SIAM J. IMAGING SCIENCES
Vol. xx, pp. x

Reliable multi-scale and multi-window stereo matching

Antoni Buades* and Gabriele Facciolo†

Abstract. We consider the two-images stereo disparity problem favoring correctness of matches over density. We will deal with high resolution images which permit an accurate matching in textured zones, but which might present, as any stereo pair, ambiguities and occlusions. Global variational methods can estimate a dense map based on regularity assumptions about the disparity function. However, if these assumptions are incorrect this may lead to erroneous interpretations and mismatches. The availability of tri-stereo or even multi-view stereo imagery permit to combine the disparities from different pairs, allowing for a reliable densification not based on regularity assumptions.

Local methods are suitable for this purpose since they permit to check the validity of each match. The main disadvantages of local methods are the matching ambiguity and the failures of the front-parallel hypothesis (at places like discontinuities and slanted surfaces). We advocate, in this work, for the use of oriented windows in order to deal with slanted surfaces and discontinuities. Unlike adaptive support windows the oriented windows permit to correctly estimate disparities on non front-parallel surfaces.

Several parameter-less techniques for detecting mismatches are presented. The incorporation of these validation techniques in a coarse-to-fine multi-window algorithm, allows to obtain fairly dense results with few mismatches. An extensive comparison, including classical stereo pairs, high resolution satellite images and images from the KITTI benchmark, illustrates the performance of the proposed method.

1. Introduction.

Stereovision consists in estimating the depth of a scene from two or more images taken from slightly different viewpoints. The principle of stereovision is that the apparent motion (or parallax movement) induced by camera movement, is larger for objects closer to the camera and smaller for far away objects. Thus, computing the apparent motion of corresponding points in the two images permits to estimate the depth. For any point $\mathbf{x}$ in one image, it can be shown [24] that the corresponding point in the second image lies on the so-called *epipolar* line of $\mathbf{x}$. A stereo pair (consistent with the pinhole camera model) can be *stereo-rectified* [24] so that the all parallax motion is horizontal, in which case it is referred to as *disparity*.

Stereo matching algorithms aim at estimating the disparity of as many image points as possible. The stereo matching methods are classified [44] in *local* and *global methods*. The local methods compute the disparity by correlating a small window (or patch) along the epipolar lines, being the *Sum of Squared Differences* (SSD) the most common matching cost used for this purpose. The lack of texture or information in the image is the main challenge of local methods. That is, if an image patch lacks texture or its signal-to-noise ratio is too small, then the computed disparity will likely be incorrect. *Global methods* cope with these ambiguities by restricting the disparities to some class of smooth functions, which usually permit to derive reasonable estimations even on textureless areas. Most global methods could even extend these estimations to half-occluded regions (regions which are visible only in one of the images). However, the result of a global method makes no explicit distinction between bad

*Universitat de les Illes Balears, Spain (toni.buades@uib.es)
†CMLA, ENS Cachan, France (facciolo@cmla.ens-cachan.fr)

and good matches, confusing where the disparity is factual or not. Moreover, the smoothness of the solution is always controlled by some parameter, whose optimal value usually depends on the scene being processed.

The choice between local and global methods is certainly application dependent. In some cases dense disparity maps are needed, regardless of the possible extrapolation errors that a global method may introduce. However, here we will focus on applications that favor reliability over completeness thus allowing for non matched regions. Although a dense map is desirable, in some cases it is important to distinguish reliable, and thus validated regions, from interpolated ones. This is the case, for instance, of cartographic applications from high resolution satellite stereo images [12] and urban mapping from a car-mounted stereo rig such as the ones from recent benchmark [21], for which our algorithm yields the lowest error for the estimated pixels. Besides these applications we will also consider classical stereo databases.

In this work we propose a principled block-matching algorithm for stereo, using a coarse-to-fine strategy. A multi-window algorithm with oriented windows is used to compute the disparities. This new approach permits to deal with non fronto-parallel surfaces and to match closer to the edges of the object, leading to higher matching densities. On a slanted surface the chosen window will adapt to the direction of least variation of the disparity. Whereas near a disparity discontinuity the window oriented as the edge will match with lower cost, thus correctly matching closer to the discontinuity.

To limit the mismatches we propose a set of criteria to detect and reject the incorrect matches. Each criterion is designed to cope with one of the limitations of block-maching. Match ambiguity is solved by adapting previous distinctiveness approaches [35, 41] while a new criterion is designed to cope with fattening effects due to occlusions. Unlike most match validation methods the proposed ones have no tradeoffs and are virtually parameter-less besides the size of the patch. The criteria are applied for each oriented window and scale.

We summarize the plan of the paper as follows. Section 2 introduces local methods and their main limitations. The literature including state of the art methods and how they deal with local methods weakness are also reviewed. In Section 3 we introduce the multi-scale and multi-window block-matching. In the same section we describe the criteria used to reject incorrect matches. Lastly, Section 4 compares the proposed algorithm with state-of-the-art local methods for non-dense block-matching. It is shown that the proposed method yields the lowest mismatch rates among all the considered algorithms while producing fairly dense disparity maps. This validation is mainly done using the high resolution Middlebury images [27] for which a ground truth is available. We also compare the results obtained on satellite images[1] and street-level urban images from the recent KITTI benchmark [21]. An on-line demonstration of our algorithm is also made available[2].

2. State of the art in local methods and limitations.

We denote by $\mathbf{x} = (x_1, x_2)$ the position of a point in a continuous image domain $\Omega \subset \mathbb{R}^2$, and by $u(\mathbf{x}) = u(x_1, x_2)$ and $v(\mathbf{x})$ the images of a rectified stereo pair. For non-integer positions $\mathbf{x} \in \Omega$, $u(\mathbf{x})$ is computed interpolating the discrete samples of u. Local methods estimates the disparity $\mathbf{d}$ of the point

[1] Pléiades satellite images kindly provided by the Centre National d'Études Spatiales (CNES).

[2] On-line demo of the proposed method is available at: http://dev.ipol.im/~facciolo/msmw/

$\mathbf{x}$ by the winner-take-all strategy

$$\mathbf{d}(\mathbf{x}) := \arg\min_{\mathbf{d}' \in p\mathbb{Z}} c(\mathbf{x}, \mathbf{x} + \mathbf{d}'),$$

where $c(\cdot, \cdot)$ is a cost function computed by comparing patches of u and v centered respectively at $\mathbf{x}$ and $\mathbf{x} + \mathbf{d}'$. The factor p is the subpixel disparity step, usually $p = \frac{1}{n}$ with $n \in \mathbb{N}^+$.

Many matching costs have been proposed for local block-matching methods among them: sum of squares distances (SSD), normalized cross correlation, robust costs, mutual information, and census. For a comprehensive review on matching costs we refer to [27] and references therein. Regardless of the cost used for comparing windows, there are some limitations intrinsic of local methods, namely: match ambiguity, inadequacy of geometric model, and disparity sampling. In the following paragraphs we review the stereovision literature on local methods highlighting how these limitations are dealt with.

Match Ambiguity. The lack of texture in the image is one of the main sources of matching errors for the local methods. Similar issues are observed in presence of the aperture problem, that is when the direction of the main geometrical structure in the patch coincides with the epipolar line, or if the patch belongs to an object with a repetitive pattern. The latter case is quite common for man made objects, as for example building facades. The common denominator for all these cases is the ambiguity of the matches.

A straightforward way to reduce ambiguity is to use larger comparison windows. However, this increases the risk of incurring in other problems. For instance, larger windows are more likely to contain objects with different depth or half occluded regions. For this reason algorithms using larger window sizes usually adapt the shape of the windows by taking into account image edges or objects [29, 32, 53, 56, 38, 6, 22, 54, 39]. These methods identify depths discontinuities with image discontinuities, which is not always the case.

Even for large window sizes the extracted patches can be non-informative as for instance in presence of repeated patterns. A common criterion used to detect these cases compares the costs of the best and second best match candidates [43, 26, 10, 28]. This criterion is also used in the SIFT keypoint matching algorithm [33] in order to reject mismatches. For that a threshold on the ratio (or difference) of the first and second best costs has to be set. Different detections might be obtained depending on the value of this critical parameter. Manduchi and Tomasi [35] observed that the content of the reference image could be used to predict how ambiguous the match would be. For that they define the *distinctiveness* of an image position as the perceptual distance to the most similar other position in the same image within a search window. Distinctive points are not necessarily rich in texture, but their features are unique as they look like nothing else in the image (at least along the epipolar line). Particularly, in [35] the authors used the auto-SSD function (Sum of Squared Differences computed in the same image), rejecting pixels with auto-SSD above the average cost value for the entire image. However, this test also depends on the critical choice of the threshold parameter.

A parameter-less approach was proposed by Sabater et al. [41]. The authors proposed a new measure based on the number of false alarms (NFA) that a patch similarity takes by chance, inspired by the general framework proposed in [15]. The algorithm uses a principal component analysis (PCA) to represent patches and the histogram of PCA coefficients of

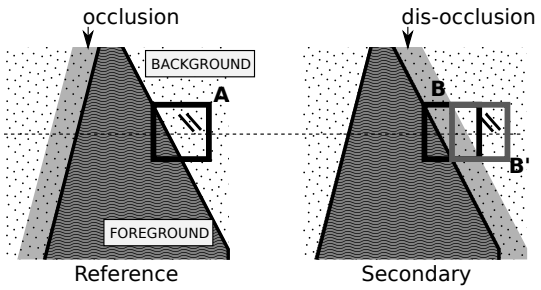


Figure 1. *Depth discontinuities and foreground fattening. The center of the window (A) is on the background but the window contains the depth boundary, thus some points in the window match at the foreground disparity (B), while others match at the background disparity (B'). The best match will be for the region containing the stronger horizontal texture, which is usually the object boundary itself. Since the object boundary has the foreground disparity, a strong preference for the foreground disparity is created. Leading to the well-known "foreground fattening" effect.*

the whole image as an a priori model. The proposed model matches only correct pairs but discards matches whenever the patches are slightly different, for example in slanted zones. In the same work, the authors presented a distinctiveness based criterion, comparing for each pixel the SSD cost of the best match in the second image and the auto-SSD proposed in [35]. We will modify this approach adapting it into our multi-scale and multi-window framework for detecting ambiguous matches.

Inadequacy of geometric model. Three-dimensional geometric discontinuities of the scene produce occlusions and dis-occlusions, which cause spurious matches or foreground fattened results. This occurs because patches overlapping discontinuities are only partially visible in one image [46, 31], thus the right match may not be found. Figure 1 illustrates the *foreground fattening* phenomenon. A similar phenomenon is observed on slanted surfaces, there the block-matching is usually imprecise and may even produce incorrect matches. This is because the patches on slanted surfaces appear distorted (by an affinity in the case of rectified images) in the second image [16]. The overall reason for these failures is the inadequacy of the fronto-parallel hypothesis for block-matching methods. That is, the assumption that any image patch can be found undistorted in the second image.

The usual way to cope with depth discontinuities is to use adaptive windows (also known as adaptive support weights) that avoid image discontinuities as was first proposed by Kanade and Okutomi [29]. Similar works by Lotti and Giraudon [32] and more recently by Wang et al. [53] pre-compute the image edges and iteratively grow a comparison window that avoids discontinuities. Patricio et al. [38] and Yoon and Kweon [56] select an adaptive window containing only pixels with a grey level similar to the reference one, like in neighborhood and bilateral filters [50, 55]. Recently, the Cost Volume Filter [39] has allowed to compute such adaptive windows very efficiently by using a local linear model to filter the volume of matching costs using the reference image as guide. Other methods limit the size and shape of the window by imposing an image segmentation [6, 22, 54]. All these methods identify depth discontinuities with image discontinuities, which is not always the case.

Other approaches do not try to explicitly avoid the discontinuities of the reference image.

Instead, they choose an adaptive window using a minimum matching cost criterion. The underlying idea is that windows that do not contain depth discontinuities will match with a smaller cost. Fusiello et al. [20] choose among all the windows containing the reference pixel the one that yield the minimal matching cost. This well known technique can be computed efficiently by the so-called *min-filtering*, which is a post-process of the costs computed with centered windows [44]. This technique significantly reduces the *foreground fattening* near the occlusion boundaries of the scene. However the *min-filter* post-process creates shocks on regions with smooth disparities and sometimes it may also propagate incorrect disparities near textureless areas. Veksler [51] applied the same strategy but considering square windows of different sizes. A more elaborated version by Hirschmüller et al. [26] adapts the shape of the window by dividing its support into small sub-windows and by selecting those for which the minimum matching cost is attained.

The methods based on adaptive windows deal effectively with discontinuities, however they are bound to the fronto-parallel hypothesis as they don't contemplate the case of slanted surfaces. To cope with the slanted surfaces some methods propose [23] to explore all possible surface orientations, at each image point. For rectified images the transformations between planar patches has three degrees of freedom: translation, horizontal tilt, and shear, which may be unaffordable in many situations. Some recently proposed methods [7, 34] efficiently explore the space of affine transformations by using the PatchMatch [3] strategy. In this work we improve the matching on non-frontal surfaces by matching using elongated windows with different orientations. This family of windows has a single degree of freedom: the orientation.

Recently, the mathematical analysis of local methods presented in [14, 42, 5] has clarified the presence of subpixel fattening errors in regular parts of the disparity function. These works were motivated by the study of subpixel precision in low baseline stereo for satellite imaging. This micro-fattening introduces subpixel errors in the computed disparity whenever the image gradient is not well spread over the comparison window. In [5] the authors proposed a solution for this kind of fattening effect. However this solution is only valid for regular parts of the disparity function and therefore cannot be applied on occlusion and dis-occlusion parts of the stereo pair.

In this work we propose a new technique, related to the *min-filter*, for detecting occluded and dis-occluded pixels.

Sampling issues. Inadequate sampling of the disparity can also result in mismatches (see Figure 2). That is, because of the image sampling we can't expect to find in the second image an exact copy of reference patch (unless the true disparity is integer). For this reason many block matching methods estimate subpixel disparities [48].

Birchfield and Tomasi [4] proposed a matching cost (based on the SSD) that is insensitive to the sampling of the patches. This cost, which is estimated at integer positions, is proven to be stable with respect to subpixel translations of the patches. This is particularly useful when used within global methods. However, as seen in Figure 2, for a winner-take-all method this cost [4] yields a limited performance gain with respect to the subpixel SSD cost.

Comparing in Figure 2 the disparity estimates at 1-, $\frac{1}{2}$-, and $\frac{1}{4}$-pixel steps we confirm that integer disparity steps yield incorrect results on textured areas. Failures are less frequent when sampling at $\frac{1}{2}$-pixel intervals, but can still occur. While sampling at $\frac{1}{4}$-pixel intervals

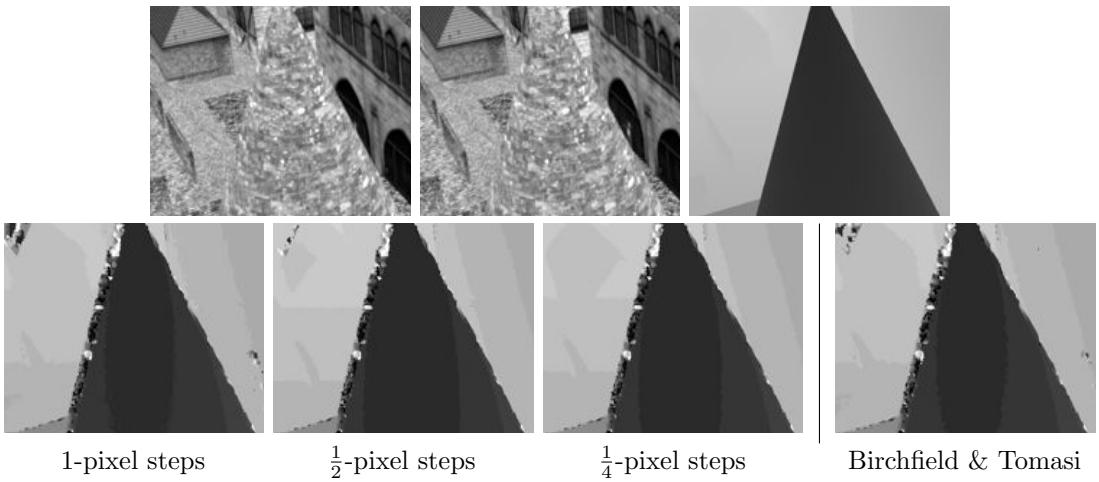


Figure 2. *Top: synthetic stereo pair and ground truth disparity. Bottom, from left to right: disparity computed using SSD by exhaustive search with* 1*-,* $\frac{1}{2}$*- and* $\frac{1}{4}$*-pixel steps. Note that increasing the precision eliminates false matches from the upper-left corner and right portion of the image. The last image is computed using the sampling insensitive cost proposed by Birchfield and Tomasi [4] with integer disparity steps, note that the result is comparable to SSD at* 1 *pixel. All disparity images are quantified at 1 pixel for visual comparison.*

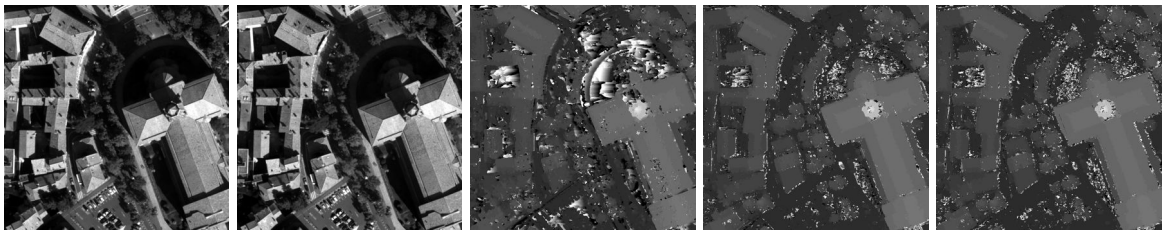


Figure 3. *Comparison of the SSD, ZSSD, and NCC matching costs using square patches of* 5×5 *pixels. From left to right: the stereo pair, the disparity estimated obtained using SSD, ZSSD, and NCC. The three methods give similar results when there is no contrast changes from one image to the other. Small intensity differences present in the dark areas of the images above yield to SSD mismatches while are correctly matched by both ZSSD and NCC. We observed that for aerial images ZSSD and NCC perform similarly [27] thus we use the former.*

these mismatches are seldom. Although more expensive, the subpixel sampling permits to identify the true minima of the matching cost. In addition, to further refine the disparity, the over-sampled cost function could be interpolated using compact interpolation filters [48]. In this work we will estimate the disparities with quarter-pixel precision.

3. Multi-scale multi-window block matching algorithm. As cost function we use the zero-mean SSD (ZSSD) [27] which removes the average intensity of each patch rendering the comparison independent of the mean intensity. This cost is defined as

$$\mathrm{ZSSD}(\mathbf{x},\mathbf{y}) := \frac{1}{|B_r|}\sum_{\mathbf{t}\in B_r}\left|u(\mathbf{x}+\mathbf{t}) - v(\mathbf{y}+\mathbf{t}) - \overline{u|_{\mathbf{x}+B_r}} + \overline{v|_{\mathbf{y}+B_r}}\right|^2,$$

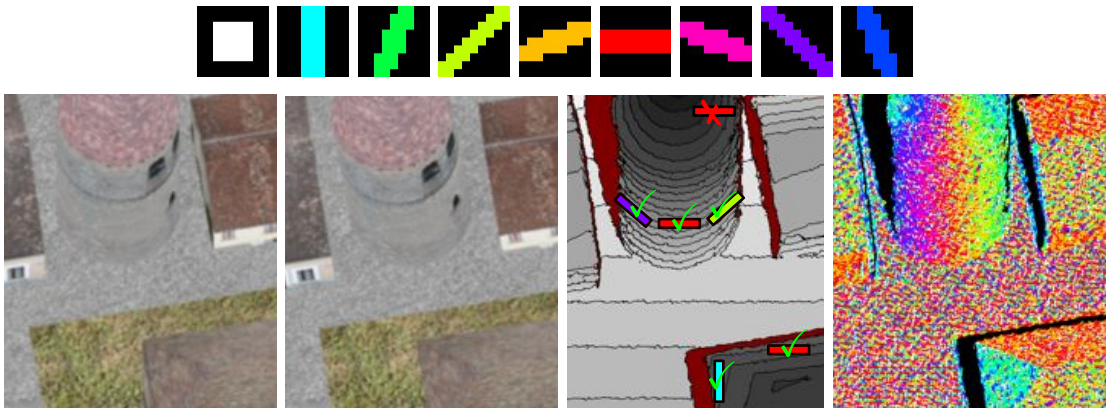


Figure 4. *Correlation windows with different orientations better adapt to different slanted planes, because with the correct window orientation the depth of the underlying surface is almost constant. Top row: oriented windows used in the experiments (with color code), corresponding to a* 5×5 *window. Bottom row from left to right. The stereo pair. The computed disparity map with some isolines highlighted, overimposed are shown some windows whose orientation best fit the surfaces. The rightmost image illustrates the minimum-cost orientation determined by the algorithm.*

where $\overline{v|_{\mathbf{y}+B_r}}$ denotes the average of image v over the window centered at $\mathbf{y}$. In Figure 3 we show that the ZSSD is able to deal with changes of intensity between the images.

3.1. Matching with multiple window orientations. Most block-matching methods implicitly assume the fronto-parallelism of the scene by testing only for translations of the matching windows. However, this hypothesis is seldom correct. If the disparity function is not constant within a window, then the corresponding region appears distorted in the second image. Thus two image patches should not be compared directly. The SSD minimization is robust against slight deformations, but fails in presence of stronger ones. An extreme case of this failure occurs when the window contains a depth discontinuity (two different depths). In this case it is impossible to find a unique correspondence leading to the occlusions and dis-occlusions artifacts.

The consequences of this failure of the fronto-parallel assumption are two fold. First, mismatches appear on highly slanted surfaces because of to the strong distortions suffered by the image within the matching window. Second, many points are mismatched in the proximity of depth discontinuities.

To cope with both problems we propose to match each pixel using elongated windows with different orientations and then to choose the window yielding the minimum matching cost. This approach differs from the shiftable windows proposed in [2, 8, 44]. We propose to use only windows centered at the reference pixel but with different shapes [26]. The shape of the window giving the minimum matching cost intuitively will be the one for which the disparity function varies the least. That is, the window being as fronto-parallel as possible. This is confirmed by the experiment shown in Figure 4.

Using elongated and oriented windows improves the matching on slanted surfaces and close to the depth discontinuities. Near a depth discontinuity the chosen window will be aligned with the discontinuity, thus producing valid matches closer to the discontinuity. Whereas

on a slanted surface the chosen window will adapt to the direction of least variation of the disparity, thus matching with a lower cost.

3.2. Multi-scale algorithm. The disparity range considered by the block-matching directly affects the computation time, but it also affects the ambiguity of the matching. That is, as the search range grows so does the probability of mismatch due to repetitive patterns or noise. In general, if a wide search range is considered, then larger patches are needed in order to keep the ambiguities from growing. Thus, the patch must be large enough to contain the information to match meaningfully, but it should also be small to avoid fattening artifacts.

The proposed multi-scale approach (see Algorithm 1) uses the disparity computed at a coarser scale in order to restrict the search range at the current one. This permits to keep a small window size while retaining meaningfulness of the matches. To prevent the propagation of mismatches, rejection criteria (presented in Section 3.3) are applied at every scale. The points rejected at the previous (coarser) scale are searched over the full disparity range. While the valid points are assigned a narrow disparity range that depends on the minimum and maximum disparity of valid points in the window.

Each multi scale level is obtained by convolution with a Gaussian kernel of standard deviation $\sigma = 1.2$ and subsampling of factor two starting from the initial pair. This means that the noise is reduced at each subsampling step making the matching more robust at coarser scales. The disparity computed at coarser scales is up-sampled by bicubic spline interpolation and used to constrain the search range at finer scales.

A classical problem with multi-scale procedures is related to small scale structures with large disparities. This is a well-known problem in optical flow computation where small objects moving fast disappear at lower scales of the pyramid making impossible to estimate the displacement for these objects. Our algorithm is able to correctly deal with this problem due to its locality and the validation criteria introduced in next section. If an incorrect search range is predicted by a coarser scale, the match will be invalidated at the first scale the object appears. Therefore, a full search range will be used for these pixels at the next fine scale.

3.3. Match validation criteria. Most stereo matching algorithms use the classic Left-Right consistency check [11, 19] to reject incorrect matches. This test rejects a match whenever the left-based disparity is not the inverse mapping of the right-based disparity. That is, a match is rejected if $|\mathbf{d}_R(\mathbf{x} + \mathbf{d}_L(\mathbf{x})) + \mathbf{d}_L(\mathbf{x})| > \tau$, where the threshold τ is usually set to 1 and $\mathbf{x} + \mathbf{d}_L(\mathbf{x})$ denotes the position in the right image of the homologous point of $\mathbf{x}$ (in the left image).

The Left-Right consistency check permits to detect many occluded pixels, but not all of them (see Figure 7). Some errors due repetitive patterns and textureless areas might be detected as well, but the foreground fattened pixels associated to dis-occlusions are usually not detected.

In the following paragraphs we present two tests specifically designed to detect ambiguous matches and dis-occlusion fattening artifacts.

Match ambiguity. Sabater et al. [41] proposed a criterion for detecting ambiguous matches that compares the costs of the best candidate match in the secondary image c_1 and the cost of the best match in the reference image itself c_{auto} (see Figure 5). The idea is that when the auto-similarity cost of a patch (c_{auto}) is smaller than the cost of its best

Algorithm 1: Recursive multi-scale block matching algorithm. Initially $dMin$ and $dMax$ are constant images set to the global minimum and maximum ranges and $s = 0$.

$MultiscaleChain(u,\ v,\ dMin,\ dMax,\ nScales,\ s)$:

Input: image pair u and v.
Input: min & max disp. range images $dMin$ and $dMax$.
Input: number of scales $nScales$.
Input: current scale s.

Output: Fixed precision disparity $disp$.
Output: Updated $dMin$ and $dMax$.

if $s < nScales$ **then**
  $G_\sigma \leftarrow$ Gaussian kernel of std $\sigma = 1.2$
  $Gu = G_\sigma * u$; $su = Gu \downarrow 2$;
  $Gv = G_\sigma * v$; $sv = Gv \downarrow 2$;
  $GdMin = G_\sigma * dMin$; $sdMin = GdMin \downarrow 2$;
  $GdMax = G_\sigma * dMax$; $sdMax = GdMax \downarrow 2$;
  $sdMin = 0.5 * sdMin$; $sdMin = 0.5 * sdMin$;
  disp $= MultiscaleChain(su,\ sv,\ sdMin,\ sdMax,\ nScales,\ s+1)$;
  $dMin = sdMin \uparrow 2$; $dMax = sdMax \uparrow 2$;
  $dMin = 2 * dMin$; $dMin = 2 * dMin$;
end if

$disp = BlockMatching(u,\ v, dMin, dMax)$

for each pixel p in $disp$ **do**
  if p is rejected **then**
    Set $dMin(p)$ and $dMax(p)$ to the global minimum and maximum range.
  else
    Set $dMin(p)$ and $dMax(p)$ to the minimum and maximum of validated disparities inside the correlation window.
  end if
end for

match in the second image (c_1), then patch is likely to be part of a repetitive pattern and thus non-distinctive. We adapt this criterion, introducing a sampling correction term $c_{sampling}$, rejecting a match when

$$c_1 > c_{auto} - c_{sampling}, \tag{3.1}$$

where the window shape and disparity step p (in our case the disparity step is $p = 1/4$ of pixel) are the same as in the matching step. The computation of the cost c_{auto} only takes into account windows in the reference image at a distance larger than one pixel of the reference window. The correction term $c_{sampling}$ is computed as the maximum of the costs of matching the reference window with itself but shifted by $p/2$ and $-p/2$.

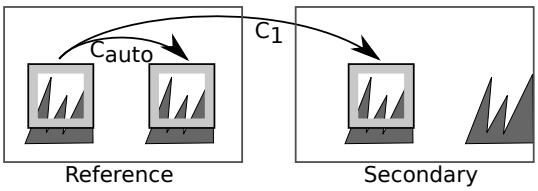


Figure 5. *Match ambiguity criterion. Compares the cost c_1 of the best match candidate in the secondary image and the cost c_{auto} of the best match in the reference image itself minus a small sampling compensation term. If $c_{auto} - c_{sampling} < c_1$ then the match is considered ambiguous.*

The sampling term is needed to compensate for the differences due to the samplings in both images. That is, c_1 and c_{auto} are directly comparable only if the patches are sampled at the same sub-pixel position, which is usually not the case.

Note that c_{auto} acts as an adaptive threshold for the match. If the patch belongs to a repetitive structure, then the test is more restrictive. Textureless regions are also detected as ambiguous by this criterion. Moreover, unlike other criteria based on the relation of the first and second best match, this test needs no parameter tuning.

The distinctiveness measure proposed by Manduchi and Tomasi [35], and later refined by Yoon and Kweon [57], rejects pixels based on a global threshold on the value of c_{auto}. Instead the proposed test uses c_1 as an adaptive threshold. Egnal [17] proposes to take two images in quick succession from each viewpoint, then use the second image to estimate an adaptive confidence for the matching. However, this procedure is only useful for static camera systems where it is possible to take two images, which is not the case for instance in satellite imaging. Instead, the used match ambiguity criterion does not require any supplementary image.

Fattening detection. We design a test to detect fattening artifacts due to occlusions, dis-occlusions, and errors on slanted surfaces. The proposed method shares with the shiftable window method by Fusiello et al. [20] that it gives more confidence to the pixel in the neighborhood having the lowest matching cost $\mathbf{x}_{MC}$.

The proposed method computes a planar approximation of the disparity around the current patch center $\mathbf{x}$. This approximation is computed by a modified RANSAC [18] strategy. We fix the point $\mathbf{x}_{MC}$ and randomly select two different pixels in the neighborhood of $\mathbf{x}$ and compute the plane passing though the coordinates and associated disparities of the fixed triplet. This process is iterated and among all the tested planes we keep the one correctly approximating the larger quantity of disparities in the neighborhood.

The reference pixel x is not necessarily one of the correctly represented disparities by the selected plane. The pixel $\mathbf{x}$ is rejected if its disparity differs too much from the estimated planar approximation Π

$$|\mathbf{d}(\mathbf{x}_{MC}) - \mathbf{\Pi}(\mathbf{x})| > s, \tag{3.2}$$

where s a precision threshold set in practice to 1. Figure 6 illustrates the behavior of this criterion in the 1D case.

Finishing touches. Isolated matches are also rejected. These are valid matches surrounded

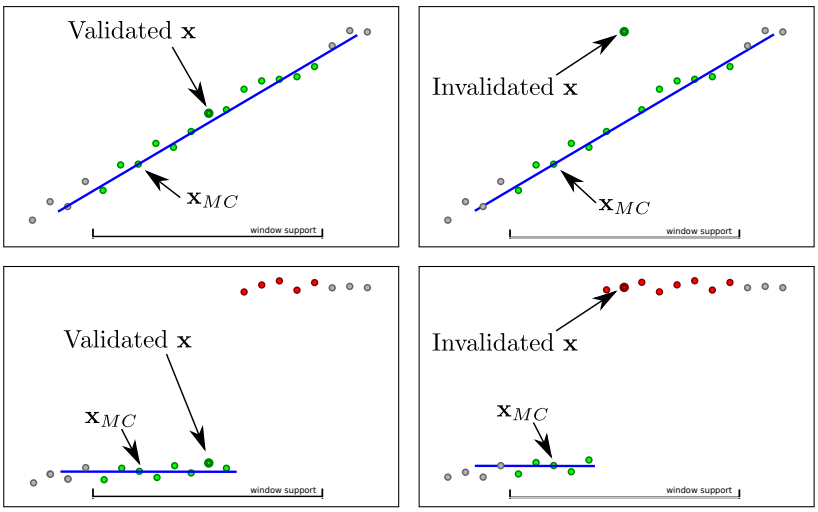


Figure 6. *Illustration of the fattening detection behavior in 1D. First row: typical validation/invalidation of x in a slanted surface. Bottom row: typical validation/invalidation of x in a depth discontinuity.*

by many rejected points (rejected by the previous criteria). The rationale of this test is that a valid but isolated point is more likely to be a mismatch than a proper match. In the eventuality of being a proper match, it would still be a too small feature for the current window size. We apply a grain filter [52, 36, 9] to remove validated regions with area smaller than the area of the correlation window (5×5 pixels in all the experiments).

Figure 7 shows the disparity computed using ZSSD and the pixels rejected by each one of the described criteria. As expected the Left-Right consistency test rejects points on the occlusion and dis-occlusion parts of the image because in those regions the same patch doesn't exist in both images (extract (a) in Figure 7). However, some of the occlusion and dis-occlusion points may produce consistent matches because of the fattening effect. In those cases the correspondence match is dominated by the occluding edge, which is visible in both images. The fattening detection rejects points mainly in dis-occlusions parts of the scene as seen in the extract (b) in Figure 7. The self-similarity criterion correctly rejects points in ambiguous areas (see the rooftop in the top left corner of the image in Figure 7 and in the extract (c)). It might also reject pixels at occluded parts, since it might happen that the auto-similarity cost is smaller than the matching cost just because the patch does not exist in the secondary image. Moreover, as mentioned before, the self-similarity test also rejects textureless regions, leaving only some isolated points, which are then removed by the isolated pixel criterion.

3.4. The final MSMW algorithm. The final algorithm uses the multi-window strategy (Algorithm 2) at each scale of the multi-scale method described in Algorithm 1. At each scale the validation criteria of Section 3.3 are applied for each window orientation. We denote the resulting algorithm multi-scale multi-window (MSMW).

Importance of multi-window. Figure 8 compares the results of ZSSD block-matching using 5×5 windows and the proposed multi-window algorithm using 5 and 9 window orientations,

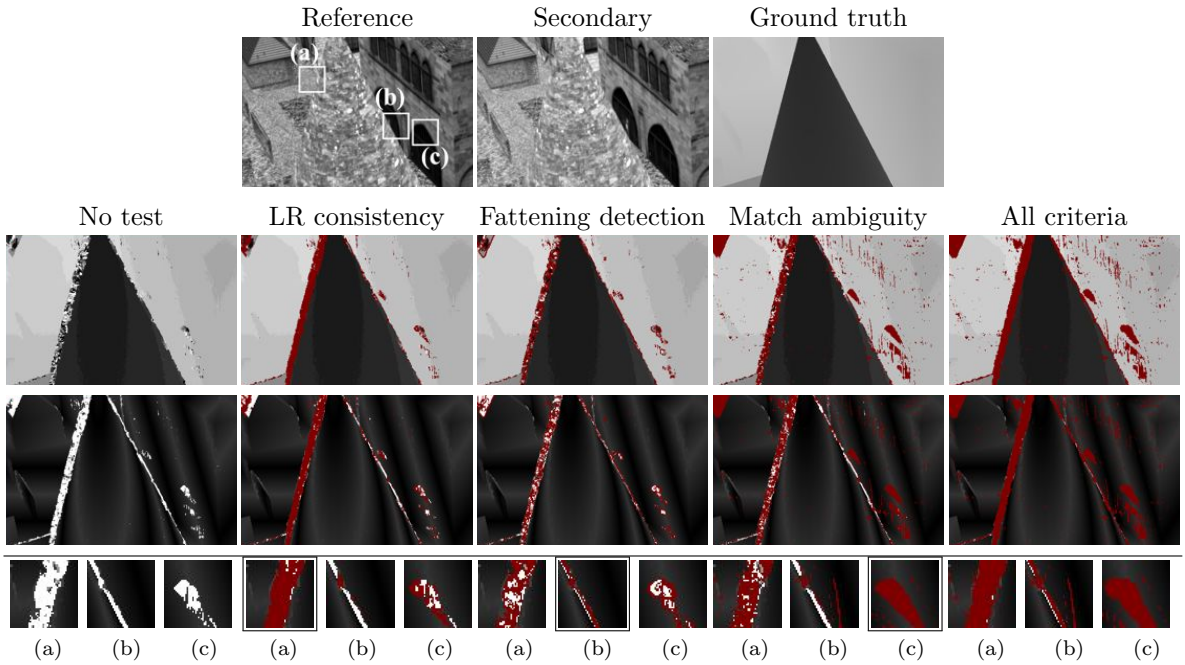


Figure 7. *Application of reject criteria with a 5×5 patch and a single scale. The disparity is displayed with rejected pixels by each criterion marked in red. From top to bottom and left to right: pair images, ground truth, disparity with no criteria applied, LR consistency, fattening detection, self-similarity and all criteria at once. The third row shows the corresponding errors computed with the ground truth, displayed in range [0, 2.5]. The last row shows the details (a), (b), and (c) for each one of the results. The extract (a) contains an occlusion, (b) contains a dis-occlusion, and (c) contains a repetitive pattern.*

where all windows have roughly the same area. We note that more pixels are correctly matched nearer the occlusions and dis-occlusions boundaries and on slanted surfaces of the scene. The orientations of the windows, illustrated with the color code of Figure 4, coincide as expected with the direction of least variation of the disparity.

Importance of the rejection criteria. The rejection criteria of Section 3.3 are straightforwardly adapted to any window shape. Moreover, we observed that the matches are much easily validated for the correct window shape, thus improving the overall performance of the rejection stage. For each window shape, the matching and rejection stages are applied independently. Then for each pixel the choice for the optimal window shape is made considering only the window orientations validated at that pixel. Algorithm 2 details the complete process. It is important to note that the rejection criteria are applied in a specific order, so that once a point has been rejected by a test it is no longer considered in the subsequent tests.

Importance of multi-scale. In Figure 9 we see that the advantages of the multi-scale add to those of the multi-window matching: the results are denser and there are fewer errors. Indeed the narrow windows allow to pick up narrow features at coarser scales, thus reducing the corresponding search ranges and considerably improving the matching. The multi-scale strategy also significantly reduces the disparity computation time.

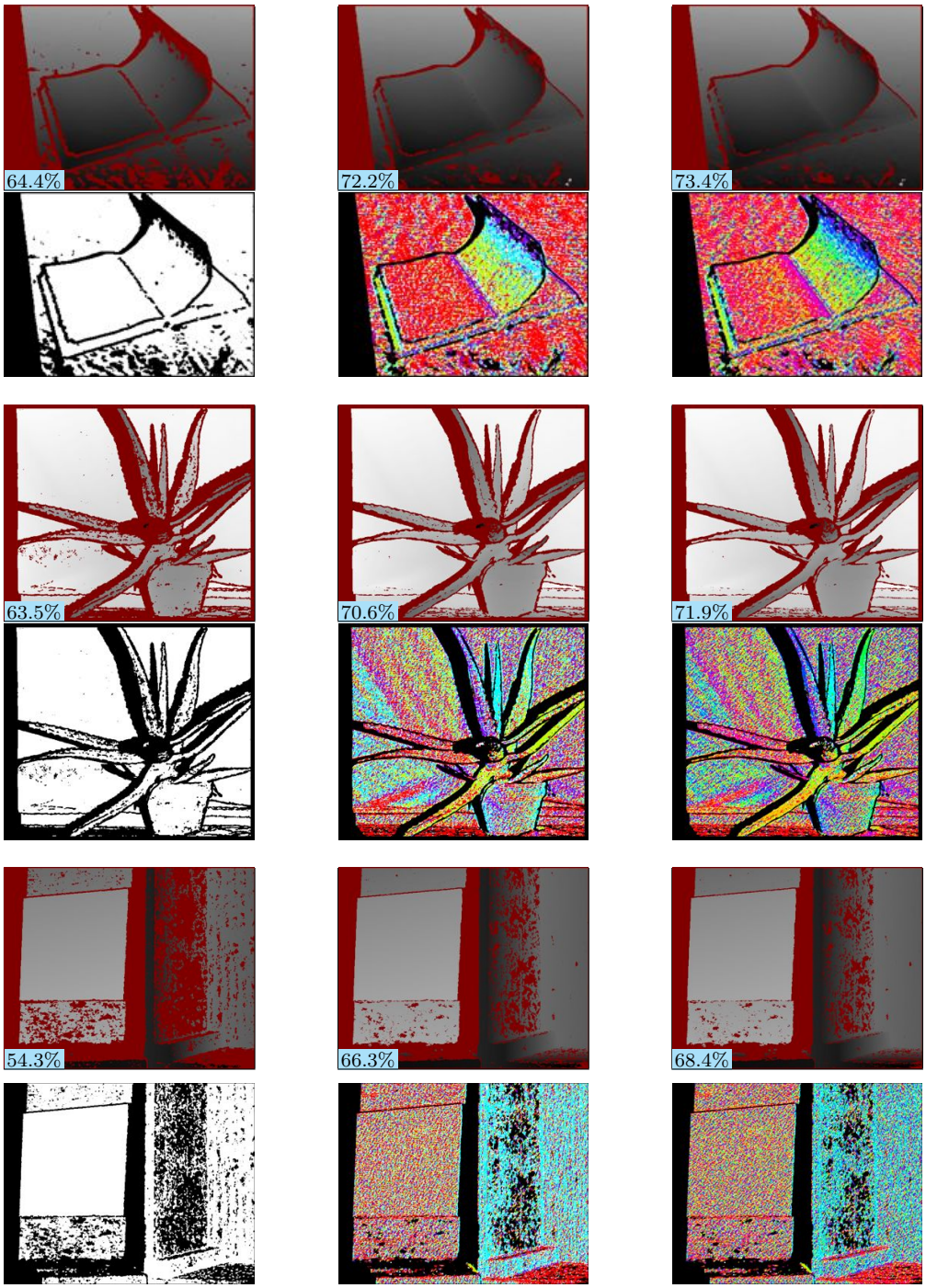


Figure 8. *Comparison of single and multiple window orientations (*book, aloe, *and* wood *pairs from Figure 11). Using respectively 1, 5 or 9 windows orientations with support equivalent to* 5×5 *windows, using all filters and exhaustive search with precision of 1/4 of pixel. This figure illustrates how the elongated windows adapt to the surface orientations, thus improving the matching and leading to denser results. The density of each disparity map is indicated in the lower left corner.*

Algorithm 2: Multi-window block-matching algorithm. $dMin$ and $dMax$ are images indicating the search ranges for all the pixels. $disp$ contain both disparity maps LR and RL.

$MultiWindowMatching(u,\ v,\ dMin,\ dMax)$:

Input: image pair u and v.
Input: min & max disp. range images $dMin$ and $dMax$, minimum and maximum disparity for each pixel.

Output: Fixed precision disparity $disp$.

for each window **w** **do**
  `// Compute left and right disparity maps`
  $disp_{\mathbf{w}} = ZSSDMatching(u,\ v, dMin, dMax, \mathbf{w})$

  `// Apply the rejection criteria to` $disp_{\mathbf{w}}$
  Update $disp_{\mathbf{w}}$ applying the Fattening detection
  Update $disp_{\mathbf{w}}$ applying the Match Ambiguity detection
  Update $disp_{\mathbf{w}}$ applying the LR criterion
  Update $disp_{\mathbf{w}}$ removing the ISOLATED MATCHES
end for

$disp$ = Combine all the $disp_{\mathbf{w}}$

`// Ensure consistency of` $disp$ `after combining`
Update $disp$ applying the LR criterion
Update $disp$ removing the ISOLATED MATCHES
Update $disp$ FILLING-IN small holes (optional)

Filling-in small holes. An optional post-process is applied in order to fill-in small holes in the disparity map. A hole is a connected component of rejected pixels (with 4-connectivity). For each hole an affine model is estimated from its boundary values using least squares. A hole is interpolated only if the regression model has a slope below 5 and fits the boundary disparities with an error below 3. This usually removes only small holes in the disparity map without introducing errors (see Figure 10).

4. Experiments and comparison with the state-of-the-art. In this section we compare the proposed algorithm with methods from the state-of-the-art. Even if our focus is the estimation on real scenes, we first evaluate using studio cases for which the ground truth is available. Then, we compare and discuss the results on satellite and urban images. Satellite images are courtesy of the French Spatial Agency (CNES) while the urban images are from the KITTI benchmark [21]. We consider the following methods as representatives of the state-of-the-art.

- The *graph cuts* method (GC) introduced in [30] is representative of global optimization methods. The algorithm writes as the minimization of a discrete energy containing three terms: one penalizing the pixel-wise radiometric differences between the two

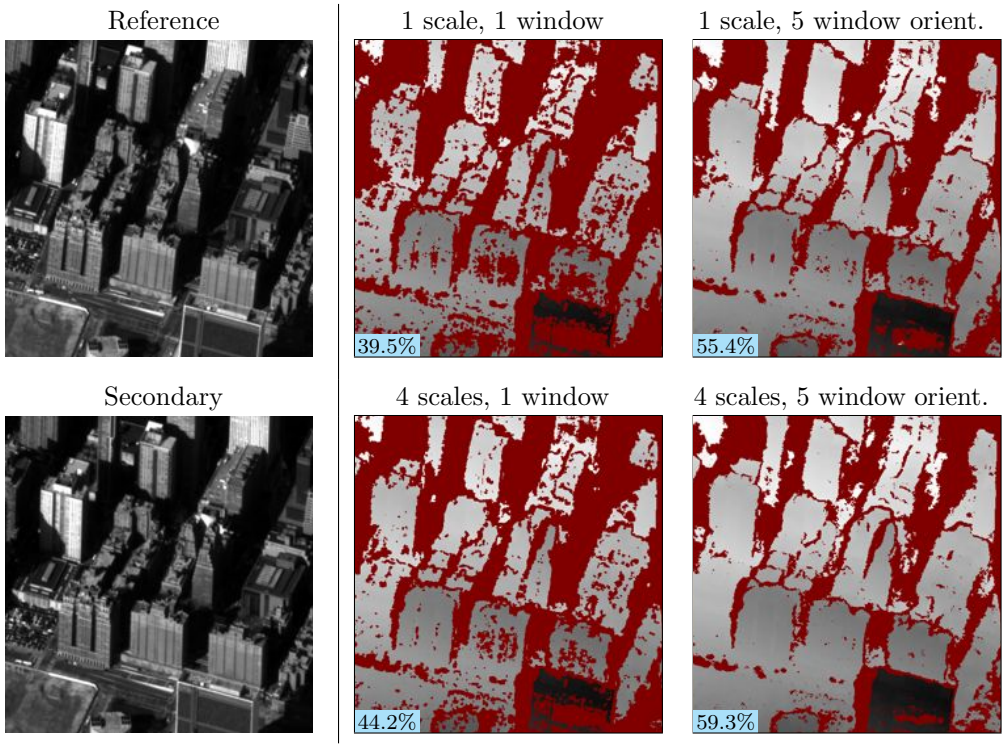


Figure 9. *Interaction between multi-scale and multi-window matching. In the left-most column is shown the image pair. The second column shows results obtained using a* 5×5 *pixels window with 1 and 4 scales, note that the multi-scale result is denser. The third column shows results obtained using windows with 5 orientations with 1 and 4 scales. The density of each disparity map is indicated in the lower left corner. With 5 orientations the single scale result is already denser than the one obtained with a single squared window, and the multi-scale result is even denser. The multi-scale algorithm is also faster than the single scale one.*

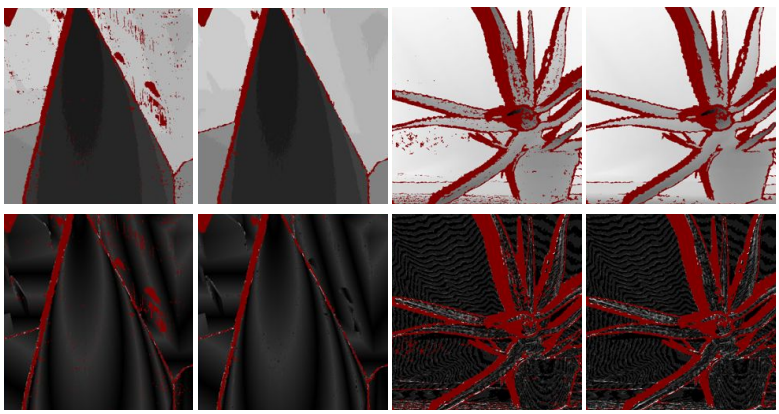


Figure 10. *Filling-in small holes. This post-processing step removes small holes from the disparity maps by extrapolating from the boundary of the hole. In the first row are shown the computed disparity maps before and after the filling-in, in the second row are shown the corresponding error maps represented in the range [0, 2.5]. The disparity map has been computed with a single scale, single window and pixel precision SSD algorithm combined with the proposed reject tests.*

images, a second one imposing a penalty for making a pixel occluded, and a smoothness term forcing neighboring pixels to have similar disparities. Two parameters must be set in the graph cuts method, one for setting the regularity of the solution and one for the amount of occlusions to be detected.

- Cech and Sara's *growing correspondence seeds* (GCS) [10] and the *five regions correlator* (5REG) proposed by Hirschmuller et al. [26] are representative of local non dense methods. The GCS algorithm, as ours, favors reliability over completeness of the results. The algorithm performs a region growing from "sure" matches chosen as Harris interest points. The 5REG algorithm considers five overlapping 5×5 sub-windows for each pixel and the matching cost is determined adding the cost of the center sub-window plus the two sub-windows yielding the smallest costs. The matches are then filtered applying the Left-Right consistency test and comparing the matching cost with the cost of the second local minimum. We used the implementation of 5REG provided in [1] which lacks the fattening refinement described in the original paper [26].
- The *cost-volume filtering* method (CVF) [39] is representative of adaptive window algorithms. These methods use large comparing windows, for the CVF a 17×17 window is used. The window support is then adapted taking into account the reference image in order to select pixels belonging to the same object. The CVF algorithm also proposes a densification post-processing, which is disabled in the comparison. We used the implementation from Tan and Monasse [49] of CVF which permits to disable the densification stage of the algorithm.
- Hirschmuller's *semi global matching* (SGM) [25] is a strategy for minimizing a global energy that comprises a pixel-wise matching cost and a global smoothness constraint. In SGM the two-dimensional smoothness constraint is efficiently approximated as the average of one-dimensional line optimization problems. We use here the implementation of SGM [25] included in OpenCV[3].

For MSMW we use 9 window orientations with areas comparable to a 5×5 pixel window (the windows are shown in Figure 4), for the multi-scale we consider 4 levels. As commented in section 3 the subpixel precision is set to $p = 1/4$ and the threshold for the fattening detection is set to $s = 1$. We consider two variants of MSMW one without post-processing and a second, denoted MSMW w/fill, including the optional filling-in of small holes from section 3.4.

4.1. Evaluation on Stereo Pairs with Ground Truth. In Figures 12 – 13 we compare the results of the methods mentioned above for stereo pairs with ground truth, both real [27, 45, 47] and synthetic [37, 40]. The images show the computed disparities and the errors with respect to the ground truth. The reference images and ground truth are shown in Figure 11.

We observe that the graph cuts algorithm fails at detecting the occlusions in most of the examples with complex geometry (i.e. the leaves in the Aloe image in Figure 12). We also observe an over-quantization of the disparity computed by graph cut. This is due to the fact that discontinuities of disparity are penalized independently of its amplitude in the regularity term, see [30] for more details. Both, the GCS and 5REG algorithms produce fairly dense

[3] *StereoSGBM* module in OpenCV 2.4.8 (`http://opencv.org`) and with default parameters. To filter more outliers we compute a second disparity map reversing the reference and secondary images and enforce the consistency of both maps [25].

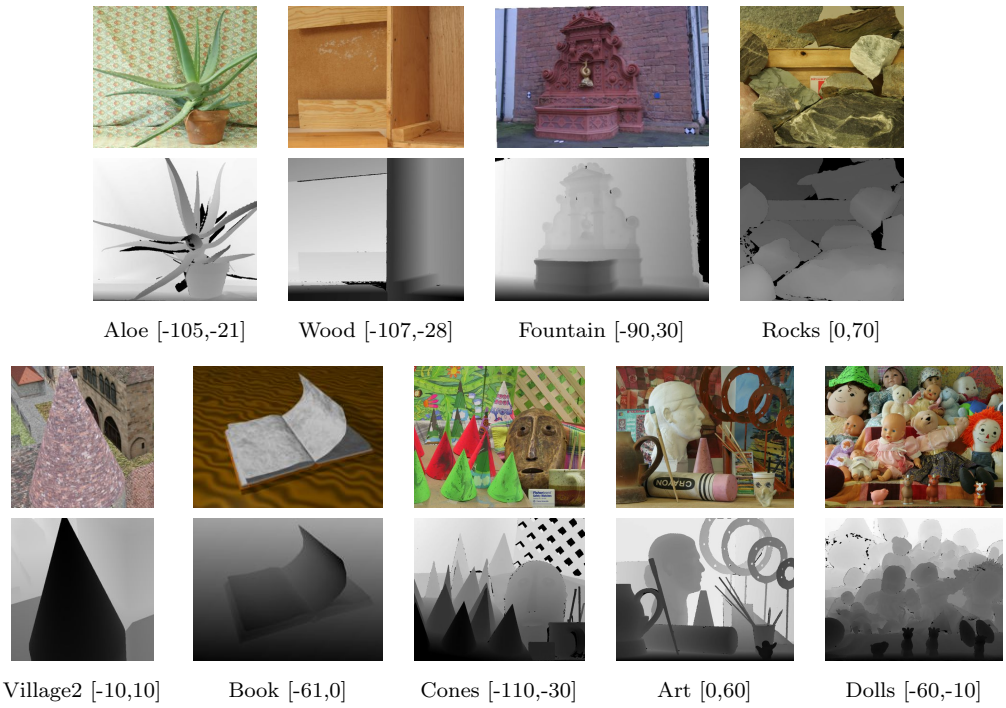


Figure 11. *Reference images and ground truths for the stereo pairs used in the experimentation section. For each stereo pair the disparity range* $[dMin, dMax]$ *is indicated in the caption.*

results. However, they contain many spurious matches and false detections (i.e. the background of the Aloe image in Figure 12). Both use ambiguity detection criteria based on fixed thresholds of the ratio between first and second local minima of the correlation function. In comparison, the proposed match ambiguity criterion combined with the multi-scale scheme of MSMW significantly reduces the mismatches while obtaining denser results. The CVF, SGM, and MSMW produce sharp disparity maps consistent with the image geometry. However, CVF has a limited performance on self-similar regions (upper-left corner of the Village2 image in Figure 13). This is again due to the lack of ambiguity detection.

The quality of the results of MSMW comes close to SGM and CVF on textured regions and close to the object edges (as in the Aloe image). Only, the textureless areas such as in the Cones image are handled worse. On slanted surfaces the results are much denser than the other algorithms.

Quantitative analysis. To compare the methods we evaluate the density of the disparity maps and the number of mismatches produced by each method. Note that a density of 100% cannot be attained unless the disparity maps are interpolated. A mismatch is a point not rejected by an algorithm and whose disparity differs from the ground truth by more than a certain threshold. Two mismatch thresholds are considered: 1 and 3 pixels. We compute these indicators on three regions of each image [44]: all the pixels (ALL), the non-occluded pixels (NONOCC), and the occluded pixels (OCC). Errors at occluded pixels can be computed because a the ground-truth is known also in occluded areas.

In table 1 we report the results of the algorithms for each image, evaluated on the region

containing with ALL the pixels. For brevity the tables corresponding to OCC and NONOCC are not reported as the same conclusions can be drawn from the averages seen in table 2. It is observed that the proposed algorithm attains the lowest mismatch rate on non-occluded pixels, while keeping a relatively high density. Beside the GC method, which is a global, the highest densities are attained by SGM and MSMW with small differences between the two. The 5REG method also yields high densities but at the expense of a higher mismatch rates. GCS has low mismatch rates but also low density. We can confirm that the filling in post process for MSMW increases the final density without too much impact on the mismatch statistics. We also see that for strongly fronto-parallel scenes the CVF method yields the lowest errors, while for the scenes with slanted surfaces the error is much lower for MSMW.

Performance on slanted surfaces. The issue of non-frontal surfaces is evident on the ground planes of the Aloe, Wood, and Fountain images in Figure 12. The adaptive window methods such as CVF are bound to the fronto-parallel hypothesis. Although the shape of the window is determined from the image content, its is not necessarily adapted to the geometry of the surface, leading to poor results.

A bias towards the constant disparity is also observed in the results of SGM. This is probably due to the discontinuity penalty term P_1 in the line optimization of SGM [25]. On the same slanted surfaces the results of MSMW are denser and have less mismatches than all the consider methods.

Errors on occluded areas. We have observed that many of the errors of MSMW come from the occlusion regions (OCC) in the scenes. According to table 2 all the window based methods (except for CVF) have a comparable performance on these regions. These errors are manly due to the fattening effect, and the adaptive windows used by CVF greatly reduce it. The fattening detection used in MSMW also mitigates the fattening errors, which is lower than GC, 5REG and SGBM. Yet, it is clear that adaptive windows constitute a powerful tool when it comes to removing the foreground fattening, and its incorporation in the current framework must be subject of future works.

4.2. Evaluation on Stereo and Tristereo Satellite Images. One of the interests of this paper is the estimation of digital elevation models from satellite images. The existence of high-resolution quasi-simultaneous satellite imagery is leading to an increase in the demand of automatically generated and reliable digital elevation models.

Figures 14-15 compare the performance of previous algorithms applied to stereo pairs kindly provided by the CNES (Centre national d'Études spatiales). The first one is an airborne pair taken over Toulouse, while the second is a view of the Grand Central Terminal in New York acquired by the *Pléiades satellite*. These experiments illustrate that only SGM and MSMW are able to produce a reasonable disparity map free of artifacts. Both solutions are very similar for the nadir aerial images in Fig. 14, while MSMW performs better on the slanted view of Grand Central, Fig. 15, yielding a denser result.

Figure 16 shows an application of the proposed algorithm to compute a digital elevation model from a tri-stereo dataset. These images have been acquired by the *Pléiades satellite* and are courtesy of the CNES and Astrium DS. Two disparity maps are computed forming nadir-left and nadir-right stereo pairs using the s2p pipeline [13]. Note that the images contain significant occluded regions in the vicinity of high structures. These occlusions are inevitable

Table 1

Evaluation of the results. The table below compares the outputs of the proposed algorithm (MSMW without and with the filling-in post-process) with: the Graph Cuts algorithm (GC) from Kolmogorov and Zabih [30], Cech and Sara's [10] Growing Correspondence Seeds (GCS), the 5 regions algorithm (5REG) from Hirschmüller et al. [26], the Cost-Volume Filtering algorithm (CVF) from Rhemann et al. [39], and SGBM (OpenCV's implementation of Hirschmuller's Semi Global Matching (SGM) [25]). We report the density (D) of the obtained disparity maps and the mismatches as percentage of pixels yielding errors above one pixel (E1) and above 3 pixels (E3).

	GC [30]			**GCS [10]**			**5REG [26]**		
Image	**D**	**E1**	**E3**	**D**	**E1**	**E3**	**D**	**E1**	**E3**
Aloe	90.53	14.16	6.52	73.72	4.26	2.24	79.10	2.91	1.26
Wood	87.12	25.83	10.60	73.73	1.72	0.34	75.38	2.21	0.61
Fountain	92.36	36.50	8.88	72.87	6.57	1.36	83.38	4.99	1.22
Rocks	97.61	2.07	0.62	91.89	1.67	0.41	96.32	0.95	0.38
Cones	89.79	9.53	4.72	65.09	4.00	2.02	68.20	3.83	2.33
Art	94.13	11.97	5.13	68.38	5.30	2.80	76.34	4.54	3.59
Dolls	96.93	9.70	1.37	73.59	5.22	1.45	82.11	4.73	2.01
Book	88.80	31.01	1.35	49.80	4.88	1.86	83.43	6.27	1.91
Village2	97.47	0.27	0.22	95.53	0.70	0.44	96.95	0.42	0.39
average	92.748	15.67	4.38	73.84	3.81	1.43	82.36	3.43	1.52

	CVF [39]			**SGBM [25]**			**MSMW**			**MSMW w/fill**		
Image	**D**	**E1**	**E3**	**D**	**E1**	**E3**	**D**	**E1**	**E3**	**D**	**E1**	**E3**
Aloe	80.23	2.00	1.08	82.77	1.82	0.98	**83.59**	**1.62**	**0.78**	**83.94**	1.72	**0.80**
Wood	78.29	1.74	0.63	82.46	**0.99**	0.39	**83.50**	1.20	**0.16**	**84.15**	1.26	**0.17**
Fountain	81.95	5.65	1.34	87.65	**4.38**	0.94	90.25	5.68	**0.76**	**92.39**	6.25	**0.81**
Rocks	92.92	**0.47**	**0.23**	95.75	0.83	0.37	94.64	0.86	0.25	**96.38**	1.01	0.26
Cones	78.27	**1.19**	**0.90**	**86.32**	2.29	1.64	78.21	4.06	1.66	81.63	4.49	1.73
Art	75.50	**1.87**	**1.52**	**88.75**	3.72	3.02	82.03	3.11	1.85	85.30	3.34	1.94
Dolls	84.74	**1.88**	**0.70**	**95.53**	3.17	1.08	86.08	3.76	1.07	90.72	4.15	1.10
Book	79.22	**1.93**	**0.11**	86.87	4.33	0.99	86.51	3.50	0.48	**87.52**	3.67	0.49
Village2	96.31	0.38	0.38	96.07	0.32	0.2	**97.24**	**0.23**	**0.14**	**97.35**	0.31	**0.14**
average	83.05	**1.90**	**0.76**	**89.13**	2.43	1.07	86.89	2.67	0.79	88.82	2.91	0.83

Table 2

Averages of disparity map density (D) and mismatch rates as percentage of pixels yielding errors above one pixel (E1) and above 3 pixels (E3). These values are computed for all the pixels (ALL), the non-occluded pixels (NONOCC), and the occluded pixels (OCC) [44]. We compare results of MSMW (without and with the filling-in post-process) with those of the Graph Cuts (GC) [30], Growing Correspondence Seeds (GCS) [10], Hirschmüller's 5 regions algorithm (5REG) [26], the Cost-Volume Filtering algorithm (CVF) [39], and SGBM (OpenCV's implementation of Semi Global Matching [25]).

	NONOCC			OCC			ALL		
Method	D	E1	E3	D	E1	E3	D	E1	E3
GC [30]	**98.11**	15.72	3.62	18.83	16.97	13.95	**92.75**	15.67	4.38
GCS [10]	78.17	3.30	0.92	14.21	13.09	9.44	73.84	3.81	1.44
5REG [26]	87.52	3.04	1.11	12.31	11.19	8.17	82.36	3.43	1.52
CVF [39]	88.66	**1.74**	0.61	7.50	**5.60**	**3.65**	83.05	**1.90**	**0.77**
SGBM [25]	**94.87**	1.92	0.64	**12.96**	11.94	8.15	**89.13**	2.43	1.07
MSMW	92.62	2.32	**0.47**	11.20	10.12	6.33	86.89	2.67	**0.79**
MSMW w/fill	94.62	2.54	**0.49**	12.28	11.08	6.61	88.82	2.91	0.83

Averages considering the images: Aloe, Wood, Fountain, Rocks, Cones, Art, Dolls, Book, and Village2.

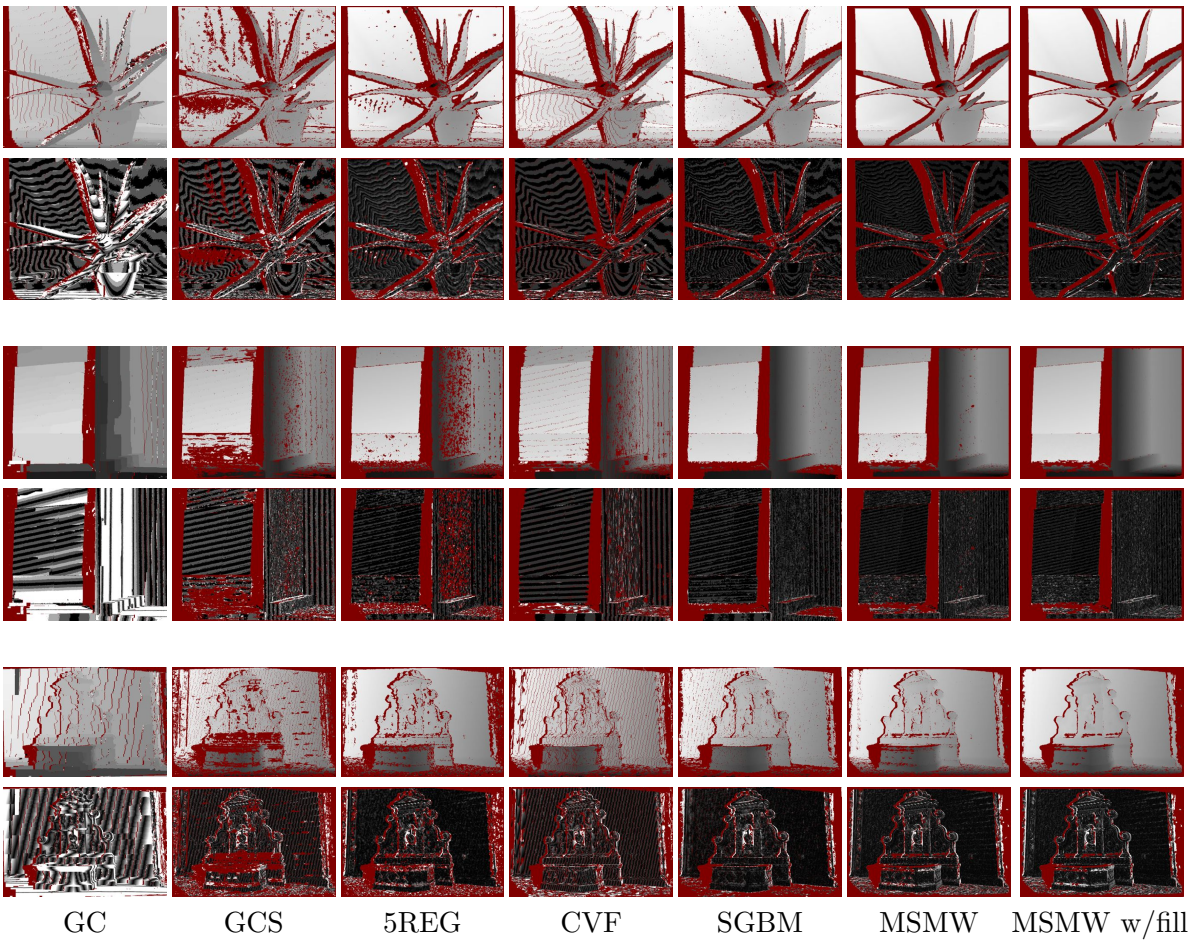


Figure 12. *Comparison of results with Aloe, Wood, and Fountain. Top, from left to right: disparity and error computed by the graph cut method (GC), the GCS, the 5REG, CostVolume, SGBM (OpenCV's implementation of SGM), and the MSMW without and with filling-in postprocess. Below: image error between computed disparities and ground truth displayed in range [0, 2.5].*

in satellite imaging. However since these regions are complementary in the two datasets it is possible to fuse the models. In order to take advantage of this complementarity, the stereo algorithm must produce few outliers, favoring correctness over density.

4.3. Evaluation on the KITTI benchmark. We finally compare the algorithms with the images provided in the benchmark KITTI [21]. These images are taken by the *AnnieWAY* autonomous driving platform, which is equipped with a stereo rig and a laser scanner for capturing accurate ground truth data. Images are captured in a mid-size city and contain rural areas and highways, including cars and pedestrians. Figure 17 displays a comparison with one of the KITTI images.

In order to rank methods, the density of disparity maps and the percent of errors over a certain threshold, typically two or three pixels, are used. A second classification is performed on the dense disparity maps obtained by applying a standard disparity interpolation to the

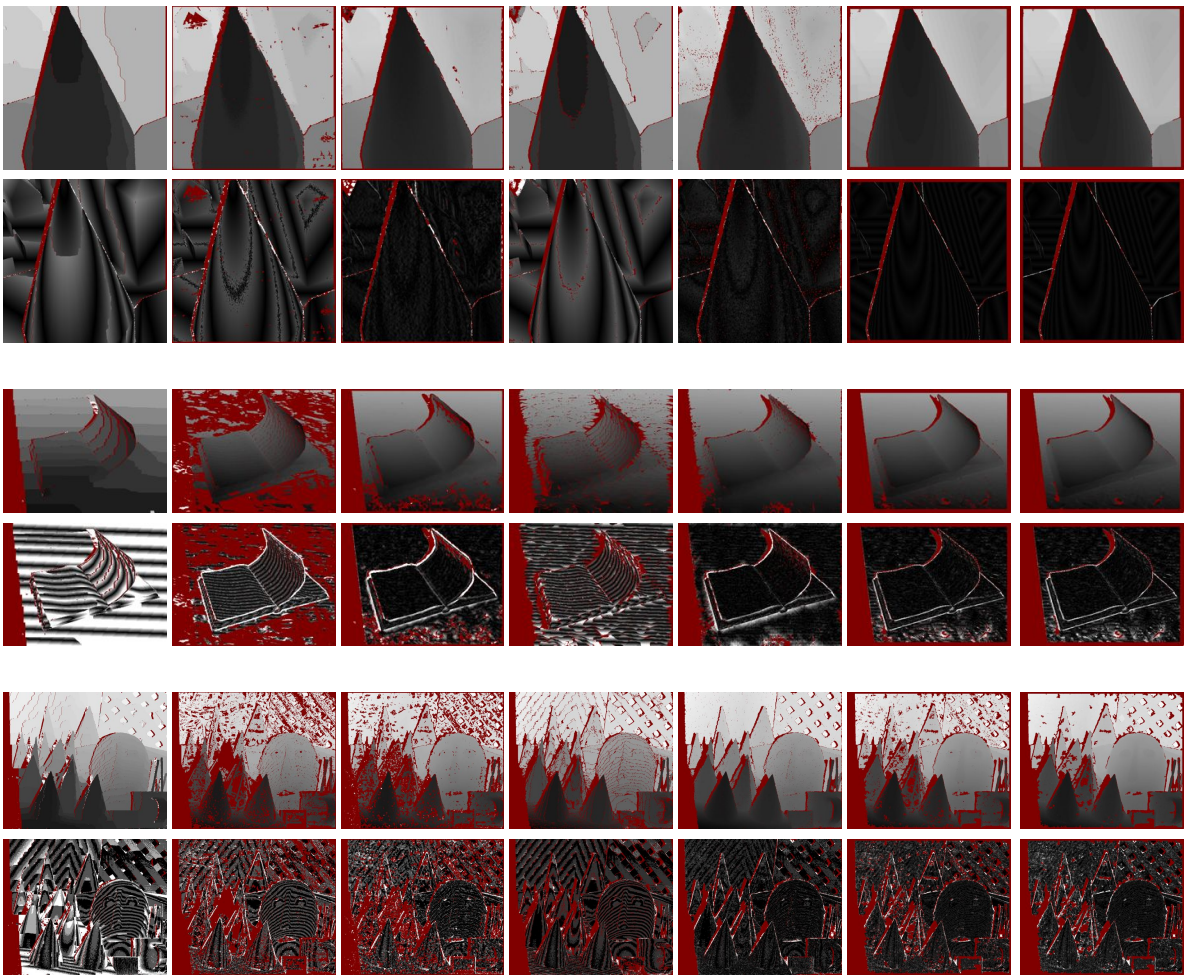
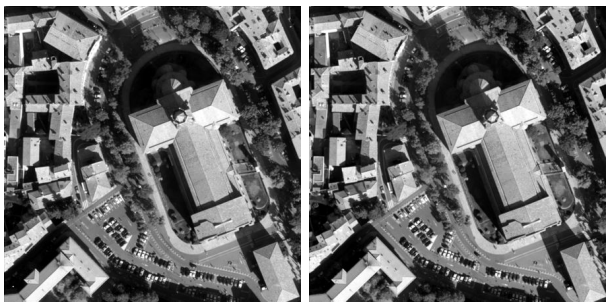


Figure 13. *Comparison of results with Village2, Book, and Cones. Top, from left to right: disparity and error computed by the graph cut method (GC), the GCS, the 5REG, CostVolume, SGBM (OpenCV's implementation of SGM), and the MSMW without and with filling-in postprocess. Below: image error between computed disparities and ground truth displayed in range [0, 2.5].*

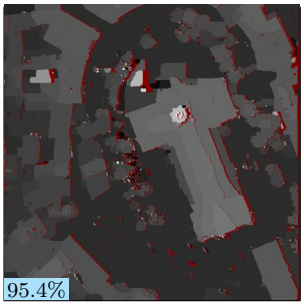
submitted results. More specifically, the disparity is interpolated line by line, and each hole is filled with the furthest value. The method proposed is this paper is currently ranked in first position, based on the estimated disparity.

5. Conclusion. We have proposed a multi-window and multi-scale block-matching algorithm for stereo. The proposed method uses oriented windows to minimize the mismatches on surfaces that don't verify the fronto-parallel assumption. The oriented windows align with the 3D geometry of the scene and unlike adaptive windows permit to match correctly on slanted surfaces.

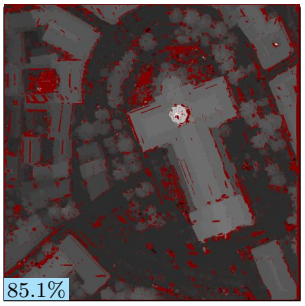
Two sources of mismatches are also identified, namely the ambiguous matches and the fattening artifacts due to scene discontinuities. Validation criteria have been incorporated



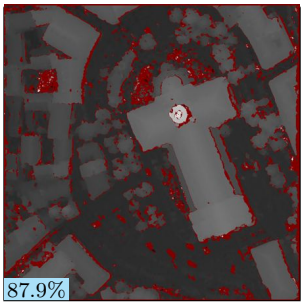
(a) stereo pair



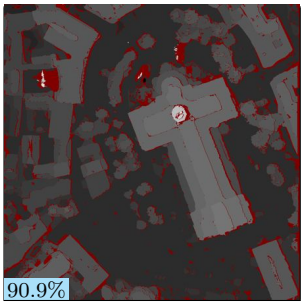
(b) GC



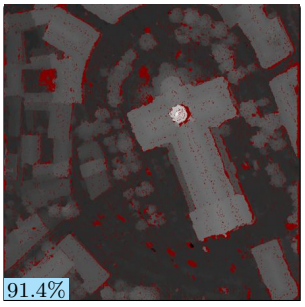
(c) GCS



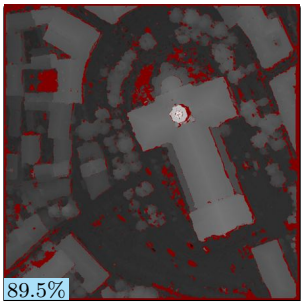
(d) 5REG



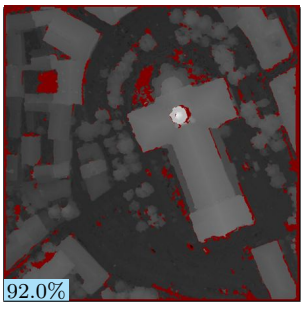
(e) CVF



(f) SGBM



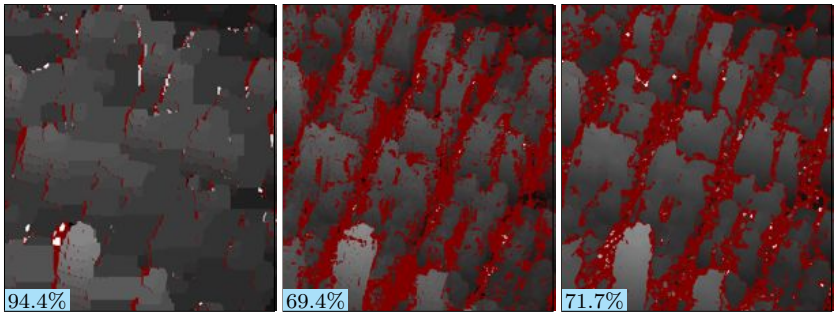
(g) MSMW



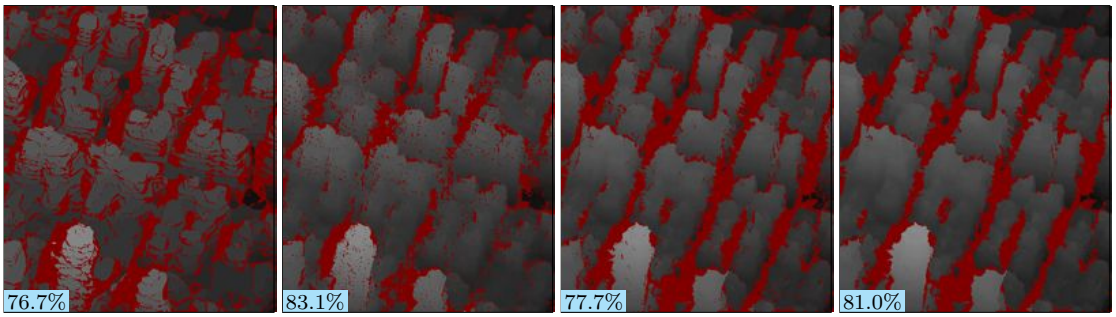
(h) MSMW w/fill

Figure 14. *Experiments on airborne pair kindly provided by the CNES* (Centre national d'études spatiales). *From left to right: the stereo pair, disparity computed by the graph cut method, the GCS, the 5REG, CostVolume, SGBM (OpenCV's implementation of SGM), and the MSMW (without and with the filling-in post-process). The density of each disparity map is indicated in the lower left corner. The disparity range for this pair is [-5, 30].*

(a) stereo pair



(b) GC (c) GCS (d) 5REG



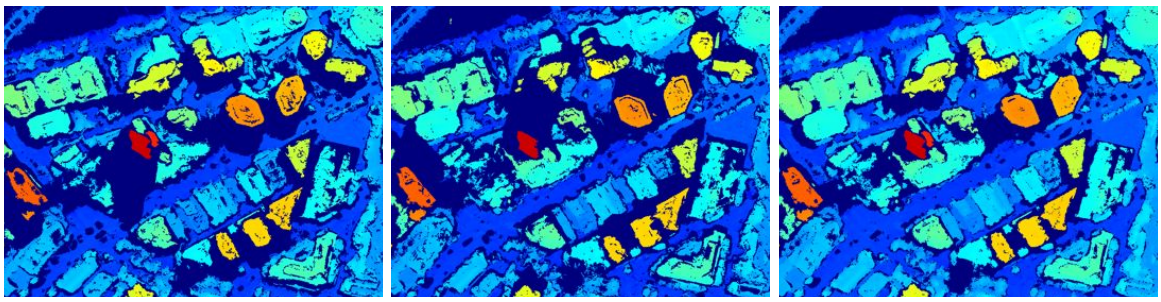
(e) CVF (f) SGBM (g) MSMW (h) MSMW w/fill

Figure 15. *Experiments on satellite stereo pair kindly provided by the CNES* (Centre national d'études spatiales). *From left to right: the stereo pair, disparity computed by the graph cut method, the GCS, the 5REG, CostVolume, SGBM (OpenCV's implementation of SGM), and the MSMW (without and with the filling-in post-process). The density of each disparity map is indicated in the lower left corner. The disparity range for this pair is [-15, 50].*

to detect mismatches caused by any of these sources. Extensive evaluation on benchmark datasets and on real images show that the proposed method has an excellent performance in terms of density and mismatch rates.

The natural continuation of the present work should, on one hand, explore the combination of information from windows of various sizes. On the other hand, improve the matching process by combining oriented windows with image adaptive windows.

(a) left image (b) right image (c) nadir image



(d) DEM from nadir-left pair (e) DEM from nadir-right pair (f) fused DEM

Figure 16. *Tristereo Pléiades satellite images of Melbourne (a-c). In (d) and (e) are shown the digital elevation models obtained with MSMW by taking image pairs with a fixed reference image (nadir), while (f) corresponds to the fusion of the two. The models are represented in false color to improve visualization, blue areas represent rejected pixels.*

REFERENCES

[1] P. Abeles, *Boofcv*. http://boofcv.org/, 2012.

[2] R.D. Arnold, *Automated stereo perception.*, tech. report, DTIC Document, 1983.

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D.B. Goldman, *PatchMatch: a randomized correspondence algorithm for structural image editing*, in SIGGRAPH '09: ACM SIGGRAPH 2009 papers, New York, NY, USA, 2009, ACM, pp. 1–11.

[4] S. Birchfield and C. Tomasi, *A pixel dissimilarity measure that is insensitive to image sampling*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (1998), pp. 401–406.

[5] G. Blanchet, A. Buades, B. Coll, J.-M. Morel, and B. Rougé, *Fattening free block matching*, Journal of mathematical imaging and vision, 41 (2011), pp. 109–121.

[6] Michael Bleyer and Margrit Gelautz, *A layered stereo matching algorithm using image segmentation and global visibility constraints*, ISPRS Journal of Photogrammetry and Remote Sensing, 59 (2005), pp. 128–150.

[7] Michael Bleyer, Christoph Rhemann, and Carsten Rother, *PatchMatch stereo - stereo matching with slanted support windows*, in Procedings of the British Machine Vision Conference 2011, British Machine Vision Association, 2011, pp. 14.1–14.11.

[8] A.F. Bobick and S.S. Intille, *Large occlusion stereo*, International Journal of Computer Vision, 33 (1999), pp. 181–200.

[9] Vicent Caselles and Pascal Monasse, *Grain filters*, Journal of Mathematical Imaging and Vision, 17 (2002), pp. 249–270.

[10] J. Cech and R. Šára, *Efficient Sampling of Disparity Space for Fast And Accurate Matching*, in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, IEEE, June 2007,

pp. 1–8.

[11] S.D. Cochran and G. Medioni, *3-D surface description from binocular stereo*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 14 (1992), pp. 981–994.

[12] C. de Franchis, G. Facciolo, and E. Meinhardt, *s2p on line demo*. http://dev.ipol.im/~carlo/s2p/, 2014.

[13] C. de Franchis, E. Meinhardt, J. Michel, J.-M. Morel, and G. Facciolo, *An automatic and modular stereo pipeline for pushbroom images*, in Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (SPRS), 2014.

[14] Julie Delon and Bernard Rougé, *Small baseline stereovision*, Journal of Mathematical Imaging and Vision, 28 (2007), pp. 209–223.

[15] A. Desolneux, L Moisan, and J.-M. Morel, *From gestalt theory to image analysis: a probabilistic approach*, vol. 34, Springer, 2007.

[16] F. Devernay and O. Faugeras, *Computing differential properties of 3-D shapes from stereoscopic images without 3-D models*, in Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, IEEE, June 1994, pp. 208–213.

[17] G. Egnal, *A stereo confidence metric using single view imagery with comparison to five alternative approaches*, Image and Vision Computing, 22 (2004), pp. 943–957.

[18] M. A. Fischler and R. C. Bolles, *Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography*, Commun. ACM, 24 (1981), pp. 381–395.

[19] P. Fua, *A parallel stereo algorithm that produces dense depth maps and preserves image features*, Machine Vision and Applications, 6 (1993), pp. 35–49–49.

[20] A. Fusiello, V. Roberto, and E. Trucco, *Efficient stereo with multiple windowing*, International Journal of Pattern Recognitiond and Artificial Intelligence, 14 (1997), pp. 858–863.

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, *Vision meets robotics: The kitti dataset*, International Journal of Robotics Research (IJRR), (2013).

[22] Mark Gerrits and P. Bekaert, *Local stereo matching with segmentation-based outlier rejection*, in Computer and Robot Vision, 2006. The 3rd Canadian Conference on, IEEE, June 2006, p. 66.

[23] A. W. Gruen, *Adaptive least squares correlation: a powerful image matching technique*, South African Journal of Photogrammetry, Remote Sensing and Cartography, 14 (1985), pp. 175–187.

[24] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second ed., 2004.

[25] Heiko Hirschmüller, *Stereo processing by semiglobal matching and mutual information*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (2007), pp. 328–341.

[26] H. Hirschmüller, P.R. Innocent, and J. Garibaldi, *Real-time correlation-based stereo vision with reduced border errors*, International Journal of Computer Vision, 47 (2002), pp. 229–246.

[27] H. Hirschmüller and D. Scharstein, *Evaluation of stereo matching costs on images with radiometric differences*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (2009), pp. 1582–1599.

[28] X. Hu and P. Mordohai, *Evaluation of stereo confidence indoors and outdoors*, in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, June 2010, pp. 1466–1473.

[29] T. Kanade and M. Okutomi, *A stereo matching algorithm with an adaptive window: Theory and experiment*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16 (1994), pp. 920–932.

[30] V. Kolmogorov and R. Zabih, *Computing visual correspondence with occlusions using graph cuts*, in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 2, IEEE, 2001, pp. 508–515.

[31] J.J. Little and W.E. Gillett, *Direct evidence for occlusion in stereo and motion*, in Computer Vision — ECCV 90, O. Faugeras, ed., vol. 427 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1990, pp. 336–340.

[32] J. Lotti and G. Giraudon, *Correlation algorithm with adaptive window for aerial image in stereo vision*, In Image and Signal Processing for Remote Sensing, 1 (1994), pp. 2315–10.

[33] D. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60 (2004), pp. 91–110.

[34] Jiangbo Lu, Hongsheng Yang, Dongbo Min, and M. N. Do, *Patch match filter: Efficient Edge-*

Aware filtering meets randomized search for fast correspondence field estimation, in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, June 2013, pp. 1854–1861.

[35] R. Manduchi and C. Tomasi, *Distinctiveness maps for image matching*, Proceedings of the International Conference on Image Analysis and Processing, (1999), pp. 26–31.

[36] S. Masnou and J.-M. Morel, *Image restoration involving connectedness*, 3346 (1998), pp. 84–95.

[37] L. Moisan, *Paires stéréo simulées*. personal communication, 2010.

[38] M.P. Patricio, F. Cabestaing, O. Colot, and P. Bonnet, *A similarity-based adaptive neighborhood method for correlation-based stereo matching*, in International Conference on Image Processing, vol. 2, 2004, pp. 1341–1344.

[39] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, *Fast cost-volume filtering for visual correspondence and beyond*, in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 3017–3024.

[40] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N.A. Dodgson, *Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid*, in Computer Vision–ECCV 2010, Springer, 2010, pp. 510–523.

[41] Neus Sabater, Andrés Almansa, and J-M Morel, *Meaningful matches in stereovision*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34 (2012), pp. 930–942.

[42] Neus Sabater, J-M Morel, and Andrés Almansa, *How accurate can block matches be in stereo vision?*, SIAM Journal on Imaging Sciences, 4 (2011), pp. 472–500.

[43] D. Scharstein and R. Szeliski, *Stereo matching with nonlinear diffusion*, International Journal of Computer Vision, 28 (1998), pp. 155–174.

[44] ———, *A taxonomy and evaluation of dense two-frame stereo correspondence algorithms*, International journal of computer vision, 47 (2002), pp. 7–42.

[45] ———, *High-accuracy stereo depth maps using structured light*, in Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition, CVPR'03, Washington, DC, USA, 2003, IEEE Computer Society, pp. 195–202.

[46] Anselm Spoerri, *The Early Detection of Motion Boundaries*, PhD thesis, 1990.

[47] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, *On benchmarking camera calibration and multi-view stereo for high resolution imagery*, in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–8.

[48] R. Szeliski and D. Scharstein, *Sampling the disparity space image*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26 (2004), pp. 419–425.

[49] P. Tan and P. Monasse, *Stereo disparity through cost aggregation with guided filter*, tech. report, IPOL Preprint, 2013.

[50] C. Tomasi and R. Manduchi, *Bilateral filtering for gray and color images*, in Proceedings of the Sixth International Conference on Computer Vision, vol. 846, Citeseer, 1998.

[51] O. Veksler, *Fast variable window for stereo correspondence using integral images*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1 (2003), pp. 556–561.

[52] Luc Vincent, *Grayscale area openings and closings, their efficient implementation and applications*, in First Workshop on Mathematical Morphology and its Applications to Signal Processing, 1993, pp. 22–27.

[53] L. Wang, M. Liao, M. Gong, R. Yang, and D. Nister, *High-quality real-time stereo using adaptive cost aggregation and dynamic programming*, in Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission, 2006, pp. 798–805.

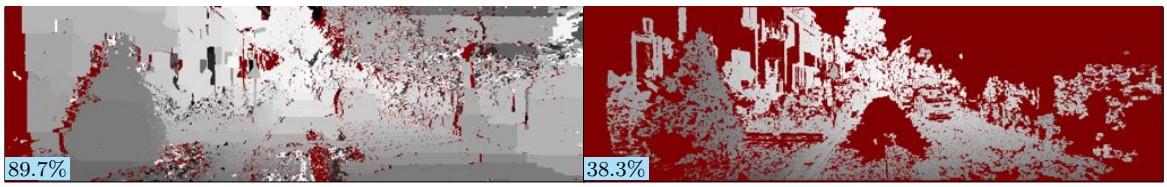
[54] Zeng-Fu Wang and Zhi-Gang Zheng, *A region based stereo matching algorithm using cooperative optimization*, in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, June 2008, pp. 1–8.

[55] L. Yaroslavsky and M. Eden, *Fundamentals of digital optics*, 2003.

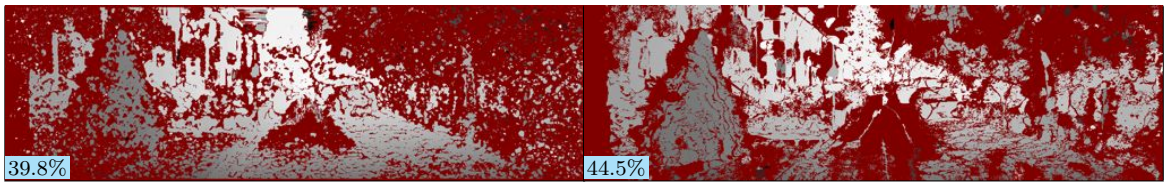
[56] K.J. Yoon and I.S. Kweon, *Adaptive support-weight approach for correspondence search*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (2006), pp. 650–656.

[57] ———, *Stereo matching with the distinctive similarity measure*, in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, Oct. 2007, pp. 1–7.

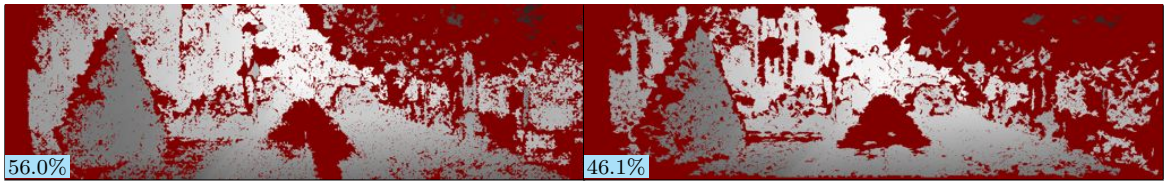
(a) stereo pair



(b) GC (c) GCS



(d) 5REG (e) CVF



(f) SGBM (g) MSMW

(h) MSMW w/fill

Figure 17. *Experiments on KITTI benchmark [21]. Top, from left to right: the stereo pair, the disparity computed by the graph cut method (GC), the GCS, the 5REG, CostVolume, SGBM (OpenCV's implementation of SGM), and the MSMW (without and with the filling-in post-process). The density of each disparity map is indicated in the lower left corner. The disparity range for this pair is [-120, 0].*