μ_p, that is, $\mu_p^j = (d_{\hat{p}}; t_{\hat{p}})$ with $\hat{p}$ being any of the four immediate neighbors of p. Note that μ_p^j does not necessarily belong to $\mathcal{M}$.

Algorithm 1 describes the growing process. Instead of referring to pixel p with disparity d_p and correction t_p, we directly refer to meta-disparities. The algorithm starts by sorting the elements of $\mathcal{M}$ based on their energy value, while it initializes

[6]It would be possible to detect all areas if one knows a priori that the constraint is *everywhere* valid or invalid.



(a) (b) (c)

Fig. 7: (a) The validity of ordering constraint implies $\Omega_{DO} \subset \Omega_{SO}$ when the depth camera is mounted between the color cameras; the shorter the baseline is, the more coincident these regions are. (b) Ω_{DO} and Ω_{SO} do not overlap when the scene contains very thin foreground objects and the ordering constraint is invalid. In (c), Ω_{SO}, Ω_{DO} for the example of Fig 2 are shown. Points that have been removed during the refinement step (Sec. IV-A) are also marked as depth-occlusions, while outliers in the initial maps can produce false positives for Ω_{SO}, e.g. the red points on the right arm of the person.

Algorithm 1 Stereo-Depth Fusion

Require: Image-pair I_L, I_R, set $\mathcal{D}\downarrow$, Threshold T.
Transform $\mathcal{D}\downarrow$ into a set $\mathcal{M}$ of meta-disparities with $t_p{=}0$.
Compute the initial disparity map D^0.
Sort $\mathcal{M}$'s elements based on their energy $E(\mu)$.
Set both the visit and assignment flags $f_v(\mu)$, $f_a(\mu)$ to false, for all candidate meta-disparities, including $\mathcal{M}$.
repeat
  Consider μ_p with minimum energy and false $f_v(\mu_p)$
  Set $f_v(\mu_p) = \text{true}$
  for all $\mu_p^j \in \mathcal{N}(\mu_p)$ with $f_a(\mu_p^j) = \text{false}$ **do**
    $\mu_p^{j*} = \arg\min E(\mu_p^j)$ (Eq. 20)
    if $E(\mu_p^{j*}) < T$ **then**
      Set $f_a(\mu_p^{j*}) = \text{true}$
      Push μ_p^{j*} in $\mathcal{M}$ w.r.t. sorting
    end if
  end for
until card($\mathcal{M}$) is fixed
Compute the dense disparity map D from $\mathcal{M}$
return D.

to false the visit- and assignment-flag of all candidate meta-disparities. Next, it considers the lowest-energy μ_p with false visit-flag and switches this flag to true. Then, it assigns values at *each* $\mu_p^j \in \mathcal{N}(\mu_p)$ *with false assignment-flag* based on the following minimization scheme

$$(d_{\hat{p}}^*; t_{\hat{p}}^*) = \underset{t_{\hat{p}}, |d_{\hat{p}} - d_p^*| \leq r}{\arg\min} E(d_{\hat{p}}; t_{\hat{p}}), \tag{20}$$

where

$$E(d_{\hat{p}}; t_{\hat{p}}) = \eta_{\hat{p}}^S E_S(d_{\hat{p}}; t_{\hat{p}}) + \eta_{\hat{p}}^D E_D(d_{\hat{p}}) \tag{21}$$

and d_p^* is the optimum disparity of the parent node of $\hat{p}$. Note that (21) extends (11) by adding the subpixel disparity parameter into the stereo term, and making the fusion pixel-dependent. In other words, what we do for each neighbor $\hat{p}$ with false assignment-flag is the following. For each candidate integer value $d_{\hat{p}} \in [d_p^* - r, d_p^* + r]$ the optimum $t_{\hat{p}}^*$ that minimizes the term $E_S(d_{\hat{p}}; t_{\hat{p}})$ is obtained based on (16) or (15) and the local energy $E(d_p; t_p^*)$ is computed from (21). Among the $2r+1$ values, the disparity minimizer $d_{\hat{p}}^*$ is finally chosen and the corresponding subpixel correction is assigned, as it is shown in (20). Recall that r has a low value in contrast to conventional stereo algorithms where it equals the whole disparity range. If $E(d_{\hat{p}}^*; t_{\hat{p}}^*) < T$, then the meta-disparity $(d_{\hat{p}}^*; t_{\hat{p}}^*)$ is pushed into $\mathcal{M}$ with respect to the sorting, and its assignment-flag becomes true. The above process is repeated with unvisited meta-disparities until the cardinality of $\mathcal{M}$ remains fixed. The visiting order depends on local energies, since the lowest-energy meta-disparity is always picked from the stack. The final set $\mathcal{M}$ corresponds to the final dense disparity map while a post-filling method can deal with missing disparities; their number depends on threshold T. It is important to note that the algorithm cannot get stuck in a loop, because it propagates disparities in a tree structure, and a true visit-flag can never be reset to false.

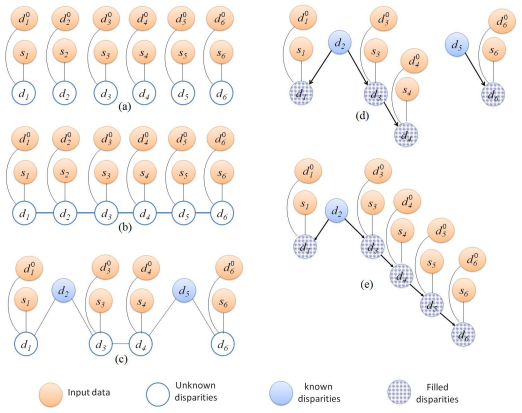


Fig. 8: Dependency network for (a) WTA and (b) MRF models in an 1D example. Initial and two final candidate graphs for the proposed scheme are shown in (c), (d) and (e) respectively.

E. Comparison with Other Inference Formulations

State-of-the-art stereo or fusion methods adopt an MRF model that provides a straightforward way to integrate multiple sensor data. If we recall equation (6), the term $P(\mathcal{D})$ can be written as a Gibbs distribution whose energy is a sum of potential functions over maximal cliques [47], hence a sum of pairwise potentials when a 4-pixel neighborhood is considered. This in turn offers a smoothness term in the global energy equation that leads to piecewise smooth disparity map. There exist exact solutions for such models, under very specific conditions [23]. Their computational complexity, however, becomes more and more severe as the number of states increases, thus becoming prohibitively expensive in the case of HR-stereo. Therefore, one has to focus on feasible approximate inferences that decompose the global optimization problem into a series of local optimization problems.

MRF stereo generalizes the winner-take-all (WTA) approach [22] which can be seen as the simplest inference. Fig. 8(a) and (b) show an 1D-grid example of the MRF and WTA networks. In essence, links between disparity nodes in WTA disappear and each node is connected only to input data (observations). This implies a uniform prior distribution $P(\mathcal{D})$ and the maximization of $P(\mathcal{D}|\mathcal{S}, \mathcal{D}^0)$ reduces to a set of independent pixel-wise maximization problems of type $P(d_p|s_p, d_p^0)$. As with the proposed method, we tacitly assume, that s_p are not intensities but they represent the stereo data of local windows centered at p and its candidate correspondence. This is necessary in WTA since the likelihood cannot count on single pixels only, while it can be optionally used in global MRF solutions as well. It is worth noticing that our algorithm switches to a WTA solution when $r = N$, since the visiting order of nodes and their connectivity becomes meaningless and the uniform distribution $P(d_p^{pa}|d_p)$ is not truncated anymore.

Another MRF approximation that finds pixel-wise optimizers without breaking the connectivity between nodes is the iterated conditional modes (ICM) method [48]. After a proper initialization of all disparity nodes, ICM-like schemes visit one node (pixel) at a time and assign the disparity value that maximally contributes to the global posterior distribution. On

the contrary, our approach requires the initialization of few nodes only, at least one in principle. Moreover, the graph in our case is a set of directed *trees*, i.e., a forest, so that each non-root node has only one parent. It is important to note, however, that it is the output of our algorithm that defines the final network since nodes may be re-ordered, links may be cancelled and arrows may be reversed during the inference.

To be more specific, let us consider the graph of Fig. 8(c) and let us assume the clamping of d_2 and d_5 to the observation d_2^0 and d_5^0. Starting from d_2, we can 'propagate' d_2^0 to its neighbor d_3 (and d_1), i.e., to look for the optimum value of d_3 but being strictly conditioned by the value d_2^0. Note that ICM would look for the best assignment of d_3 value by taking into account both initial values d_2^0 and d_4^0. Moreover, ICM would search, in principle, among all (here N) states for the optimum d_3's assignment, while our scheme looks only around d_2^0, namely d_3 takes values in $\{d_2^0 - r, ..., d_2^0 + r\}$, with r being a small integer. Once d_3's assignment is done, d_2^0 can be passed to d_1 in a similar way. Next, another node is visited, here one of d_3 and d_5, and its disparity is propagated to its neighbors. As a result, the principal graph of Fig. 5 is iteratively considered. Note that ICM assigns a new value to d_3 anyway, while our scheme invokes a criterion that validates the assignment. If d_3's assignment is not valid, our algorithm will possibly assign a new disparity value to d_3 only after d_4's assignment. Fig 8(d) and (e) show possible final graphs obtained by different realizations of our algorithm. In the example of Fig 8(e), the initial disparity of d_5 was cancelled by the validation process, thus all nodes were filled due to d_2^0.

It is now clear that, as opposed to WTA solution, the final inference we obtain depends on the visiting order of nodes. The visiting order of the ICM scheme is either fixed in advance (e.g. raster-scanning), or random. Inspired by [47] and [20], however, we instead adopt a highest confidence first (HCF) scheme that suggests visiting the nodes based on their local evidence (energy). This means that we keep the nodes sorted with respect to their energy, and we visit each time the least-energy node that has not been visited yet. For instance, in Fig. 8(d), starting with d_2 implies that d_2 is more confident than d_5. The above validation process relies on thresholding the local energy as explained in Sec. V-D. Table II summarizes some properties of MRF-based solutions (Graph-cut [49] and ICM [48]), WTA, and our approach.

VI. Experiments

In this section, we evaluate our algorithm, and quantitatively compare it to the state-of-the-art, based on both simulated and real data-sets. We also test our algorithm and provide qualitative comparison on a new and challenging dataset.

A. Simulated Data

We use the Middlebury database, and focus on a challenging data-set which contains 1.5MP images (1300×1100) along with ground-truth maps (GTM) [50]. To simulate an LR disparity map from another viewpoint, we proceed as follows. Given the calibration parameters, we transform the ground-truth disparities into 3D points, as viewed from the midpoint of the baseline. We then apply a 3D rotation to the point-set and we downsample the points by a factor of 10. The rotation is such that the average disparity bias is more than 2 pixels. Finally, we back-translate the points into sparse biased disparities and we add colored noise, that is, a 2D mid-frequency sinusoidal signal with peak-to-peak distance equal to 2σ, where σ denotes the noise deviation. Note that [15], [31], [34] only downsample the GTMs in their experimental setup. Algorithm performance is quantified in terms of the so-called *bad matching pixels* (BMP) percentage in the non-occluded areas [22], i.e., $(1/N_o)\sum_p(|D_p - G_p| > \delta)$ where G is the GTM and N_o is the number of the non-occluded pixels. While the threshold level $\delta = 1$ is mostly used for mid-resolution images, HR stereo justifies the value $\delta = 2$ as well [21]; a value $\delta < 1$ is chosen when subpixel accuracy is to be evaluated.

The same parameter settings are used for our method, in all of the experiments. The radius of the upsampling filter is 20 and the values γ_c and e_c are 10 and 0.2 respectively. Because of the propagation strategy, we choose a relatively small window, i.e., 9×9. The local energy in (11) and the weights in (17) are obtained with $\lambda = 0.01$ and $\gamma_d = 5$, respectively. We enable subpixel correction when the normalized entropy in the (left) window is above 0.4. As mentioned, we use a fixed (and strict) search range around the disparity parent, that is $r = 1$, which leads to the most efficient solution. As for the threshold, we set $T = 0.5$. Note that the energy validation threshold implies a tradeoff between accuracy and density. We recommend setting a middle threshold value and post-filling sparse missing disparities, e.g., with the upsampling filter, rather than using a high threshold that incorporates erroneous disparities in a fully dense map. The density obtained with this strategy is about 90% in HR images. As with all algorithms, large remaining gaps are filled with a streak-based filling [21]. We refer to our methods as Fusion-ECC (F-ECC) and Fusion-EMCC (F-EMCC).

Before comparing with the state-of-the-art, we show the performance gain in terms of the new modules that are integrated compared to EPC method [15], thus quantifying the contribution of the new data-term and the adaptive fusion. Note that [15] (EPC) follows a seed-growing approach by using a quadratic model for both stereo and depth consistency terms respectively. To better evaluate the contribution of the enhanced correlation coefficients presented in the data-terms, we use LR images (450×375 on an average) whose GTM contains subpixel disparities. The noise deviation in the sparse map is $\sigma = 2$. We intentionally do not fill any remaining large gaps, in order to assess the net contribution of each module, and we compute the error for the filled area only (85% density). Table III shows the BMP error averaged over eight images. We also evaluate our approach when none of the modules are enabled, i.e. pure correlation is used along with a fixed fusion of the terms; we refer to this method as *simple fusion*. All of the variants start from the same initial depth map, obtained by our upsampling method. As can be seen, both the data-term and the adaptive fusion process contribute to a better reconstruction, compared to simple fusion. The use of the proposed energy data-term leads to more accurate

TABLE II: Properties of inference algorithms in stereo and/or depth-stereo fusion.

	MRF (graph cuts)	MRF (ICM)	WTA	Proposed
Dependency network	undirected graph (MRF)	undirected graph (MRF)	independent minor graphs	independent directed trees (forest)
Inference	exact*	approximate	approximate	approximate
Prior distribution	Gibbs	Gibbs	uniform	truncated uniform
Invariance to visiting order	yes*	no	yes	no
Disparity search range	full ($r = N$)	full ($r = N$)	full ($r = N$)	narrow ($r \ll N$)
Complexity in HR stereo/fusion	too high	high	high	low

*under specific conditions [23], [49]

results, while the adaptive fusion eliminates large errors (see the error with $\delta = 2$). Even the simple fusion has a lower BMP than EPC, owing to the different stereo- and depth-consistency terms (we use a linear model for the latter). The two proposed criteria behave similarly, with the F-ECC being slightly more accurate, since it achieves higher correlation values (see also Table IV). Note that our results were systematically worse when the initialization of [15] was used.

To compare with the state-of-the-art, we implemented the upsampling methods of [34] and [31] (two-view version), as well as the MRF-based fusion [12] by modifying the MRF-stereo toolbox of [23], referred here to as F-MRF. While [12] uses belief propagation, we experimentally found that Graph-Cuts [49] perform better. Specifically, we tried ten different parameter settings and we found that the best performing algorithm is the expansion mode with Birchfield-Tomasi cost [51] truncated at 7, linear disparity differences truncated at 5 and quadratic cost for the smoothness-energy; the weights for the stereo-, depth- and smoothness terms were 1, 1.2 and 10. We refer to [12] instead of [13] for MRF-based fusion since the latter relies on a TOF-based reliability fusion which cannot be implemented here.

Fig. 9 plots the BMP curves of the upsampling and fusion algorithms as a function of the noise deviation for the challenging low-texture HR image *Lampshade1*. We just add noise here in the down-sampled GTMs. Except for [34], all schemes start from the same HR map, obtained by a *naive interpolation*, while its BMP curve is plotted as well. As can be seen, F-MRF and EPC are more affected by the noisy prior disparity, in contrast to the proposed algorithm, which is less sensitive to initialization. It is clear that the pure up-sampling methods provide acceptable results only when the initial LR disparity map is very accurate.

We now proceed with a detailed comparison including efficient and well-known stereo algorithms as well. Specifically, we include four recently proposed methods, [21] (ELAS), [20] (CGS), [25] (FastAgg), [27] (NonLocalAgg) and two MRF-based stereo algorithms, Graph Cuts (GC) [52] and constant-space belief propagation (CSBP) [28]. The top-performing local algorithms, FastAgg and NonLocalAgg, build cost volumes that depend on both the image size and the disparity range. This leads to a huge memory footprint ($\sim$ 3GB) in the case of HR images, and the authors' implementations could not be run as is. In order to be able to run FastAgg, the cost volume has been split into slices and cached on disk. The NonLocalAgg was run using the maximum allowed resolution while the disparity map produced by the algorithm was finally upsampled. Note that both methods invoke a left-right consistency checking, combine color and gradient information and enable refinement steps. Authors' implementations for ELAS, GCS, GraphCut, CSBP (local-minima version+bilateral post-processing) were used in the comparisons, with the default settings suggested by the authors. We also implemented the ICM algorithm for the MRF-based fusion using the same parameters with GC. For a fair comparison, all fusion schemes merge stereo data with the *same* initial disparity map, which is obtained here by our upsampling process. Due to the simulated experimental setup, however, the error of the initial map obtained from this process is very close to that of Kopf *et al.*'s method [34] (the average difference of their BMP error is below 0.5) and is thus omitted.

Table IV provides the BMP error for eight HR images with error threshold $\delta = 1$, while the corresponding table for $\delta = 2$ is given in the appendix. Bold and underlined numbers mark the lowest and second lowest errors per column. Weakly textured scenes (*Lampshade1*, *Monopoly*) seem to be problematic for conventional stereo algorithms, while cylindrical surfaces (*Bowling2*, *Baby1*) present another challenge. It is not surprising, however, that stereo methods outperform fusion methods in highly textured images (e.g. *Rocks2*), or in images with many thin objects (e.g. *Art*), since the sparse noisy initial map negatively affects the fusion. ELAS and FastAgg behave better than other stereo algorithms. Similar results would be expected from the NonLocalAgg method, if we were able to run it at full resolution.

TABLE III: Contribution of various modules of the proposed algorithm (BMP error averaged over eight LR images).

	BMP (%) for $\delta = 0.5$ / $\delta = 1$ / $\delta = 2$		
	F-ECC	F-EMCC	EPC [15]
Simple Fusion	22.4 / 7.7 / 3.4	23.3 / 7.9 / 3.5	33.0 / 11.6 / 3.7
Data-term	16.8 / 6.4 / 3.2	17.2 / 6.7 / 3.3	
Data-term+ Adap. fusion	14.6 / 5.8 / 2.4	15.0 / 6.2 / 2.7	

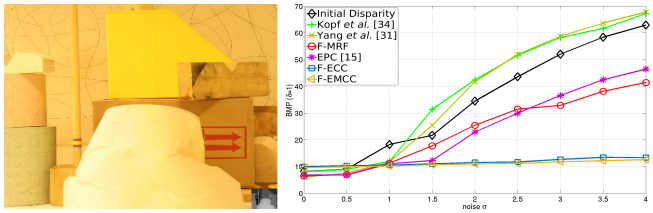


Fig. 9: Left image with superimposed depth map (left) and BMP curves (right) for the HR stereo pair *Lampshade1*. The lower curves show the robustness of the proposed schemes to the noise of the initial depth map.

TABLE IV: BMP for *high-resolution* disparity maps of the Middlebury dataset with $\delta = 1$.

		Lampshade1	Art	Monopoly	Rocks2	Reindeer	Bowling2	Baby1	Moebius	Average
Stereo	GCS [20]	25.1	20.0	52.4	4.5	14.1	25.4	18.1	19.5	22.4
	ELAS [21]	15.2	12.2	38.1	2.6	**5.6**	14.6	10.4	14.0	14.1
	GraphCuts [52]	32.3	26.3	62.1	10.8	28.8	43.1	15.6	19.5	29.8
	CSBP [28]	39.4	26.6	61.8	8.3	20.3	31.9	16.4	20.0	28.1
	FastAgg [25]	26.8	**10.8**	44.6	**2.0**	8.1	15.6	11.5	14.9	16.8
	NonLocalAgg [27]	25.0	23.5	32.6	10.7	24.3	28.8	18.0	17.2	22.5
Upsamping	Yang *et al.* [31]	46.9	47.4	48.5	41.8	42.9	46.3	36.9	41.9	44.1
	Kopf *et al.* [34]	43.7	46.5	45.4	39.4	40.9	42.1	35.1	41.8	41.8
Fusion	EPC [15]	20.7	17.8	20.7	4.2	11.8	20.8	11.9	12.0	14.9
	F-MRF (GC) [12]	16.3	17.1	25.7	4.5	12.1	20.0	10.1	10.5	14.5
	F-MRF (ICM)	23.1	27.9	43.4	14.7	20.1	24.8	17.9	20.1	24.0
	F-ECC	8.4	14.0	7.6	2.9	8.7	**8.5**	4.4	**9.0**	**7.9**
	F-EMCC	**8.2**	14.9	**7.5**	3.1	9.0	**8.5**	**4.2**	10.8	8.2

Stereoscopic and depth data are better fused in general by the proposed criteria than the other fusion methods. EPC and up-sampling methods verify the sensitiveness to their initialization, with the former being more effective due to the fusion process. As far as the MRF solution is concerned, the benefit due to the depth data is verified from the results, i.e. the F-MRF (GC) scheme behaves better than pure-stereo GC. Moreover, GC in fusion provides better results than the simple ICM algorithm. Unlike the proposed criteria, F-MRF deals better with thin objects, since it does not aggregate costs in a window, hence the lower error in *Art* and *Reindeer*. Recall that [12] refines the TOF depth map based on stereo data, without increasing the resolution. By putting aside the high complexity, it seems that MRF-based solutions need to be reformulated for HR stereo-depth fusion, e.g. high-order connectivity might be more helpful, semi-global solutions could be investigated, and conditional random fields [50] might need to be extended to the fusion framework.

Between the two proposed criteria in our fusion scheme, ECC and EMCC, it is the image content that makes one outperform the other. While F-ECC may be slightly better on average, F-EMCC deals better with images of very low texture, e.g. *Monopoly*, which supports Moravec's argument for introducing MCC [43]. Moreover, F-EMCC is more affected by the filling, as it provides less dense maps than F-ECC, provided that the threshold is the same.

Fig. 10 shows the average performance of all fusion competitors as a function of the error threshold δ, while the stereo baseline of FastAgg is added for reference. In essence, this figure reflects the distribution of errors. Evidently, the contribution of the LR depth prior in the fusion schemes is verified, as opposed to the stereo baseline whose performance is bounded. The proposed schemes are more accurate compared to the fusion baselines. However, F-MRF provides lower errors when the tolerance is not that strict ($\delta > 2$) owing to its global smoothness constraint. Note that the performance of [34] approaches the performance of the fusion schemes as δ is increasing. The other upsampling method of [31] seems to produce large errors.

Computational efficiency is an important feature of any depth-stereo fusion method. Table V shows the execution times of the algorithms for the simulated data, as well as some of their implementation details. A combined Matlab-C version of our fusion algorithm requires 2.0s per image triplet while

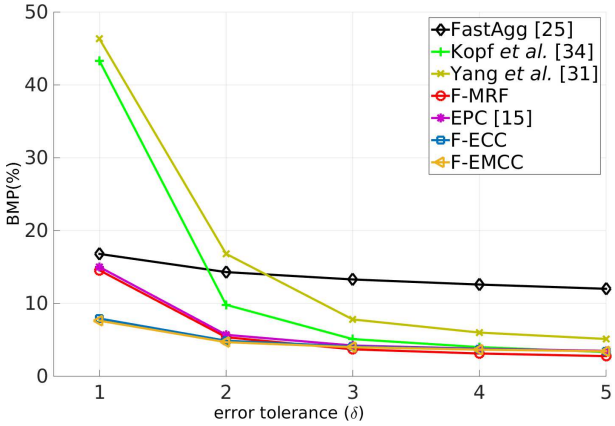


Fig. 10: The BMP error of fusion algorithms averaged over the eight test images as a function of the error tolerance δ.

it takes 0.3s (for 2MP images) when GPU hardware is used for some initializations, and the SSE instruction set accounts for the online computation of correlation values between windows. We also developed a GPU-based implementation of a triangulation-based upsampling (interpolation) that takes less than 50ms. This allows one to envisage real-time execution of the proposed depth-stereo fusion framework, despite the high resolution. Note that it is the Matlab implementation that makes FastAgg and upsampling schemes slow, while the time for the NonLocalAgg method is based on matching at half resolution. We also point out that CSBP attains a solution in

TABLE V: Average execution times of algorithms for the HR Middlebury data-set [50] in a 2.6GHz machine.

	Time(sec)	Matlab	C/C++	GPU	SSE
GCS [20]	1.2	✓	✓		
ELAS [21]	1.0		✓		✓
GC [52]	$> 10^3$		✓		
CSBP [28]	15		✓		
FastAgg [25]	$> 10^3$	✓			
NonLocalAgg [27]	5.0		✓		
Yang et al. [31]	$> 10^2$	✓			
Kopf et al [34]	88	✓			
EPC [15]	1.5	✓	✓		
F-MRF (GC) [12]	$> 10^3$		✓		
F-MRF (ICM) [12]	$> 10^2$		✓		
Our upsampling	95	✓			
F-ECC, F-EMCC	2.0	✓	✓		
Upsampling of [15]	$< 0.1*$		✓	✓	
F-ECC, F-EMCC	$0.3*$		✓	✓	✓

* time needed for the 2MP images of our real data-set (see Fig.13)

TABLE VI: Evaluation on the dataset of Dal Mutto *et al.* [18]

	MSE of disparity estimation			
	Scene A	Scene B	Scene C	Average
Stereo [25]	97.52	5.78	93.94	65.74
Initialization	9.33	6.34	5.62	7.09
Our upsampling	9.96	6.54	5.52	7.34
Kopf *et al.* [34]	10.23	7.45	5.68	7.78
Dal Mutto *et al.* [18]	**3.76**	6.56	8.69	6.34
EPC [15]	8.54	6.61	5.72	6.95
F-MRF [12]	8.96	4.67	6.18	6.67
F-ECC	6.98	**4.19**	**5.39**	**5.52**

a reasonable time, despite its MRF-based formulation.

B. Real Data

Dal Mutto *et al.* [18] provide real TOF-stereo data along with ground truth disparities, shown in Fig. 11. To be consistent with [18], we upsample the depth in a similar way, using a bilateral filter, which benefits from color segmentation. This procedure is used to initialize the depth in *all* fusion schemes. Table VI shows the mean square error (MSE) of the disparity estimation for several algorithms. The contribution of the stereo data in fusion is unquestionable in scenes A and B, where the depth varies locally. All of the fusion methods obtain a more accurate map than the initial one. However, scene C contains only planar objects, and the upsampling methods provide good results. Although the proposed scheme does not perform best in all examples, it always improves the initial estimate, which demonstrates the advantage of adaptive fusion (very similar numbers are obtained with F-EMCC). As expected, the use of stereo data only (e.g., [25]) performs well only with the textured scene B. Our upsamping filter is more accurate than [34] and less accurate than the filter proposed by [18]. Note that our implementation of [34] achieves better results than those reported in [18].

We also assess the performance of the fusion methods on the HCIbox data-set [17]. The scene shows the interior of a box that contains some objects (Fig. 12). Note that there is no texture, apart from some horizontal lines on the stairs and the ramp, hence stereo methods tend to fail. We follow the experimental setup of [17], thus evaluating the depth estimation based on some statistics of the absolute error, after excluding inter-reflection areas (see Fig. 12). We do not include the results of [17], since the authors provided a different inter-reflection mask with larger support area than the one used in [17].[7] Table VII shows the error statistics of the algorithms. All of the fusion methods start from the same initial map, obtained by our upsampling method. The proposed fusion method achieves the lowest mean and median error (similar results are obtained with F-EMCC). The variance of F-MRF is increased (a local bias was observed owing to the global smoothness), while its median remains low. Because of the depth discontinuities, [34] yields a less accurate result compared to our upsampling.

We also captured our own challenging TOF-stereo data-set using a synchronized camera setup, developed in collaboration with 4D View Solutions[8]. Two HR (1624×1224) color cameras and a MESA Imaging SR4000 TOF (range) camera (176×144) are mounted on a rail. The stereo baseline is 50cm approximately while the TOF camera is mounted in the middle. The algorithm of [19] allows us to transform the depth-sensor measurements into a very sparse stereo disparity map (1% sparsity), with an average error of 1.0 pixel. The sparse map is refined and upsampled as discussed in Sec. IV (see also Fig.3). For a fully real-time upsampling, we use our GPU-based implementation of the triangulation-based interpolation [15]. Note that our stereo rectification puts the principal points in the same position, rather than making the optical axes parallel; this maximizes the overlap between the images, given the relatively wide baseline.

[7]Personal communication with R. Nair.

[8]http://www.4dviews.com

(A) (B) (C)

Fig. 11: The three scenes (cropped) of the dataset used in [18].

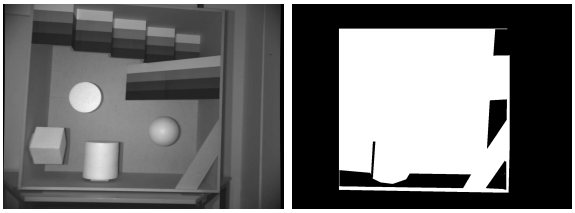


Fig. 12: The left image(left) and the mask that excludes inter-reflections (right) of the HCIbox dataset [17].

Our 'MIXCAM' data-set contains challenging cases, e.g., periodicities, weakly textured areas, thin objects, depth discontinuities, and so on. The fusion algorithm merges the stereo and depth data and the outcome is a dense HR disparity map. We reuse our upsampling filter in a post-processing step to fill missing disparities. A streak-based method fills any remaining gaps in all algorithms. The results are shown in Fig. 13. White areas denote unmatched pixels, while black areas mark the detected TOF-occlusions. The left column shows the left image, with the TOF image shown in the bottom-right corner at the true scale. Next columns show the results of ELAS, FastAgg, F-MRF (GC) and F-ECC algorithms; the last column show the disparity maps of F-ECC after post-processing.

ELAS fills local areas, surrounded by textured points, through an interpolation scheme. We intentionally show the results of ELAS before the streak-based filling; as opposed to FastAgg, where missing disparities after the left-right consistency check are filled. Clearly, a pure stereo algorithm cannot deal with large untextured areas, and the post-filling is unreliable. F-MRF provides fully dense results. Note that

TABLE VII: Evaluation on HCIBox dataset *et al.* [17]

	Mean	St. D.	1^{st} Quart.	Median	3^{rd} Quart.
Kopf *et al.* [34]	3.23	**4.00**	0.93	2.15	3.85
Initialization	3.00	4.21	0.84	1.94	3.61
EPC [15]	3.01	4.32	0.90	1.93	3.17
F-MRF [12]	2.95	4.96	0.83	1.85	3.17
F-ECC	**2.53**	4.25	**0.62**	**1.38**	**2.64**

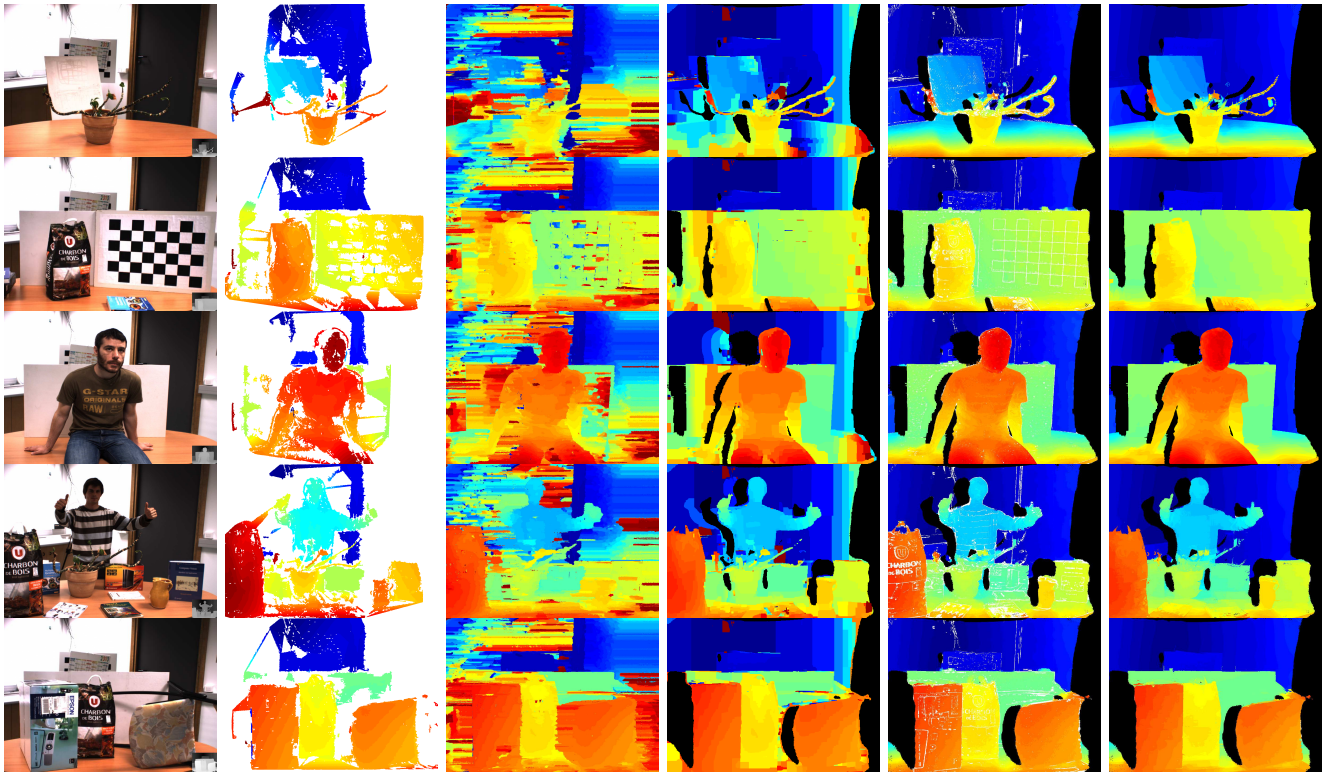


Fig. 13: HR images and disparity maps for MIXCAM dataset obtained by (from left to right) ELAS, FastAgg, F-MRF (GC), F-ECC, and F-ECC after post-processing.

we run F-MRF with half-resolution images (812×612), due to its tremendous memory requirements. Moreover, we set a fixed value along the disparity range for the data-term of all TOF-occlusion points, so that the global inference becomes independent of this area. F-MRF provides artifacts in stereo occlusions, that are next to TOF-occlusions when the scene contains large foreground objects. As with [13], the results of F-MRF scheme verify the lack of an adaptive fusion of the depth- and stereo-consistency data terms, as opposed to our methods. However, F-MRF seems to deal better with very thin objects (e.g. the branch of the plant), as already discussed above. Note that the biased range measurements of very slanted surfaces (e.g. the table-top) negatively affect the fusion schemes, in particular when the table surface lacks texture (e.g. first example). The proposed scheme provides very good results on average, especially after the post-processing step, which fills the gaps and refines the disparities. We obtain very similar results with the F-EMCC method, while EPC provides results *visually* close to ours, but with more gaps. The bilateral upsampling of [34] provides visually good results, but with blurred depth discontinuities (see also Fig. 3).

VII. Conclusions

We have presented a high-resolution stereo matching algorithm that is guided by low-resolution depth data, thus helping the algorithm to compensate for its difficulty in estimating disparities over weakly textured areas. We cast the problem into a MAP formulation whose inference is obtained through a series of local optimization problems, solved hierarchically in a seed-growing manner. The latter characteristic yields an intrinsically efficient solution that allows for near real-time matching of 2.0MP images. The data-term of the energy function benefits from a correlation function that is capable of providing scores at subpixel disparities, from an adaptive cost aggregation step inside the window based on the depth data, and from an adaptive fusion of stereo- and depth-consistency terms based on the scene texture and the camera geometry. These properties lead to a more selective growing process that prevents the algorithm from propagating incorrect disparities. As a result, a low-complexity method builds an accurate high-resolution disparity map. A quantitative comparison against pure stereo and stereo-depth fusion algorithms, as well as a qualitative assessment on real data, has validated the strong performance of the proposed method. Future research will include the optimum visiting order for seeds in the growing framework, as well as an adaptive window size, based on the local surface orientation.

References

[1] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using kinect-style depth cameras for dense 3D modeling of indoor environments," *IJRR*, vol. 31, no. 5, 2012.

[2] Y. M. Kim *et al.*, "Multi-view image and tof sensor fusion for dense 3D reconstruction," in *ICCV Workshops (3DIM)*, 2009.

[3] M. D. Bergh and L. V. Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *WACV*, 2011.

[4] G. D. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *ECCV Workshops*, 2014.

[5] J. Stückler and S. Behnke, "Combining depth and color cues for scale- and viewpoint-invariant object segmentation and recognition using random forests," in *IROS*, 2010.

[6] M. Hansard, S. Lee, O. Choi, and R. Horaud, *Time-of-flight cameras: principles, methods and applications*. Springer, 2012.

[7] F. Remondino and D. Stoppa, Eds., *TOF Range-Imaging Cameras*. Springer, 2013.

[8] J. Geng, "Structured-light 3D surface imaging: a tutorial," *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, 2011.

[9] R. Nair *et al.*, "A survey on time-of-flight stereo fusion," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, M. Grzegorzek *et al.*, Ed. Springer, 2013, vol. 8200, pp. 105–127.

[10] S. A. Gudmundsson, H. Aanaes, and R. Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3D estimation," *Int. J. Intell. Syst. Technol. Appl.*, vol. 5, no. 3/4, 2008.

[11] U. Hahne and M. Alexa, "Combining time-of-flight depth and stereo images without accurate extrinsic calibration." *IJISTA*, vol. 5, no. 3/4, pp. 325–333, 2008.

[12] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *CVPR*, 2008.

[13] J. Zhu, L. Wang, R. Yang, J. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," *IEEE TPAMI*, vol. 33, no. 7, pp. 1400–1414, 2011.

[14] J. Fischer, G. Arbeiter, and A. Verl, "Combination of time-of-flight depth and stereo using semiglobal optimization," in *ICRA*, 2011.

[15] V. Gandhi, J. Cech, and R. Horaud, "High-resolution depth maps based on tof-stereo fusion," in *ICRA*, 2012.

[16] K. Ruhl, F. Klose, C. Lipski, and M. Magnor, "Integrating approximate depth data into dense image correspondence estimation," in *CVMP*, 2012.

[17] R. Nair, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, and D. Kondermann, "High accuracy tof and stereo sensor fusion at interactive rates," in *ECCV Workshops*, 2012.

[18] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. Cortelazzo, "Locally consistent tof and stereo data fusion," in *ECCV Workshop (CDC4CV)*, 2012.

[19] M. Hansard, G. Evangelidis, Q. Pelorson, and R. Horaud, "Cross-calibration of time-of-flight and colour cameras," *CVIU*, 2014, to appear.

[20] J. Čech and R. Šára, "Efficient sampling of disparity space for fast and accurate matching," in *BenCOS Workshop-CVPR*, 2007.

[21] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *ACCV*, 2010.

[22] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, 2002.

[23] R. Szeliski *et al.*, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE TPAMI*, vol. 30, no. 6, 2008.

[24] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE TPAMI*, vol. 28, no. 4, 2006.

[25] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *CVPR*, 2011.

[26] L. De-Maeztu, S. Mattoccia, A. Villanueva, and R. Cabeza, "Linear stereo matching," in *ICCV*, 2011.

[27] Q. Yang, "A non-local cost aggregation method for stereo matching," in *CVPR*, 2012.

[28] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," in *CVPR*, 2010.

[29] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *IJCV*, vol. 33, no. 3, 1999.

[30] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch stereo - stereo matching with slanted support windows," in *BMVC*, 2011.

[31] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *CVPR*, 2007.

[32] E. Z. Psarakis and G. D. Evangelidis, "An enhanced correlation-based method for stereo correspondence with sub-pixel accuracy," in *ICCV*, 2005.

[33] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, 1998.

[34] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *SIGGRAPH*, 2007.

[35] K. Kuhnert and M. Stommel, "Fusion of stereo-camera and pmd-camera data for real-time suited precise 3D environment reconstruction," in *ICIRS*, 2006.

[36] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool, "A hierarchical symmetric stereo algorithm using dynamic programming," *IJCV*, vol. 47, no. 1–3, pp. 275–285, 2002.

[37] L. Wang and R. Yang, "Global stereo matching leveraged by sparse ground control points," in *CVPR*, 2011.

[38] M. Hansard, R. Horaud, M. Amat, and S. Lee, "Projective alignment of range and parallax data," in *CVPR*, 2011.

[39] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (tof) cameras: A survey," *IEEE SENS J*, vol. 11, no. 9, 2011.

[40] M. Reynolds, J. Doboš, L. Peel, T. Weyrich, and G. J. Brostow, "Capturing time-of-flight data with confidence," in *CVPR*, 2011.

[41] G. Petschnigg *et al.*, "Digital photography with flash and no-flash image pairs." *ACM T on Graphics*, vol. 23, no. 3, pp. 664–672, 2004.

[42] R. Adams and L. Bischof, "Seeded region growing," *IEEE TPAMI*, vol. 16, no. 6, pp. 641–647, 1994.

[43] H. P. Moravec, "Toward automatic visual obstacle avoidance," in *IJCAI*, 1977.

[44] G. Egnal, M. Mintz, and R. P. Wildes, "A stereo confidence metric using single view imagery with comparison to five alternative approaches," *IVC*, vol. 22, no. 12, pp. 943 – 957, 2004.

[45] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow, "Learning a confidence measure for optical flow," *IEEE TPAMI.*, vol. 35, no. 5, pp. 1107–1120, 2013.

[46] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum, "Symmetric stereo matching for occlusion handling," in *CVPR*, 2005.

[47] P. Chou and C. Brown, "The theory and practice of bayessian image labeling," *IJCV*, vol. 4, no. 3, pp. 185–210, 1990.

[48] J. Besag, "On the statistical analysis of dirty pictures," *Royal Statistical Soc.*, vol. 48, no. 3, pp. 259–302, 1986.

[49] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," in *IEEE TPAMI*, vol. 23, no. 11, 2001, pp. 1222–1239.

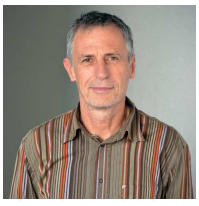
[50] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *CVPR*, 2007.

[51] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE TPAMI*, vol. 20, no. 4, 1998.

[52] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *ICCV*, 2001.

Georgios D. Evangelidis received his BSc, MSc and PhD degree in computer science in 2001, 2003 and 2008 respectively from the University of Patras, Greece. During 2009-2010, he was a post-doctoral researcher of the Fraunhofer IAIS in Sankt Augustin, Germany. Currently, he is a researcher of the Perception Team of INRIA Grenoble, France. His research interests are in the area of computer vision and include 3D reconstruction, gesture recognition and image/video alignment.

Miles Hansard is a lecturer in Computer Science at Queen Mary, University of London. He is a member of the Vision Group, and of the QMUL Centre for Intelligent Sensing. His research interests include 3D scene modelling, depth cameras, and human vision. He has BSc, MRes and PhD degrees from University College London.



Radu Horaud received the B.Sc. degree in electrical engineering, the M.Sc. degree in control engineering, and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble, Grenoble, France. Currently he holds a position of director of research with the Institut National de Recherche en Informatique et Automatique (INRIA), Grenoble Rhône-Alpes, Montbonnot, France, where he is the founder and head of the PERCEPTION team. His research interests include computer vision, machine learning, audio signal processing, audiovisual analysis, and robotics. He is an area editor of the *Elsevier Computer Vision and Image Understanding*, a member of the advisory board of the *Sage International Journal of Robotics Research*, and an associate editor of the *Kluwer International Journal of Computer Vision*. He was Program Cochair of the Eighth IEEE International Conference on Computer Vision (ICCV 2001). In 2013, Radu Horaud was awarded a five year ERC Advanced Grant for his project *Vision and Hearing in Action* (VHIA).