

AUTOMATIC LOW BASELINE STEREO IN URBAN AREAS

L.IGUAL[†], J.PRECIOZZI[‡], L.GARRIDO[†], A.ALMANSA^{‡*},
V.CASELLES[†] AND B.ROUGÉ^{*}

[†]Dept. de Tecnologia, Universitat Pompeu Fabra, 08003 Barcelona, Spain

[‡]Fac. de Ingeniería, Universidad de la República, 11300 Montevideo, Uruguay

*CMLA, Ecole Normale Supérieure de Cachan, 94235 Cachan cedex, France

*Centre National d'Etudes Spatiales (CNES), 31055 Toulouse, France

(Communicated by Jean-Michel Morel)

ABSTRACT. In this work we propose a new automatic methodology for computing accurate digital elevation models (DEMs) in urban environments from low baseline stereo pairs that shall be available in the future from a new kind of earth observation satellite. This setting makes both views of the scene similarly, thus avoiding occlusions and illumination changes, which are the main disadvantages of the commonly accepted large-baseline configuration. There still remain two crucial technological challenges: *(i)* precisely estimating DEMs with strong discontinuities and *(ii)* providing a statistically proven result, automatically. The first one is solved here by a piecewise affine representation that is well adapted to man-made landscapes, whereas the application of computational Gestalt theory introduces reliability and automation. In fact this theory allows us to reduce the number of parameters to be adjusted, and to control the number of false detections. This leads to the selection of a suitable segmentation into affine regions (whenever possible) by a novel and completely automatic perceptual grouping method. It also allows us to discriminate *e.g.* vegetation-dominated regions, where such an affine model does not apply and a more classical correlation technique should be preferred. In addition we propose here an extension of the classical "quantized" Gestalt theory to continuous measurements, thus combining its reliability with the precision of variational robust estimation and fine interpolation methods that are necessary in the low baseline case. Such an extension is very general and will be useful for many other applications as well.

1. INTRODUCTION

Computing the depth of objects in a scene from two or more images taken from different points of view is one of the key problems in computer vision known as stereo vision. Its numerous applications make it the object of current research, see [36] and [6] for an account of it.

2000 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases: stereo vision, subpixel urban DEMs, piecewise affine, a contrario region-merging approach.

A. Almansa, J. Preciozzi and L. Igual acknowledge partial support from PDT project number S/C/OP/17/01 (Uruguay). A. Almansa acknowledges additional support from CSIC (Uruguay), CNES, CNRS and ENS Cachan (France), and UPF (Spain). V. Caselles acknowledges partial support by the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya and by PNP GC project, reference BFM2003-02125. The authors would like to especially thank Jean-Michel Morel, Gabriele Facciolo, Julie Delon, Rafael Grompone, Jérémie Jakubowicz, for their help and fruitful discussions, as well as CNES for providing the data and part of the software used for this article.

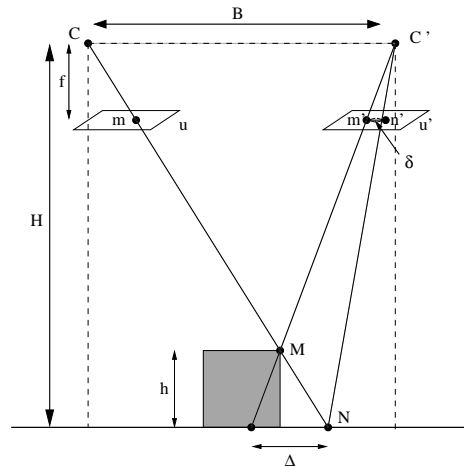


FIGURE 1. Sketch of the capturing process performed by two cameras C , and C' . The image planes are u and u' . Disparity between m and m' is equal to distance between m' and n' denoted by δ . The *baseline* or distance between viewpoints is B , and f is the focal length. By triangle similarity arguments we get the relationships $\Delta = \frac{H}{f}\delta$ and $\Delta = \frac{B}{H-h}h \simeq \frac{B}{H}h$ where the last approximation assumes that the cameras are very high ($H \gg h$). Now, if r is the size of a pixel on the image plane, the size of a pixel projected on the ground is $R = \frac{H}{f}r$, which is a useful measure of resolution of the system. Combining the previous relationships we obtain the main result shown in Equation (1).

The depth (or height) estimation from stereo pairs involves several steps. In this article we concentrate mainly on matching stereo pairs of satellite or aerial images that have been already rectified to epipolar geometry and where the altitude of the camera is high enough for the parallel projection model to be accurate. In such a setting the relationship between the height h (measured in meters) of a point in the 3D scene, and the disparity δ (measured in pixels) between the two projections x and $x + \delta(x)$ in the reference and secondary images can be described by [13, 23, 25]

$$(1) \quad \delta[\text{pixels}] \simeq \frac{B}{H} \frac{1}{R} h[\text{meters}].$$

where B is the *baseline* or distance between viewpoints, H is the height of the cameras with respect to the ground, and $R[\text{meters/pixel}]$ is the size of an image pixel projected on the ground (see figure 1 for an explanation of this geometric setting).

1.1. LOW BASELINE STEREO: ADVANTAGES AND CHALLENGES. This relationship means that the accuracy in the height measurements h is directly proportional to the accuracy in the disparity measurements δ and the resolution R , and inversely proportional to the B/H ratio. If we assume that the accuracy in the disparity measurements δ is limited by the pixel size, and that the resolution R is limited by hardware constraints, then the only way to improve the accuracy in height measurements h is to use a system with a relatively large B/H ratio of about 1, which corresponds to cameras at a viewing angle of 45° . For this reason the great majority of the literature on stereo vision has concentrated in the high B/H case

where sub-pixel estimations of the disparity are not necessary to obtain a reasonable depth estimation.

Using large B/H values, however, presents several disadvantages, especially when computing digital elevation models in urban areas, which is our focus in this article. First, streets and low buildings will be occluded by higher buildings in at least one of the images, making it very difficult to estimate the height in those occluded areas. And secondly, but most important, when obtaining both images of the stereo pair with a single satellite, a large B/H imposes a significant time-frame between the two shots, during which illumination conditions changed, shadows moved, and several other changes may have occurred. All these geometric and radiometric changes make the search for reliable matching points much more difficult and error-prone.

New earth observation devices could be capable of automatically computing DEMs in urban environments from very low $B/H \approx 0.05$ stereo pairs taken from the zenith. Taking into account the speed of the device, the time-frame between shots could be just a few seconds, thus obtaining quasi-simultaneous views, and avoiding many of the disadvantages of the large B/H setting.

On the other hand, in order to make height measurements accurate enough, several challenges have to be addressed. First, in addition to using high resolution acquisition systems, it is also necessary to be able to interpolate both images very accurately so that a precision of about 0.1 pixels in the computation of the disparity becomes feasible. See [44, 4, 2] for an account of restoration, and microvibration estimation and correction techniques that can be used to make this possible, by assuring that the bandlimiting conditions of Shannon's sampling theorem are satisfied, while at the same time improving resolution as much as possible. Second, the *adherence* phenomenon may have to be modeled and corrected, as explained in [13].

The advantage of using stereo pairs with low B/H values can be formulated more precisely in the following manner. In the ideal case where $B/H \rightarrow 0$, no occlusions occur. Given an image pair u and \tilde{u} under these conditions the relation between them can be modeled as a simple image deformation model

$$(2) \quad \tilde{u}(\mathbf{x}) = g(u(\mathbf{x} + \delta(\mathbf{x}))) + n,$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a (non-decreasing) contrast change, $\delta(\mathbf{x}) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is the disparity induced from one image to the other by the urban surface (*i.e.* proportional to the elevation map $h(\mathbf{x})$ as seen before), and the noise n is usually modeled as a zero-mean Gaussian.

This model is more and more accurate as $B/H \rightarrow 0$. Even if the images are not taken from the zenith, the small B/H ensures that no significant change of the occlusion occurs between both shots, so that they are linked by a simple "deformation" model like equation (2).

In addition, the fact that both images from the pair are almost simultaneous implies that virtually no illumination changes occurred, so that the contrast change g can be well approximated by a linear one. For this reason we considered in this article both an algorithm that is robust to general contrast changes (see RAME, Section 2.2) and a second algorithm which is optimized to the case of locally linear contrast changes (see MARC, Section 2.1).

1.2. RELATED WORK ON ACCURATE STEREO. Our choice of the MARC and RAME methods is based on the fact that the majority of other stereo methods are not devoted to subpixel disparity computation. The applied techniques often use local

correspondences because the information needed for the matching cannot be accurately and reliably extracted from a single pixel graylevel. Block matching methods with large windows, or feature locations determined by interpolating the images are therefore used to produce subpixel maps [12, 46], thus being potentially corrupted by the adherence artifact described and analyzed in [14]. These effects are most often minimized by the use of very small windows, which leads to inaccurate and unreliable results. Even though such inaccuracies can be corrected by the use of some kind of regularization, *i.e.* either local or global energy minimization, this would require the use of very specialized techniques to obtain a computationally well performing algorithm. The optical flow methods (locally regularized) are fast and obtain accurate results [27, 35, 7, 11], but initial experiments with these methods did not show enough accuracy when applied to low B/H stereo pairs of urban scenes [43, 20]. A more exhaustive evaluation may however be necessary to confirm this conclusion. The methods that use discretized space/disparity grids like dynamic programming [30] or graph cuts [45, 32, 46] are reported to produce the most accurate results with the fastest computational performance in the large B/H case. But in the case of low-baseline stereo, where precisions of the order of 0.1 pixel or less are required, applying such algorithms would require refining this grid by a very large factor (either in the spatial or disparity domain or both), thus becoming computationally too expensive to be applied with subpixel accuracy. Such shortcomings may be solved by recent advances in graph-cut algorithms for energy minimization, which avoid the fixed discretization of the disparity values, in favour of a dyadic search strategy [10].

Nevertheless the application of graph-cut based methods is more fundamentally limited to minimizing a certain class of regularization energies [33]. To the best of our knowledge, all graph-cut compatible energies that have been explored and experimented so far, tend to favour piecewise-constant solutions, thus producing serious staircasing artifacts whenever the depth map does not agree with this model, as it has been shown for instance in [34]. The piecewise-affine model is much better adapted to the kind of urban environments we are interested in (where slanted roofs are ubiquitous), and is the one that will be adopted in this work, for the reasons explained in section 2.

The Multiresolution Algorithm for Refined Correlation (MARC) introduced in [13, 24] is a new correlation-based method which performs well in subpixel case, however the result is a non dense disparity map. We propose to interpolate the missing data using a criterion coherent with the underlying urban model. We analyze here a region-based approach which assumes that the transformation of the regions corresponding to structures of the reference image can be modeled by affine transformations.

We also study the performance of the Region-based Affine Motion Estimation approach (RAME) presented in [29] applied to the estimation of disparity maps in stereo image pairs. This approach is based on aligning gradient orientations between both images, thus it is completely independent of MARC's disparity computations, but is region-based and also assumes an affine motion model for each region.

1.3. MODELING URBAN ENVIRONMENTS. The aim of the work presented in this paper is threefold: *(i) Denoising* the disparity results of MARC and RAME to obtain a better *accuracy* as discussed in the previous paragraph is only one aspect of the problem. In this work we address in addition, and simultaneously the following two problems: *(ii) Validating* the disparity measurements, to be sure that they are

reliable; and *(iii)* providing a slightly *higher level description* of the urban scene than just a disparity map.

In this last sense our aim is similar to the work of Descombes et. al. [41, 42] where highly complex building models are fitted to a previously measured elevation/disparity map. In our low B/H case, however, we have to deal with much sparser and noisier disparity measurements. In such a situation, the manual detection threshold tuning, and difficult to achieve convergence conditions reported in [41] are not acceptable.

For these reasons, and following the computational Gestalt theory introduced in [19], we adopt here an *a contrario* approach to select the best urban scene model among a previously defined family of possibilities, and to test whether such a model is a valid explanation of the data or not. In addition to ensuring reliable results (in the sense that the number of false detections is controlled in a certain sense) this strategy allows to automatically fix detection thresholds, thus avoiding manual parameter tuning to a large extent. The challenge here is to obtain simultaneously accurate and reliable results whenever the scene correctly matches the model, and a mask of non-validated areas (such as vegetation or curved surfaces). The *a contrario* approach is an adequate tool for validating and selecting among different models, which is also quite robust to outliers, but the significance measure (or NFA) that has been used so far in this context is based on a quantization of the search space which limits its accuracy. On the other hand, robust estimation techniques are more specialized in producing highly accurate results even in the presence of outliers. In this work we propose to combine both approaches by using a "continuous" version of the significance measure or NFA (which is inspired from robust statistics) both for selecting the best model (and denoising the data), and for validating this choice.

In order to study in more detail the relationships between the three aspects *(i)*, *(ii)* and *(iii)* mentioned before, we adopt here a much simpler urban scene model than in [41], namely a piecewise-affine one, which is still adequate for urban scenes, and can be later further grouped to form more complex structures in the hierarchical fashion suggested by Gestalt psychophysicists.

1.4. OVERVIEW. Summarizing, after briefly introducing in Section 2 the two algorithms (MARC and RAME) that we use to obtain raw disparity estimations, and how we use them to obtain a piecewise affine description of the DEM, we introduce in Section 3 an *a contrario* framework to validate the meaningfulness of each facet of this piecewise affine model. The core of the proposed approach is given in Section 4 which extends the previous *a contrario* framework in such a way that it also allows us to select the best segmentation into affine regions among a sufficiently large number of possibilities. The main idea consists in a decision rule to determine whether two neighboring regions should be kept separate (with a different affine model for each) or joined into a single region (with a common affine model for both). The details on how the statistical background model is defined and used to compute or estimate probabilities are deferred to Section 5. In particular this section introduces a new "continuous NFA" which allows to treat the estimation and validation stages in a uniform and coherent way. Finally the experimental results in Section 6 show how the proposed segmentation, estimation and validation algorithm is useful in the two different contexts (MARC and RAME).

2. ESTIMATING PIECEWISE AFFINE MODELS FOR URBAN DEM IMAGES

A common characteristic of urban scenes is that they are composed by geometric structures corresponding to buildings, streets, and so on. On the other hand, recall that rigid motions of planar objects in 3-D space induce homographic motion models in 2-D images [47]. This homographic motion model is a good approximation when the depth of the objects is small compared to their distance to the camera. Moreover, the affine motion model is a good approximation under parallel projection. In this context, it seems reasonable to model the disparity map by a piecewise affine transformation.

Thus, we consider a partition \mathcal{R} of the reference image domain into connected and disjoint regions. This partition is found here by the piecewise constant Mumford-Shah minimization in [31] and by the approach proposed in [5] based on the *Mumford-Shah functional subordinated to the level lines of the image*. Mumford-Shah based segmentations are motivated by the Lambertian hypothesis, which states that a planar surface patch with uniform reflectance properties produces uniform luminance on the corresponding region of the image. Here we use the inverse Lambertian hypothesis which associates a planar surface patch to each uniform luminance region in the image. This inverse hypothesis most often leads to an over-segmentation of the image, since most planar surface patches are composed of several sub-patches with different luminance levels. This over-segmentation is dealt with by the subsequent region-merging algorithm.

On the other hand the inverse hypothesis does not always hold true. For instance a house roof composed of two planes with the same reflectance and at the same angle with respect to the observer and the illumination source will appear as a uniform region with almost constant gray-level in the image. Even though such cases are quite rare they may justify the use of polygonal segmentations where not all the sides of the polygon need to be well contrasted. This possibility is explored in [48, 43].

Let us denote by $T(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x}))$, $\mathbf{x} = (x, y)^T$, a general transformation which may be written as $T(\mathbf{x}) = \mathbf{x} + \delta(\mathbf{x})$, being $\delta(\mathbf{x})$ the disparity map. A six-parameter affine transformation has the following form [47]:

$$(3) \quad T^A(\mathbf{x}) = \begin{pmatrix} T_1^A(\mathbf{x}) \\ T_2^A(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix},$$

where e, f are the translation parameters and a, b, c, d are the parameters that model the linear transformation (thus, including scaling in both directions, rotation and shearing) [47].

From now on, we shall assume that images have been aligned and horizontal lines are in epipolar correspondence. In this case T_2 is the identity transformation and T^A has the form

$$(4) \quad T^A(\mathbf{x}) = \begin{pmatrix} T_1^A(x, y) \\ T_2^A(x, y) \end{pmatrix} = \begin{pmatrix} ax + by + e \\ y \end{pmatrix}.$$

Let us denote by \mathcal{A} this class of transformations which have three parameters: the translation parameter e , and the linear transformation parameters a and b .

As mentioned before we consider in this work two different ways of estimating the affine transformation T^A in each region from the image data. The first one (Section 2.1) is based on normalized correlation maximization, and the second one

(Section 2.2) is based on matching the level lines of both images in the stereo pair and is therefore contrast invariant.

2.1. RAF-MARC: ROBUST AFFINE FITTING TO POINT CLOUDS. MARC is an algorithm that implements a multiwindow multiscale correlation. Invented at CNES by B. Rougé [24] it was mathematically analyzed by B. Rougé and J. Delon [15, 13] and coded by Nathalie Camlong [8] and Vincent Muron [39]. As most of the matching algorithms proposed in the literature, it generates a non dense disparity map, but for most applications a dense disparity map is needed. To compute it we use the piecewise affine model for disparity to interpolate the missing data.

Given the segmentation \mathcal{R} , the class of affine maps \mathcal{A} , the computed disparity map $M(\mathbf{x})$ and a continuous function $\rho : [0, \infty) \rightarrow [0, \infty)$ with a unique minimum at zero (being locally convex there), for each region $R \in \mathcal{R}$ we compute the affine transformation $T_R \in \mathcal{A}$ that best fits the data in R by minimizing the error functional :

$$(5) \quad E_R(T) := \sum_{\mathbf{x} \in R^*} \rho(Y(x)) = \sum_{\mathbf{x} \in R^*} \rho(\|T(\mathbf{x}) - (\mathbf{x} + M(\mathbf{x}))\|),$$

where $\|\mathbf{v}\|$ denotes the euclidean norm of the vector in $\mathbf{v} \in \mathbb{R}^2$. Notice that the sum is only extended to $R^* \subseteq R$ which is the set of valid points where the disparity can be accurately computed (with accuracy λ) by correlation using an algorithm like MARC.

There are several estimation techniques to minimize Equation (5), some of them are less sensitive to the outliers present in the original data than others. We have chosen the M-Estimators technique [28], with a normalized Tukey function such that $\rho(x) \in [0, 1]$ for any $x \in \mathbb{R}$. It will be useful to be able to apply Hoeffding's theorem in Section 5, without affecting the result of the minimization.

Finally let us observe that even if our experiments were done with disparity data obtained from MARC, this methodology can be applied to the output of any algorithm that produces a point cloud like $\{(\mathbf{x}, (\mathbf{x} + M(\mathbf{x}))) : \mathbf{x} \in R^*\}$ that has to be matched by an affine map $(\mathbf{x}, T(\mathbf{x}))$ within the region R .

2.2. RAME : REGION-BASED AFFINE MOTION ESTIMATION. In this Section we review the motion estimation algorithm developed by Caselles, Garrido and Igual in [9, 29] which was initially conceived for motion estimation in image sequences and will be applied here to disparity computation in satellite stereo image pairs.

Usually, the energy functional whose minimum gives the disparity map is based in the brightness constancy assumption [27, 35, 7]. In [9, 29], the authors proposed a contrast invariant displacement estimation based on the following assumption: shapes move with possible affine deformation between two frames in a sequence (or two images in a stereo pair). Thus gradient orientations (which are orthogonal to the shape boundaries) should match between both images after an appropriate affine transformation.

More precisely, given two images u and \tilde{u} and a region $R \in \mathcal{R}$ from a segmentation of u , the optimal transformation $T \in \mathcal{A}$ is estimated by minimizing the error functional

$$(6) \quad E_R(T) := \sum_{\mathbf{x} \in R^*} \rho(Y(x)) = \sum_{\mathbf{x} \in R^*} \rho(\|Z(u \circ T)(\mathbf{x}) - Z(\tilde{u})(\mathbf{x})\|),$$

i.e. the mis-alignment between the gradient orientations

$$(7) \quad Z(\tilde{u}) = \frac{\nabla \tilde{u}}{\|\nabla \tilde{u}\|}$$

in the secondary image, and the gradient orientations $Z(u \circ T)$ in the transformed reference image.¹ Note that these gradient orientations are accurately defined only at those points \mathbf{x} where both gradients are large enough. Thus the sum is reduced to the points $\mathbf{x} \in R^* \subseteq R$ such that $\|\nabla \tilde{u}(\mathbf{x})\|$ is above a certain threshold (0.05 in our experiments).

For more details about the optimization and interpolation techniques used to find this minimum with high levels of accuracy we refer the reader to [9, 29]. Here we used a quadratic function $\rho(e) = \frac{1}{4}e^2$ for simplicity, but other robust functions may be used as well.

3. VALIDATION OF PIECEWISE AFFINE ELEVATION MODELS

We are assuming in this section that we have computed a piecewise affine disparity map, that is, the disparity map can be expressed on each region $R \in \mathcal{R}$ by a map $T_R \in \mathcal{A}$. The maps have been obtained using either RAF-MARC or RAME. In both cases, the maps are obtained by minimizing an energy functional which measures a matching error and we cannot ensure that the minima obtained give us the correct model of the region (we could match a donkey to a pig!). Moreover, the affine model is appropriate only for objects that can be approximated by planes. Even if this is common in urban images there may be exceptions (trees for instance, can not be modeled by affine transformations). For this reason, we need to measure how well the estimated transformation adjusts the disparity values of a region. Our approach is based on the statistical *a contrario* approach developed in [17, 18].

According to this theory a geometric event is considered as meaningful if its expected number of occurrences "by chance" is very low (typically below $\epsilon = 1$). In our case the geometric event will be given by an affine transformation T which produces an unexpectedly low matching error $E_R(T; \theta)$ in region R . Here, $E_R(T; \theta)$ denotes the error functional minimized by RAF-MARC (equation 5) or RAME (equation 6) where we made explicit the dependence of the matching error on the image observations θ (disparities $\theta = M$ in the RAF-MARC case, or gradient orientations $\theta = Z$ in the RAME case).

Alternatively, the geometric event can be written in terms of the *degree of coincidence*

$$(8) \quad k(R, T, \theta) := |R^*| - E_R(T; \theta) = \sum_{x \in R^*} (1 - \rho(Y(x))).$$

This name is justified by the fact that when choosing $\rho(y) = \mathbb{1}_{|y| \geq \alpha}$, then $k(R, T, \theta)$ actually counts the number of points $x \in R^*$ where the matching error $Y(x)$ is below the precision threshold α . This is the common *quantized* case used in most a contrario methods. However, when $\rho : \mathbb{R} \rightarrow [0, 1]$ is a more general and smooth robust function (like the quadratic or Tuckey function), then $k(R, T, \theta)$ measures a *continuous* degree of coincidence.

In order to make clear what we mean by "chance" we consider the image observations θ as a realization of a random process Θ that follows a certain statistical distribution or "background model" (to be specified more precisely in Section 5). A

¹ In the sequel \tilde{Z} will be used as a shorthand for $Z(\tilde{u})$, and Z as a shorthand for $Z(u \circ T)$.

common assumption is to consider $\Theta(x)$ as i.i.d. uniform random variables. Now, if the observed number of coincidences $k(R, T, \theta)$ is remarkably large with respect to this background model, then we consider T to be a valid choice for region R :

Definition 1 (Number of False Alarms (NFA), and ϵ -meaningful event). The *number of false alarms* $NFA(R, T)$ of assigning the affine map T to the region R is the expected number of occurrences of an event E' at least as rare as $E = [k(R, T, \Theta) \geq k(R, T, \theta)]$, where E' spans all possible choices of a region R' and an affine transformation T' in the image pair. The choice (R, T) is called ϵ -*meaningful* if its number of false alarms is lower than ϵ

$$(9) \quad NFA(R, T) < \epsilon.$$

This ideal definition is still far from a practical procedure, since the exact value of the NFA may be difficult to compute. However, in most situations, this ideal definition can be bounded from above by an expression of the form

$$(10) \quad NFA \leq \mathcal{N} \cdot \mathbb{P}[E],$$

where \mathcal{N} (or number of tests) is the number of possible configurations of the event E (see [19], or Section 4.1 for a proof in our case). For this reason, defining an event as ϵ -meaningful, whenever $\mathcal{N} \cdot \mathbb{P}[E] < \epsilon$, is still consistent with the original definition 1 and ensures that the method is robust in the sense that no more than ϵ “false detections” will be obtained due to noise.

Our practical validation procedure will therefore use the following definition instead of the previous one :

Definition 2. The Number of False Alarms of (R, T) is defined as:

$$(11) \quad NFA(R, T) := N_{tests} \mathbb{P} [k(R, T, \Theta) \geq k(R, T, \theta)],$$

where N_{tests} is the number of all possible configurations we can have for the pair (R, T) The choice (R, T) is called ϵ -*meaningful* if its number of false alarms is lower than ϵ

$$(12) \quad NFA(R, T) < \epsilon.$$

Remark 1. The NFA is a measure of the significance of an observation and permits us to compare two transformation maps for two given regions. Given two observations (R_1, T_1) and (R_2, T_2) , we say that (R_1, T_1) is more significant than (R_2, T_2) if $NFA(R_1, T_1) < NFA(R_2, T_2)$. In other words, (R_1, T_1) gives less support to the a contrario hypothesis than (R_2, T_2) . Notice that if $R = R_1 = R_2$ the previous condition allows us to compare two different models for the same region. This will be useful in the next Section 4.

Remark 2. The actual computation of the NFA will require a precise definition of the number of tests, the background model for each case, as well as a procedure to compute the probabilities $\mathbb{P} [k(R, T, \Theta) \geq k(R, T, \theta)]$. The definition of N_{tests} will be deferred to the next Section 4.1, and the definition of background models and probability computations will be deferred Section 5.

Remark 3. A case of particular interest that will be extensively used in the next section is the transformation T_i that maximizes the degree of coincidence for a given region R_i of the partition \mathcal{R} :

$$(13) \quad T_i := \arg \max_{T \in \mathcal{A}} k(R_i, T, \theta) = \arg \min_{T \in \mathcal{A}} E_{R_i}(T, \theta)$$

as well as the maximal value of the degree of coincidence for all possible transformations T

$$(14) \quad k_i := k(R_i, T_i, \theta) = \max_{T \in \mathcal{A}} k(R_i, T, \theta)$$

and the corresponding random variable

$$(15) \quad \mathcal{K}_i := k(R_i, T_i, \Theta)$$

Observe that this maximization of $k(R_i, T, \theta)$ *w.r.t.* $T \in \mathcal{A}$ is equivalent to the minimization of $NFA(R_i, T)$, since the probability $\mathbb{P}[k(R, T, \Theta) \geq k]$ is a non-increasing function of k , and the choice of T does not affect the number of tests which is constant for a given region. Therefore, $NFA(R_i, T)$ will be computed only to *decide* whether (R_i, T_i) is meaningful and whether it should be merged with another neighbouring pair (R_j, T_j) . However, the *optimization* needed to find T_i is carried out by k instead of NFA which is not as well posed numerically. Nevertheless in both cases we are dealing with *the same* optimization.

4. AUTOMATIC SEGMENTATION INTO VALIDATED PIECEWISE AFFINE REGIONS

Up to now we have dealt with disparity estimation between stereo images by dividing the first one into disjoint regions and estimating the disparity of each one of them in an independent way. Consequently, the global disparity map is described as a set of independent disparity maps, one for each region of the initial segmentation. In order to become independent of this segmentation we consider the integration of the estimated disparity maps using an iterative region merging approach. This will give coherence to disparity maps which are coincident and will eliminate outliers.

As in Section 2 we consider an initial segmentation \mathcal{R} of the reference image in a set of connected and disjoint regions that we denote R_1, \dots, R_N . For each of them we have an affine transformation in \mathcal{A} obtained using either RAF-MARC or RAME. In what follows, we assume that the estimation method is fixed. Let $(R_i, T_i), \forall i \in \{1, \dots, N\}$ be the pairs of region and associated transformation. Then, using our approach “coherent regions” are iteratively merged together. By “coherent regions” we mean those neighboring regions with a very similar affine transformation which probably means that the regions are part of the same scene structure. This region merging process is based on the significance measure defined in Section 3.

As usual, in order to define the region merging algorithm we need to set three concepts: *the region model*, *the merging criterion* and *the merging order*.

The region model: We call the region model each pair (R_i, T_i) where $R_i \in \mathcal{R}$ and $T_i \in \mathcal{A}$.

The merging criterion: Given two adjacent regions R_i and R_j , and their associated transformations T_i and T_j , we must define a criterion to decide whether we merge them or not. We consider the union of both regions $R_i \cup R_j$ and we estimate its associated transformation $T_{ij} \in \mathcal{A}$ (the one minimizing $E_{R_i \cup R_j}(T)$). Then, we compute the NFA of the joint region $R_i \cup R_j$ with its new estimated transformation T_{ij} , and compare it with the NFA of having each region separately with their independent transformations. Then the merging criterion is defined by the condition:

$$(16) \quad NFA(R_i \cup R_j, T_{ij}) \leq NFA((R_i, T_i); (R_j, T_j)).$$

If this condition is verified, then R_i and R_j are merged. Let us explain the meaning of inequality (16). We are asking if the joint region with its new estimated transformation T_{ij} is more meaningful than the two regions considered separately, each one with its own transformation.

We consider the following random variables, $\mathcal{K}_{ij} = k(R_i \cup R_j, T_{ij}; \Theta)$, and $\mathcal{K}_{\text{sum}} = \mathcal{K}_i + \mathcal{K}_j = k(R_i, T_i; \Theta) + k(R_j, T_j; \Theta)$. The first one measures the degree of coincidence of the configuration $(R_i \cup R_j, T_{ij})$. The second one measures the joint degree of coincidence of (R_i, T_i) and (R_j, T_j) .

We denote the observation of the first random variable \mathcal{K}_{ij} by k_{ij} , and the observations of \mathcal{K}_i and \mathcal{K}_j by k_i and k_j respectively.

The number of false alarms (NFA) of having two regions R_i and R_j , each one with a different associated transformation, T_i and T_j , is defined as:

$$(17) \quad \text{NFA}((R_i, T_i); (R_j, T_j)) = N'_{\text{tests}} P((R_i, T_i); (R_j, T_j))$$

where N'_{tests} is the number of all possible configurations for the pairs $((R_i, T_i), (R_j, T_j))$, and the probability

$$P((R_i, T_i); (R_j, T_j)) := \mathbb{P}[\mathcal{K}_i + \mathcal{K}_j \geq k_i + k_j]$$

is the probability of having two regions, each one with a different transformation.

Thus, the merging criterion is:

$$(18) \quad N_{\text{Tests}} P(R_i \cup R_j, T_{ij}) \leq N'_{\text{Tests}} P((R_i, T_i); (R_j, T_j))$$

where N_{Tests} is the number of tests of having a single disparity model for $R_i \cup R_j$ and N'_{Tests} is the number of tests of having two models, one for each region.

The description of both numbers N_{tests} and N'_{tests} will be given in Section 4.1. Both numbers are different. Indeed, $N_{\text{tests}} \leq N'_{\text{tests}}$, but $P(R_i \cup R_j, T_{ij}) \geq P((R_i, T_i); (R_j, T_j))$ and (18) is a kind of compromise between having two models (one for each region) with a better fit of the data, or a more regular one (one for both regions) with less good fitting properties but still good enough to compensate for the simplicity of the model. Thus the merging criterion is similar, in a certain sense, to a variational approach where $\log(N_{\text{tests}})$ plays the role of the regularization term and $\log(P)$ plays the role of the data fitting term.

It is clear that the definition of the joint probability of having two transformations, one for each region, is crucial for the merging criterion. The problem of estimating this probability was first addressed for clustering problems in [40] in the case where both regions R_i and R_j may share some points, leading to a trinomial distribution because of the non-independence of the events (instead of the usual binomial distribution that we obtained in the quantized case). Under certain hypotheses, this trinomial distribution can be approximated with a term that is easier to compute.

This is not the case of the region merging where the neighboring regions are disjoint. Thus, we have to look for another solution. In [49] the merging problem is being studied in the context of multisegment detection, and an “ideal” joint probability of the event (R_i, R_j) is defined in terms of a sum of probabilities among all pairs of events which are more meaningful than the observed one. In one dimension this reduces to a simple threshold problem, but in two dimensions the geometry of this area is unknown and numerical computation of the “ideal” joint probability can be too expensive. For this reason the authors considered a lower bound as an approximation of this ideal expression.

In our context, we found that our criterion gives better results than either the lower bound or the approximation defined in [49], and since we have a simpler explanation in terms of comparing a simple loosely fit model with a more complex tightly fit model, we kept the new merging criterion as defined in Equation (18).

The merging order: Let us define the order in which we process the regions to be merged. We can only merge adjacent regions, since the goal of the merging process is to merge regions that belong to the same physical object. Thus, we consider all possible pairs of adjacent regions, for each one of these pairs we estimate the transformation associated to the joint region, defined by the union of both regions in the pair, and then we compute the NFA of this new region. Each of these pairs is inserted in a priority queue ordered by the NFA. The merging order is defined using this NFA: we process first the pair of regions with the lowest NFA, which is the first pair in the priority queue.

Now, we can summarize our algorithm.

Algorithm 1. :

Input: Initial segmentation \mathcal{R}

Output: Final segmentation \mathcal{R}' and associated affine transformations T and $\text{NFA}(R, T)$ for all $R \in \mathcal{R}'$.

1. Initialization process:
 - (a) Build a graph $\mathcal{G} = \langle \mathcal{R}, A \rangle$ where a node $R_i \in \mathcal{R}$ represents a region of the image partition \mathcal{R} , and an edge $a_{i,j} \in A$ exists if regions R_i and R_j are adjacent.
 - (b) For each region $R_i \in \mathcal{R}$ ($i = 1, \dots, N = |\mathcal{R}|$) estimate the optimal transformation T_i which minimizes $\text{NFA}(R_i, T)$ among $T \in \mathcal{A}$ (equivalently, T_i maximizes $k(R_i, T; \theta)$, see Remark 3).
 - (c) For each joint region $R_i \cup R_j$, where R_i and R_j are adjacent regions, estimate the optimal transformation T_{ij} and which minimizes $\text{NFA}(R_i \cup R_j, T)$ for $T \in \mathcal{A}$.
 - (d) Build a priority queue of joint regions $R_{ij} = R_i \cup R_j$ ordered by $\text{NFA}(R_{ij}, T_{ij})$.
 - (e) Initialize the final segmentation $\mathcal{R}' := \mathcal{R}$, the corresponding adjacency graph $\mathcal{G}' := \langle \mathcal{R}', A' \rangle = \mathcal{G}$ and $l := N = |\mathcal{R}'|$.
2. Iterative process (Iterate until the queue is empty):
 - (a) Take the first element from the queue, call it (R_{l+1}, T_{l+1}) , and remove it from the queue. This element is the pair $(R_i \cup R_j, T_{ij})$ with the lowest NFA in the queue.
 - (b) If the pair satisfies the merging criterion defined in Equation (16) then:
 - (i) Remove R_i and R_j from \mathcal{R}' , and insert R_{l+1} instead in the partition. Update the edges in the graph $\mathcal{G}' = \langle \mathcal{R}', A' \rangle$ to represent the adjacency relationship of the new set of nodes.
 - (ii) Remove all of the entries in the queue involving either R_i or R_j .
 - (iii) Estimate the optimal transformation $T_{l+1,m}$ minimizing $\text{NFA}(R_{l+1} \cup R_m, T)$ among $T \in \mathcal{A}$ for all neighbors R_m of R_{l+1} in \mathcal{G}' .
 - (iv) Insert each of these new pairs $(R_{l+1} \cup R_m, T_{l+1,m})$ with its NFA into the queue.
 - (v) Go to step 2a incrementing l by 1.
 - (c) Otherwise (the merging criterion is not satisfied) discard (R_{l+1}, T_{l+1}) leaving \mathcal{R}' unchanged. Go to step 2a without incrementing l .

Note that several steps of this algorithm have different implementations depending on the selected approach (MARC or RAME). For instance, the transformation estimation and NFA computation in steps 1b, 1c, 2(b)iii are different for MARC and RAME.

The merging procedure presented so far has the same structure as the *iterative exclusion principle* [19], with which it shares the idea that the most significant gestalt masks all other similar gestalts involving the same atoms, and should be processed first. Particular implementations of this iterative exclusion principle to detect alignments [3, 19] and vanishing points [1] have been developed with success. A similar region merging algorithm but with a simplified merging criterion was proposed in [29] in a motion estimation context.

Now we can calculate the number of tests needed in Definition 2, equation (11), to compute the NFA, in such a way that it takes into account the whole region merging process in the previous algorithm.

4.1. THE NUMBER OF TESTS. Let \mathcal{S} be the set of pairs (R, T) for which $\text{NFA}(R, T)$ is computed by Algorithm 1 in order to test if it is maximal meaningful. We want to compute the size $|\mathcal{S}|$ of this set or at least find an upper bound which can be used as N_{tests} in equation (11).

Let's decompose this set into several parts $\mathcal{S} = \bigcup_{i=0}^{l_0} \mathcal{S}_i$, where l_0 is the final value of l when the algorithm stops, and each part is defined as follows:

$$\begin{aligned}
 (19) \quad & l = 0 \quad \mathcal{S}_0 := \{(R_i, T) : R_i \in \mathcal{R}, T \in \mathcal{A}\} \\
 & l \in [1, N] \quad \mathcal{S}_l := \{(R_l \cup R_j, T) : R_l \text{ and } R_j \text{ are adjacent in } \mathcal{G}, T \in \mathcal{A}\} \\
 & l \in [N + 1, l_0] \\
 & \mathcal{S}_l := \{(R_l \cup R_m, T) : R_l \text{ and } R_m \text{ are adjacent in } \mathcal{G}' \text{ at iteration } l, T \in \mathcal{A}\}
 \end{aligned}$$

i.e. \mathcal{S}_0 contains all pairs (R, T) tested at step 1b of the algorithm, $\bigcup_{i=1}^N \mathcal{S}_i$ contains all pairs tested at step 1c, while for $l > N$, \mathcal{S}_l contains all pairs tested at step 2(b)iii during the l 'th iteration of the loop.

We can compute an upper bound of the number of tests as $|\mathcal{S}| \leq \sum_{i=0}^{l_0} |\mathcal{S}_i|$.² From the definition of these sets it is clear that $|\mathcal{S}_0| = NN_{transf}$, and for $l > 0$, $|\mathcal{S}_l| = C(R_l)N_{transf}$ where $C(R_l)$ is the number of neighbours of region R_l just after being inserted into \mathcal{R}' , and N_{transf} is the number of transformations in a reasonable discrete sampling of \mathcal{A} (see Section 4.3 for details on how we compute it).

The final key observation needed to get an upper bound for the number of tests $|\mathcal{S}|$ is that $l_0 \leq 2N$. In fact, l starts at $l = N$ in step 1e while $|\mathcal{R}'| = N$. Each time l is incremented by one in step 2(b)v, $|\mathcal{R}'|$ is decremented by one in step 2(b)i. Since the algorithm will eventually stop (the queue will be empty) before $\mathcal{R}' = \emptyset$, we conclude that $l \leq l_0 \leq 2N$.

Finally, from the observations in the last two paragraphs we obtain

$$|\mathcal{S}| \leq NN_{transf} + \sum_{l=1}^{2N} C(R_l)N_{transf} \leq N(1 + 2\bar{C})N_{transf}$$

where \bar{C} is the mean of the graph connectivity $C(R_l)$ or an upper bound of it. Since \bar{C} may be difficult to estimate a priori from the initial segmentation \mathcal{R} , we define

² This upper bound is exact if the union is disjoint. Actually $\mathcal{S}_1, \dots, \mathcal{S}_N$ are not disjoint (each element is counted twice), but we keep this upper bound for simplicity.

the number of tests for each region R as follows:

$$(20) \quad N_{tests}(R) := N(1 + 3C(R))N_{transf}$$

The additional factor 3 instead of 2 is necessary when the number of tests is not constant (depends on the region). The reason for this will become clear in the proof of the next proposition.

The following proposition ensures that when inserting this definition of the number of tests to determine the NFA in Definition 2, the resulting ϵ -meaningful events (defined as $NFA < \epsilon$) are consistent with definition 1.

Proposition 1. *Let \mathcal{S} be the set of pairs (R, T) tested by Algorithm 1, and let the random variable $S = \sum_{(R, T) \in \mathcal{S}} \mathbb{1}_{(R, T)}$ is ϵ -meaningful count the total number of occurrences of an ϵ -meaningful event. Then the expected number of false alarms (under the corresponding background model) of Algorithm 1 is $E[S] \leq \epsilon$.*

The proof of this proposition will be given in Appendix 8.

4.2. REFORMULATION OF THE MERGING-CONDITION. The number of tests N'_{tests} for the joint NFA $((R_i, T_i); (R_j, T_j))$ of two regions is, as observed before, larger than the one for a single region (N_{tests}). The reason is simply that in this case we have to test separately all possible transformations T'_i and T'_j for each separate region instead of just testing a single set of transformations T'_{ij} for the merged region $R_i \cup R_j$. Thus, there is an extra factor N_{transf} in the case of the joint NFA:

$$N'_{tests}(R_i, R_j) := \frac{N_{tests}(R_i) + N_{tests}(R_j)}{2} N_{transf} = N(1 + 3\frac{C(R_i) + C(R_j)}{2}) N_{transf}^2$$

With this value, we can rewrite the merging criterion in equation (16) in a simpler form as

$$(21) \quad P(R_i \cup R_j, T_{ij}) \leq \frac{1 + 3\frac{C(R_i) + C(R_j)}{2}}{1 + 3C(R_i \cup R_j)} N_{transf} P((R_i, T_i); (R_j, T_j)).$$

and if we neglect the possible change in connectivity we can substitute it by

$$(22) \quad P(R_i \cup R_j, T_{ij}) \leq N_{transf} P((R_i, T_i); (R_j, T_j)).$$

The term N_{transf} can be interpreted as the cost of having a more complex model. Note that without this term, the region merging criterion is never satisfied. In fact, let R_i, R_j be two regions, both with n_i and n_j valid points. Let k_i and k_j be the maximal degrees of observed coincidences at each region that is attained for the transformations T_i, T_j (this maximization is equivalent to minimizing the NFA). When we consider the joint region $R_{ij} = R_i \cup R_j$, the number of valid points is the sum of the points at both regions:

$$(23) \quad n_{ij} = n_i + n_j.$$

Nevertheless, the maximal degree of coincidences in the joint region k_{ij} is always lower or equal than the sum of the degrees of coincidences at each region

$$(24) \quad k_{ij} \leq k_i + k_j.$$

The simple reason is that the first maximization is more constrained (to use a single transformation T_{ij} for both regions) than the second one (which uses a separate transformation for each region). Thus from (23) and (24) we deduce that:

$$P(R_i \cup R_j, T_{ij}) = \mathbb{P}[K_{ij} \geq k_{ij}] \geq \mathbb{P}[K_i + K_j \geq k_i + k_j] = P((R_i, T_i); (R_j, T_j))$$

This result is obvious in the discrete case where the distribution of \mathcal{K}_i becomes binomial (see next section). In the general case it is also true, because the probability $\mathbb{P}[k(R, T; \Theta) \geq l]$ is independent of T (it only depends on the distribution of Θ and the number of valid points $n = |R^*|$ as we shall see in Section 5) Therefore, recalling (8) both probabilities $\mathbb{P}[\mathcal{K}_{ij} \geq l] = \mathbb{P}[\mathcal{K}_i + \mathcal{K}_j \geq l]$, are equal because they both involve the same number of valid points $n_i + n_j$. This shows that without the term N_{transf} in (22), no merging would be done.

4.3. THE NUMBER OF TRANSFORMATIONS. It remains to estimate the number of transformations we can test at each region. This can be done by considering that each transformation is defined by three points. Suppose that we know that the range of the disparity values is $[M_{min}, M_{max}]$. Then, these points have values in that range and if we define a discretization step s , then we have the following number of possible points:

$$\frac{M_{max} - M_{min}}{s},$$

which leads to the following number of transformations:

$$N_{transf} = \left(\frac{M_{max} - M_{min}}{s} \right)^3.$$

Note that the discretization step s is in fact the precision we want to obtain at the final disparity map.

In the RAME algorithm we use a gradient descent approach, but the reached minimum gives the same solution as if we test all the possible transformations. Thus, we take the same number of tests.

5. CONTINUOUS AND QUANTIZED NFA FORMULATIONS

Computing $NFA(R, T)$ or the joint $NFA((R_i, T_i); (R_j, T_j))$ requires computing not only the number of tests, but also a probability of the form $\mathbb{P}(k(R, T; \Theta) \geq k(R, T; \theta))$. In the great majority of previous works using a contrario models, the NFA is defined in such a way that $k(R, T; \Theta)$ follows a binomial distribution. Recalling equation (8) it becomes clear that if we define $\tilde{\rho}(y) = \rho_\alpha(y)$ where

$$(25) \quad \rho_\alpha(y) := \begin{cases} 1 & \text{if } y > \alpha \\ 0 & \text{otherwise} \end{cases}$$

then

$$k(R, T; \Theta) = \sum_{x \in R^*} \mathbb{1}_{Y(x, T, \Theta) \leq \alpha}$$

has Binomial(n, p) distribution with $n = |R^*|$ as long as the probability $p = \mathbb{P}[Y(x, T, \Theta) \leq \alpha]$ is independent of the point x . This is the case, in fact, as it will be shown below. This choice of ρ will be called “quantized case”, and provides an interpretation of $k(R, T; \Theta)$ as the random number of coincidences.

On the other hand, it is very useful to consider a “continuous case” where $\tilde{\rho}(y) = y$ or $\tilde{\rho} = \rho$ equals the robust function from the minimization functional used in Section 2 for estimating the optimal T (see equations (5) and (6)). The main advantage of this continuous model is that it allows one to have a validation stage which is in perfect accordance with the estimation stage, thus allowing the benefit of the fine precision of the estimated parameters also in the validation stage. This can be useful for detecting barely meaningful structures whose meaningfulness is

also very sensitive to the parameters of T .

In this case $k(R, T; \Theta)$ is a continuous measure of the degree of coincidence.

In the sequel we shall use the following shorthand notations for the different random and deterministic variables involved :

$$(26) \quad \begin{aligned} Y^i &:= Y(\mathbf{x}, T; \theta) && \text{(observed degree of coincidence at point } \mathbf{x}_i) \\ \mathcal{Y}^i &:= Y(\mathbf{x}, T; \Theta) && \text{(random degree of coincidence at point } \mathbf{x}_i) \\ k &:= k(R, T; \theta) && \text{(observed degree of coincidence in region } R) \\ \mathcal{K} &:= k(R, T; \Theta) && \text{(random degree of coincidence in region } R) \end{aligned}$$

Given the observed values Y^i of the random variables \mathcal{Y}^i for all points $\mathbf{x}_i \in R^*$ we can compute the observed value k of the random variable \mathcal{K} , and the probability $\mathbb{P}[\mathcal{K} \geq k]$ under the background model. In the continuous case the probability distribution of \mathcal{K} is not known in a closed form, but it can be accurately approximated using Hoeffding's Theorem:

Theorem 1. ([26]) *Let $\mathcal{X}^1, \dots, \mathcal{X}^n$ be independent random variables with $\mu^i = \mathbb{E}[\mathcal{X}^i] \in (0, 1)$ and $\mathbb{P}[0 \leq \mathcal{X}^i \leq 1] = 1$, $i = 1, \dots, n$. Let $\mu = (\mu^1 + \dots + \mu^n)/n$. Then, for $0 < \eta < 1 - \mu$ and $\hat{\mathcal{X}} = (\mathcal{X}^1 + \dots + \mathcal{X}^n)/n$,*

$$\mathbb{P}[\hat{\mathcal{X}} - \mu \geq \eta] \leq e^{n w(\eta, \mu)},$$

where

$$w(\eta, \mu) = (\mu + \eta) \ln\left(\frac{\mu}{\mu + \eta}\right) + (1 - \mu - \eta) \ln\left(\frac{1 - \mu}{1 - \mu - \eta}\right).$$

Using this result, for $\mathcal{X}^i = \rho(\mathcal{Y}^i)$, we obtain $\hat{\mathcal{X}} = \frac{\mathcal{K}}{n}$ and we have the estimate

$$\mathbb{P}[\mathcal{K} \geq k] \leq e^{n w(\frac{k}{n} - \mu, \mu)}$$

where $\mu = \mathbb{E}[\frac{\mathcal{K}}{n}]$ and $n = |R^*|$ is the number of valid points in R . To estimate the value of μ we shall make precise the background model.

The following subsections describe the background models used for each case (RAF-MARC and RAME), and deduce the probability p needed for the quantized case, and the expected value μ needed for the continuous case.

5.1. VALIDATING THE AFFINE FIT TO MARC. Given a region $R \in \mathcal{R}$, let R^* be the set of valid points in R as explained in Section 2.1 To define the random variable $Y(x, T; \Theta)$ in the case of RAF-MARC, we consider the disparity $\Theta = \mathcal{M} : R^* \rightarrow [-h, h] \times [-v, v]$ as a random function. The disparity $\theta = M$ measured by MARC will be considered as a realization of this random variable. Then if $T \in \mathcal{A}$ is the affine transformation to be tested, we define

$$(27) \quad \mathcal{Y}^i = Y(\mathbf{x}_i, T; \mathcal{M}) = \|T(\mathbf{x}_i) - (\mathbf{x}_i + \mathcal{M}(\mathbf{x}_i))\|, \quad \mathbf{x}_i \in R^*.$$

This distance measures the degree of coincidence between the value of the tested transformation T at a point $\mathbf{x}_i \in R^*$ and a randomization of the disparity map \mathcal{M} at the same point. Then the random variable $k(R, T; \Theta)$ is defined by equation (8).

5.1.1. Background Model. Our purpose is to test if the (deterministic) affine model T assigned to the region $R \in \mathcal{R}$ is correct. To do so we define an *a contrario* or background model which assumes that the observed disparities $M(\mathbf{x}_i)$ are drawn from a random sampling in $\mathcal{M}(\mathbf{x}_i) \in [-h, h] \times [-v, v]$ without necessarily matching the hypothesized affine model $T - Id$. If according to this random sampling $NFA(R, T) < \epsilon$, meaning that T matches M too well to be explained by chance, the

random sampling hypothesis is rejected, and T is accepted as a good explanation of the measured disparities M .

Now we give more details on the computation of the probability in $NFA(R, T)$. In our case where the images have been rectified to epipolar geometry we know that $v = 0$ and h (if unknown) is estimated as the 99th percentile of the histogram of the observed $M(\mathbf{x})$ for all $\mathbf{x} \in \Omega^*$ in order to avoid outliers. Then the background model consists of considering $\mathcal{M}(\mathbf{x}_i)$ as i.i.d. random variables with uniform distribution $\text{Uni}[-h, h]$.

For the continuous case (where $\tilde{\rho} = \rho$ is the normalized Tukey function as explained in Section 2.1) we have to estimate the expected value μ of $\frac{\mathcal{K}}{n}$, in order to be able to use Hoeffding’s approximation of $\mathbb{P}[\mathcal{K} \geq k]$. According to equation (8) this amounts to computing

$$\mu = \frac{1}{|R^*|} \sum_{x_i \in R^*} (1 - \mu_i) = \frac{1}{|R^*|} \sum_{x_i \in R^*} (1 - \mathbb{E}[\rho(\mathcal{Y}^i)]),$$

If we denote by m_i the horizontal component of $(T - Id)(\mathbf{x}_i)$, and taking into account that $\mathcal{M}(\mathbf{x}_i) \sim \text{Uni}[-h, h]$ we get

$$\mu_i = \frac{1}{2h} \int_{m_i-h}^{m_i+h} \rho(y) dy \lesssim \frac{1}{2h} \int_{-\infty}^{\infty} \rho(y) dy = \frac{1}{2h} \int_{-c}^c \rho(y) dy$$

The last approximation is based on the assumption that the transformations $T \in \mathcal{A}$ to be tested will be not too far away from the identity, so that m_i will rarely be outside of the range $[-h+c, h-c]$, where $[-c, c]$ is the support of the Tukey function. This will in fact be the case as long as the range of disparities $h \gg c \sim 4\lambda$ is much larger than MARC’s precision level λ .

Alternatively we could search for an exact analytic closed form solution for $\mathbb{P}[\mathcal{K} \geq k]$ instead of using Hoeffding’s upper bound. This approach has been discarded due to the fact that the distribution of $\rho(\mathcal{Y}^i)$ shows a singularity at 0 whenever $\rho'(0) = 0$ (see [43] for more details).

In the quantized case, we only need to compute the probability

$$p = \mathbb{P}[\mathcal{Y}^i \leq \alpha] = \mathbb{P}[\|T(\mathbf{x}_i) - (\mathbf{x}_i + \mathcal{M}(\mathbf{x}_i))\| \leq \alpha]$$

since then the probability that appears in the NFA is just the binomial tail

$$(28) \quad \mathbb{P}[\mathcal{K} \geq k] = B(k, n, p) = \sum_{j=k}^n C_j^n p^j (1-p)^{n-j},$$

Using similar arguments as in the continuous case this probability p can be approximated by

$$p = \frac{1}{2h} \int_{m_i-h}^{m_i+h} \rho_\alpha(y) dy \lesssim \frac{1}{2h} \int_{-\infty}^{\infty} \rho_\alpha(y) dy = \frac{\alpha}{h}$$

5.2. VALIDATING RAME. For each region $R \in \mathcal{R}$, let $R^* := \{\mathbf{x}_i \in R : \|\nabla \tilde{u}(\mathbf{x}_i)\| > \gamma\}$. The threshold $\gamma > 0$ is used to ensure that the gradient orientations are not affected by the presence of noise. Large gradient vectors have a larger certainty on its orientation than smaller ones, and thus may be used for performing the statistical test [16, 19]. Using the notation of Section 2.2, if $T \in \mathcal{A}$ is the estimated motion parameter to be tested, then for each $\mathbf{x}_i \in R^*$ we consider the random variable

$$(29) \quad \mathcal{Y}^i = Y(\mathbf{x}_i, T; \mathcal{Z}) = \|\mathcal{Z}(u \circ T)(\mathbf{x}_i) - \mathcal{Z}(\tilde{u})(\mathbf{x}_i)\|,$$

as the random variable measuring their alignment. Since $\mathcal{Z}(u \circ T)(\mathbf{x}_i)$ and $Z(\tilde{u})(\mathbf{x}_i)$ have unit norm, the previous expression can be rewritten as

$$(\mathcal{Y}^i)^2 = 2(1 - \cos\beta)$$

where β is the angle that form the two vectors.

5.2.1. *Background Model.* Assuming that $Z(\tilde{u})(\mathbf{x}_i)$ is deterministic and that $\mathcal{Z}(u \circ T)(\mathbf{x}_i)$ are independent and uniformly distributed in $[0, 2\pi)$, we deduce that β is uniformly distributed in $[0, \pi)$.³

The random variable $\mathcal{K} = k(R, T; \mathcal{Z})$ is defined by (8). Notice that $1 - \frac{\mathcal{K}}{n}$ coincides with the average value of the energy on the region R .

Our purpose is to test if the affine model parameters assigned to the region $R \in \mathcal{R}$ are correct. For that issue we define the a contrario or background model which assumes that the observed angle β between the two vectors is a random variable with uniform distribution in $[0, \pi)$. The background or random model will be rejected if $\text{NFA}(R, T) < \epsilon$.

For the continuous case ($\rho(e) = \frac{1}{4}e^2$), we have to estimate the expected value μ of $\frac{\mathcal{K}}{n}$. The mean $\mu = (\mu^1 + \dots + \mu^n)/n$ is estimated by (see Theorem 1)

$$\mu^i = \mathbb{E}[\mathcal{X}^i] = \mathbb{E}[\rho(\mathcal{Y}^i)]$$

Thus,

$$\mu^i = \int_0^\pi \frac{1}{2\pi}(1 - \cos\beta)d\beta = \frac{1}{2}$$

Then, according to Theorem 1 we obtain $\mu = 1/2$.

In the quantized case, we consider ρ defined as in Eq. 25. In this case, ρ is a function that takes value 1 as the angle between the normals $\mathcal{Z}(u \circ T)(\mathbf{x}_i)$ and $Z(\tilde{u})(\mathbf{x}_i)$ is greater than a particular threshold, and 0 otherwise. The threshold is controlled via the parameter α of Eq. (25).

Knowing the background model for \mathcal{Y}^i we compute $p = \mathbb{P}[\mathcal{Y}^i \leq \alpha]$. Assume that the two normal vectors are aligned ($\rho(y) = 0$) if they form an angle below a threshold β . Thus, $\alpha^2 = 2(1 - \cos\beta)$ and p , the probability of alignment, is $p = \beta/\pi$.

6. EXPERIMENTAL RESULTS

6.1. *DATA SET.* Let us first introduce the data set that we use in the experiments and the error measure that we employ to evaluate the results. Our data set has been provided by CNES and consists of a pair of aerial images of the same real scene acquired with low disparity, with a B/H factor (baseline / altitude) of 0.045 and a ground resolution $R = 0.5$ meters/pixel (size of a pixel projected on the ground). This pair is shown in Figure 2. We also have the ground truth for this pair of images (Figure 2). Let us warn the reader that the images of the pair were taken with a difference of more than 20 minutes. Due to this delay some objects in the scene have an apparent motion and there are shadow movements that may cause some difficulty in estimating the disparity.

Note that these conditions do not exactly match those of the low baseline satellite application we are targeting (as described in Section 1). Quasi-simultaneous low

³ Observe that this choice ensures the independence between the segmentation \mathcal{R} of \tilde{u} (which is based on gradient orientations) and the background model for stereo matching (which only involves gradient orientations of u). Failure to satisfy this independence condition could lead to a violation of the NFA upper bound.

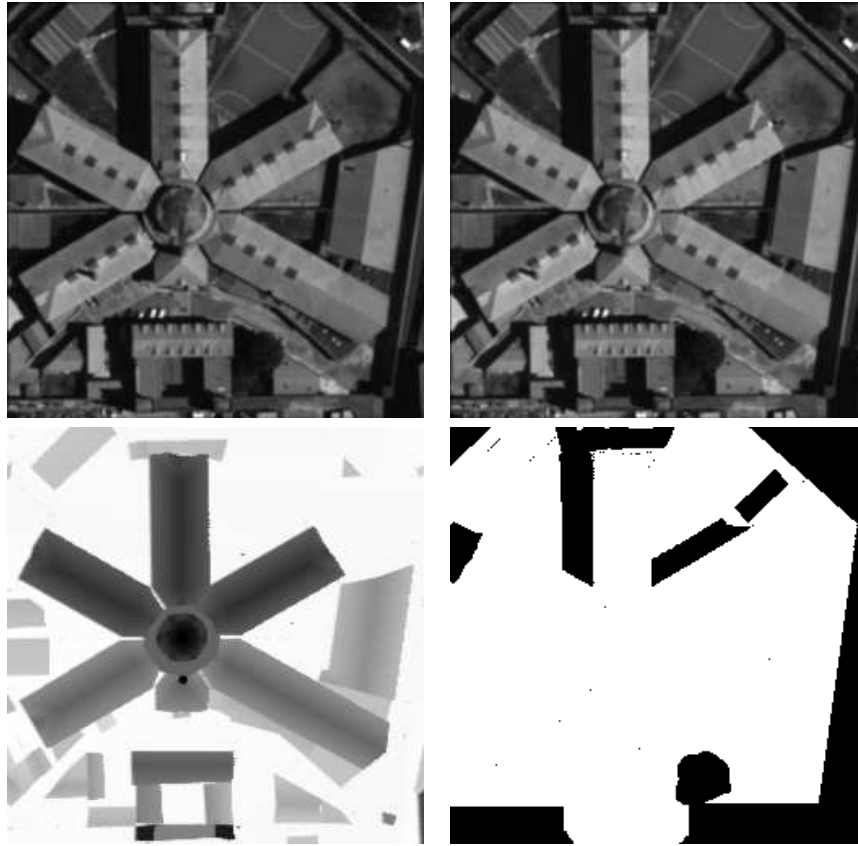


FIGURE 2. From left to right and top to bottom: the reference and the secondary images of the stereo image pair, the ground truth, and the mask without shadow zones used to compute the error.

baseline data with a little contrast change was unfortunately not available for an exhaustive evaluation at the time of writing this paper.

Our purpose in this section is to give experimental evidence of the performance of the previous algorithms. We shall consider the original MARC algorithm, a refinement of it obtained by anisotropic regularization [21], the RAF-MARC and the RAME algorithms. By displaying all these experiments we intend to show the improvements of MARC obtained by the piecewise affine region models. We shall also analyze the effect of the merging algorithm applied to the RAF-MARC and RAME results. Let us establish an abridged terminology to refer to these methods.

MARC: Original MARC version, see Section 2.1.

REG-MARC: A regularization of MARC results [21].

RAF-MARC: The Robust Affine Fitting of MARC (Section 2.1).

RAME: Region-based Affine Motion Estimation (Section 2.2)

MERGE-MARC: Region-Merging algorithm with merging criterion based on NFA over the MARC results (Section 4).

MERGE-RAME: Region-Merging algorithm with merging criterion based on NFA over the RAME results (Section 4).

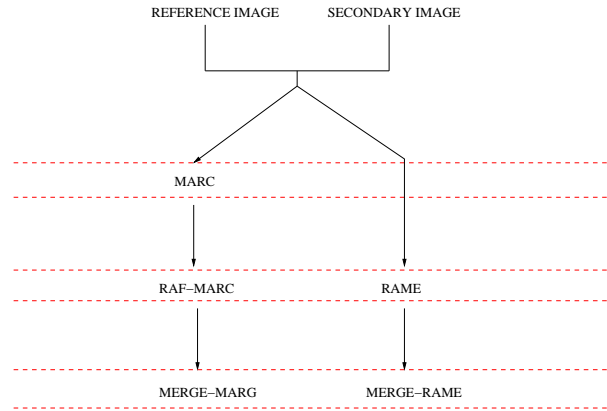


FIGURE 3. Diagram of the relations between the proposed algorithms.

Initial segmentations: The merging algorithm needs an initial segmentation. As already mentioned in the previous sections, two different simplified versions [5, 31] of Mumford-Shah segmentation algorithm [38, 37] have been used in our experiments.

Error measures: Since we have the ground truth for our image set, we can perform a quantitative analysis using the root mean squared error (RMSE) between both disparity maps, the estimated and the true ones.

To analyze our results we compute the RMSE not only on the whole image domain but also on subsets of it identified by a mask. In this work, six different masks are used. They are:

ALL: The mask is composed by all the points of the image.

SHADOW: Recall that the stereo image pair has been captured with a time difference of 20 minutes. This large time difference introduces large shadow differences between both images of the pair. Disparity measures are expected to be large in the shadow areas. Thus, we have constructed a mask where large shadow areas have been manually segmented (see Figure 2 (bottom left)).

MARC: The mask is given by the set of valid points determined by MARC algorithm. This mask is displayed in Figure 10 (top left)).

RAF-MARC: Validated regions of RAF-MARC algorithm.

RAME: A mask that includes the valid points of RAME algorithm is also constructed. These valid points result from applying the validation approach to each region of the image after estimating its disparity using RAME. This mask is shown in Figure 10 (bottom left).

MERGE1: Validated regions of MERGE-MARC algorithm (Figure 10 (top right)).

MERGE2: Validated regions of MERGE-RAME algorithm (Figure 10 (bottom right)).

In Table 1 we summarize the RMSE measurements of the different methods applied to the data set. The first column indicates the evaluated algorithm, denoted using the previously set terminology. In the second to sixth columns we report the RMSE for each of the algorithms computed on the mask specified at the top of the column. The percentage on each column represents the number of points over which the RMSE is computed. Note that the values of the RMSE are given in pixels. In order to get the RMSE in meters, we have to divide by $\frac{B}{H} \frac{1}{R} = 0.09$.

	<i>ALL</i>	<i>SHADOW</i>	<i>MARC</i>	<i>RAF-MARC</i>	<i>RAME</i>	<i>MERGE1</i>	<i>MERGE2</i>
	100 %	77.06 %	32.13 %	72.51 %	62.14 %	95.59 %	93.34 %
MARC	0.3223	0.2753	0.3195	0.3120	0.3235	0.3171	0.3223
REG-MARC	0.2733	0.2147	0.2602	0.2663	0.2761	0.2692	0.2732
RAF-MARC	0.3265	0.2941	0.3137	0.2652	0.3076	0.3239	0.3260
MERGE-MARC	0.2732	0.2221	0.2494	0.2577	0.2655	0.2666	0.2719
RAME	0.2529	0.2144	0.2346	0.2342	0.2150	0.2498	0.2532
MERGE-RAME	0.2477	0.2025	0.2290	0.2317	0.2106	0.2452	0.2439

TABLE 1. RMSE results.

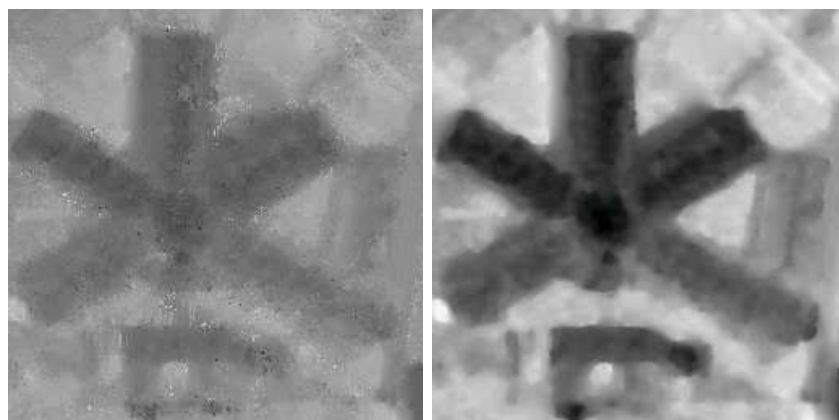


FIGURE 4. From left to right: disparity maps obtained by MARC and REG-MARC, respectively.

Analyzing the data in Table 1, we observe that the RMSE measures given by RAME method are smaller than the ones obtained by MARC, REG-MARC, and RAF-MARC. This may be explained since the real image pair has contrast differences due mainly to the motion of shadows and by the fact that the RAME method is contrast invariant. In spite of this, recent experiments seem to indicate that different MARC variants (especially MERGE-MARC and REG-MARC) perform better for almost simultaneous stereo image pairs which do not have a noticeable contrast change.

Let us also observe that the regularization used in REG-MARC and RAF-MARC improve the results of MARC, and that the merging method improves the result of the method on which it is applied, i.e., MERGE-MARC improves the results of MARC and RAF-MARC, and MERGE-RAME slightly improves the results of RAME. Let us finally point out that an improved regularization method [22] permits an improvement of the RMSE results of both methods: MARC and RAME.

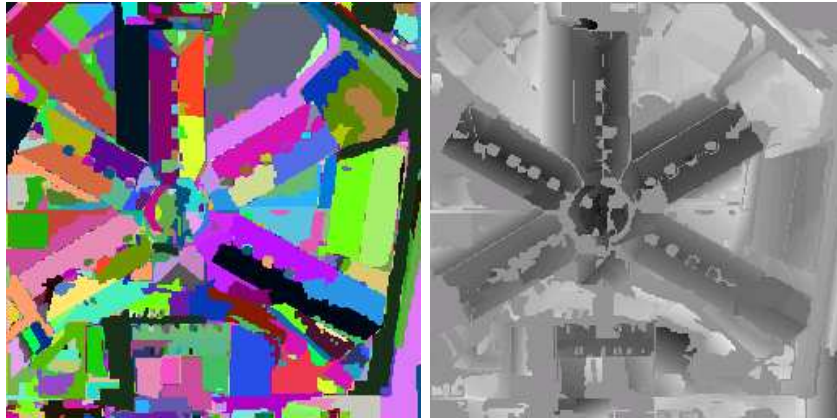


FIGURE 5. Inputs and result of RAF-MARC for the data set. From left to right: initial segmentation and disparity map obtained by RAF-MARC (non-validated regions appear as mid-gray).

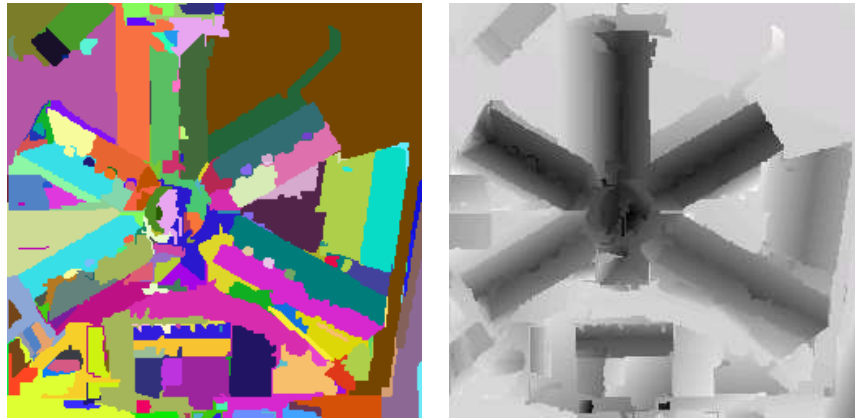


FIGURE 6. Results of MERGE-MARC for the data set. From left to right: final segmentation and disparity map obtained by MERGE-MARC.

Note that, for each method, we could take as a significant relevant measure the one computed using its corresponding mask (using MARC mask for MARC and REG-MARC methods). The different masks are displayed in Figure 10. In each image, the set of points that are found as incorrectly (resp. well) estimated are depicted in black (resp. white). Notice that the region-merging process for both methods (MARC and RAME) increases the number of validated regions, by approximately 60% for MARC and 30% for RAME as it can be seen from the percentages in Table 1.

In Figure 4 we display the disparity obtained with MARC and REG-MARC. The disparity map generated by MARC, displayed in Figure 4 (left), is used as input for REG-MARC (and also for RAF-MARC and MERGE-MARC methods). The disparity obtained with REG-MARC is displayed in Figure 4 (right). In Figure 5 we show the disparity map obtained with RAF-MARC.

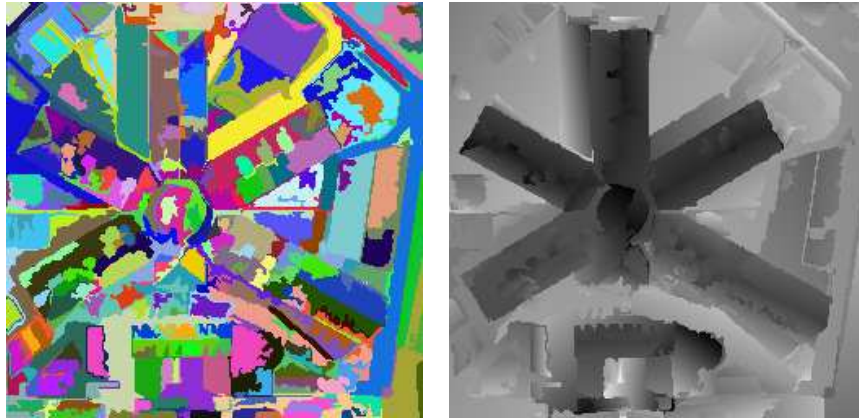


FIGURE 7. Results of RAME for the data set. From left to right: initial segmentation, and disparity map obtained by RAME.

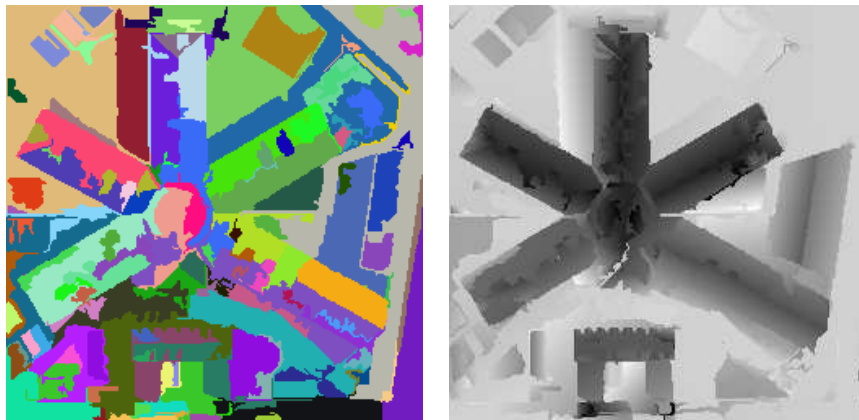


FIGURE 8. Results of MERGE-RAME for the data set. From left to right: final segmentation, and disparity map obtained by MERGE-RAME.



FIGURE 9. Detail of the results of MERGE-RAME. From left to right: original image crop, initial segmentation, and final segmentation.

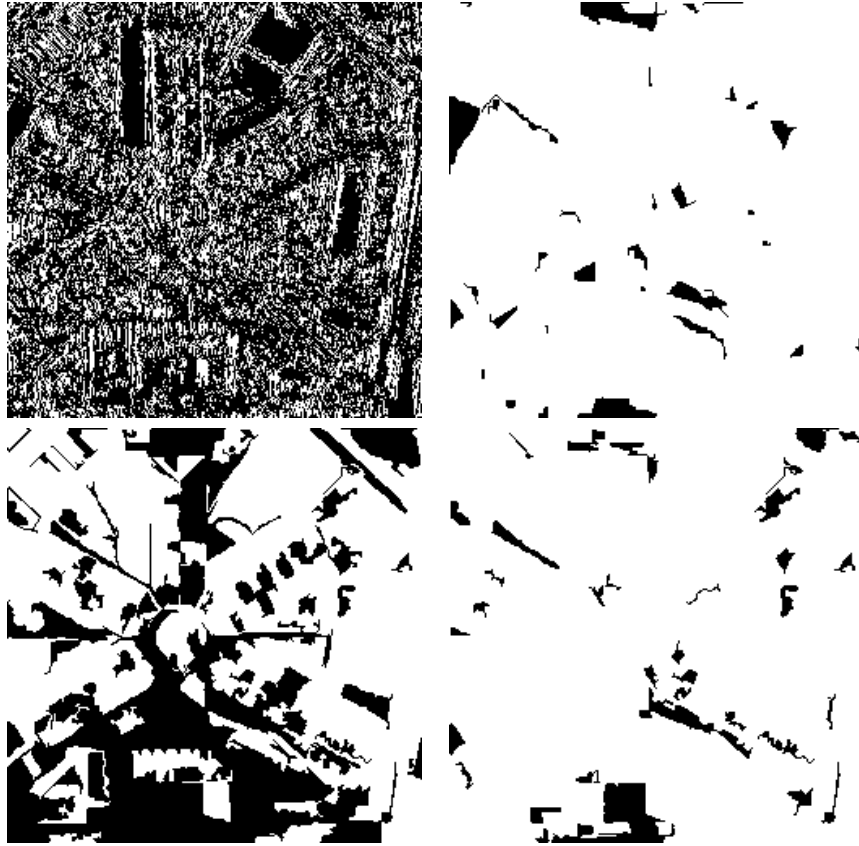


FIGURE 10. Several masks of valid points obtained by different methods. From left to right and top to bottom: mask of valid points of MARC, mask of valid regions of MARC-MERGE, mask of valid regions of RAME, mask of valid regions of MERGE-RAME.

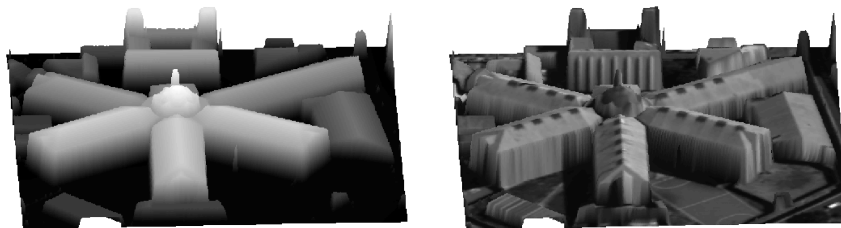


FIGURE 11. 3D representation of the ground truth of the data set. From left to right: with the disparity values as texture and with the reference image as texture.

In Figure 6 we display the disparity obtained with MERGE-MARC. As it can be appreciated, the disparity map is more regular and improves the results obtained with MARC or RAF-MARC. Moreover, we obtain in the final segmentation a good definition of the structures present in the scene.

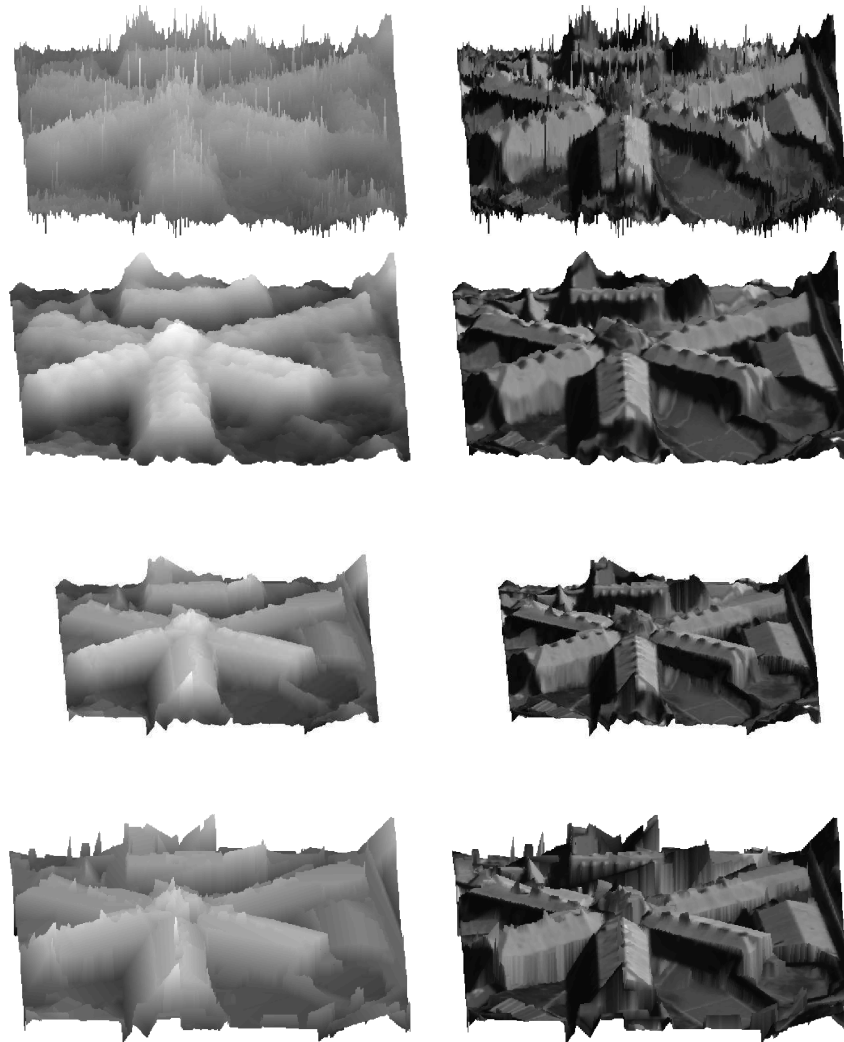


FIGURE 12. 3D representation of MARC, REG-MARC, RAF-MARC and MERGE-MARC results. In the first column we have used the disparity values as texture, in the second one we have used the reference image as texture. From top to bottom: results of MARC, REG-MARC, RAF-MARC and MERGE-MARC.

In Figure 7(right) we display the disparity obtained with RAME using the segmentation displayed in the left Figure. This segmentation of the reference real image is computed using the Mumford-Shah functional subordinated to the level lines of the image.

In Figure 8 we display the results of the MERGE-RAME algorithm: the final segmentation (left) and the final disparity map (right). In Figure 9 we have amplified a detail of the MERGE-RAME result. As it can be seen most of the mergings carried out are correct.

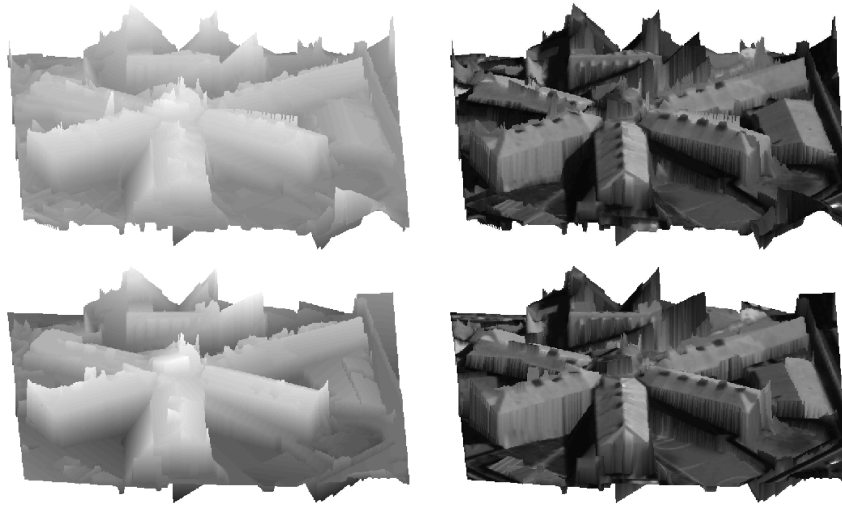


FIGURE 13. 3D representation of RAME and MERGE-RAME results. In the first column we have used the disparity values as texture, in the second one we have used the reference image as texture. From top to bottom: results of RAME, and MERGE-RAME.

Although Table 1 and the previous Figures give us an idea of the performance of the algorithms, for a better qualitative understanding we also display the reconstructed 3D scene. In Figure 11 we show the 3D representation of the ground truth. In Figures 12 and 13 we display the 3D representation for MARC and RAME, respectively.

7. CONCLUSIONS

In this work we have dealt with the computation of highly accurate subpixel disparity maps from stereo images of urban scenes. The subpixel accuracy property enables us to consider pictures almost simultaneously (small baseline), and consequently to remove most of the problems presented when the pictures are taken with a big baseline, as occlusions and object displacements.

We have evaluated an already existing method for disparity map computations (MARC, REG-MARC) which has been demonstrated to be suitable for subpixel accuracy, and proposed several ways to improve it (RAF-MARC, MERGE-MARC). These methods were compared with a second approach (RAME) which was originally proposed for motion estimation, and optimized here to the low baseline stereo case, and to be less dependent on the original segmentation (MERGE-RAME).

We have found that the RAME-based methods perform much better than the MARC based methods whenever both images from the stereo pair are quite different, due to illumination changes, which is consistent with the fact that this method is contrast invariant. For simultaneous stereo pairs, however, MARC-based methods seem to show a better performance which is in agreement with the fact that the MARC model better matches the additional information available in such data. Real quasi-simultaneous stereo pairs that will be produced by future earth observation

satellites will most likely show some aspects of both kinds of data, so our future research will try to integrate some aspects of both approaches.

We have presented a general way to define the number of false alarms of an event using a quantized and a continuous formulation. The continuous formulation allows us to combine the accuracy and robustness of variational methods with the reliability of computational Gestalt theory in terms of controlling the number of false positives, and validating the output.

In this sense the proposed approach can be generalized to further improve the accuracy by integrating into the model the exact values of the precision map instead of just a thresholded version of it. Such ideas proved very useful in the variational framework [22], and the ideas presented here shall help integrate such ideas into an *a contrario* framework.

We have applied this meaningfulness measure to define the merging order and merging criterion of a region-merging approach. This region-merging piecewise affine algorithm improves the disparity map results for both considered methods, MARC and RAME. Using this algorithm we obtain much better disparity maps in a quantitative and qualitative sense.

Thanks to the use of a contrario methods for fitting and grouping, our approach is almost parameterless. Still, its output depends significantly on the choice of initial segmentation which requires us to set its scale or number of regions by hand. To avoid this parameter that may be difficult to tune, future research should be focused on a contrario methods for a joint gray-level and disparity affine clustering approach, which performs in a single region-merging algorithm both gray-level based segmentation and piecewise affine disparity estimation and validation.

The presented merging algorithm can be seen as a first step towards a semantic description of the scene. In this sense, a simple piecewise affine model was analyzed, but the general method and the validation approach presented so far may be used in future research work for choosing –among a larger family of models with different degrees of complexity– the one which better explains the image measurements from a perception point of view.

8. APPENDIX: PROOF OF PROPOSITION 1

Let's write for convenience

$$\chi_{(R,T)} = \mathbb{1}_{(R,T) \text{ is } \epsilon\text{-meaningful}}$$

Then using the linearity of the expectation operator and Definition 2

$$\mathbb{E}[S] = \sum_{(R,T) \in \mathcal{S}} \mathbb{E}[\chi_{(R,T)}] = \sum_{(R,T) \in \mathcal{S}} \mathbb{P}[\hat{\mathcal{E}}_R \geq \hat{E}_R(T)] \leq \sum_{(R,T) \in \mathcal{S}} \frac{\epsilon}{N_{tests}(R)}$$

Now separating the set $\mathcal{S} = \cup_{l=0}^{l_0} \mathcal{S}_l$ and observing that for all $(R, T) \in \mathcal{S}_l$ we have $R = R_l$

$$\mathbb{E}[S] \leq \sum_{(R,T) \in \mathcal{S}_0} \frac{\epsilon}{N_{tests}(R)} + \sum_{l=1}^{l_0} \sum_{(R,T) \in \mathcal{S}_l} \frac{\epsilon}{N_{tests}(R_l)}$$

The first sum ($l = 0$) has NN_{transf} terms and the denominator $N_{tests}(R) = N(1 + C(R))N_{transf} \geq 4NN_{transf}$, since $C(R) \geq 1$ except in the trivial case where the partition \mathcal{R} has only one member.

The second sum ($l > 0$) has at most $2N$ terms (since $l_0 < 2N$), and the l -th term is

$$\sum_{(R,T) \in \mathcal{S}_l} \frac{\epsilon}{N_{tests}(R_l)} = \frac{C(R_l)N_{transf}\epsilon}{N(1+3C(R_l))N_{transf}} \leq \frac{\epsilon}{3N}$$

since $C(R_l) \geq 0$. Hence

$$E[S] \leq \frac{\epsilon}{4} + \frac{2\epsilon}{3} < \epsilon.$$

□

REFERENCES

- [1] A. Almansa, *Échantillonnage, interpolation et détection. Applications en imagerie satellitaire*, PhD thesis, École Normale Supérieure de Cachan, December 2002.
- [2] A. Almansa, V. Caselles, and G. Haro, *Total variation regularized image restoration: the case of perturbed sampling*, *Multiscale Modeling and Simulation*, **5** (2006)235 – 272.
- [3] A. Almansa, A. Desolneux, and S. Vamech, *Vanishing point detection without any a priori information*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(2003), 502–507.
- [4] A. Almansa, S. Durand, and B. Rougé, *Measuring and improving image resolution by adaptation of the reciprocal cell*, *Journal of Mathematical Imaging and Vision*, **21**(2004), 235–279.
- [5] C. Ballester, V. Caselles, L. Igual, and L. Garrido, *Level lines selection with variational models for segmentation and encoding*, *Journal of Mathematical Imaging and Vision*, 2005.
- [6] M. Z. Brown, D. Burschka, and G. D. Hager, *Advances in Computational Stereo*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(2003), 993–1008.
- [7] A. Bruhn, J. Weickert, and C. Schnörr, *Lucas/kanade meets horn/schunck: Combining local and global optic flow methods*, *IJCV*, **61**(2005), 211–231.
- [8] N. Camlong. *Description du multiresolution algorithm to refine correlation*, Technical report, CNES, 2001, Report cssi/111-1/cor-et-marc-2.
- [9] V. Caselles, L. Garrido, and L. Igual, *A contrast invariant approach to motion estimation*, *International Conference on Scale Space*, 2005.
- [10] A. Chambolle, *Total variation minimization and a class of binary mrf models*, Technical Report R.I. 578, Ecole Polytechnique, Centre de Mathématiques Appliquées, 2005.
- [11] H. Chen, P. Belhumeur, and D. Jacobs, *In search of illumination invariants*, *Int. Conf. on Computer Vision and Pattern Recognition*, (2000), 254–261.
- [12] D. Szeliski D. Scharstein, *Middlebury college stereo vision research page*, <http://www.middlebury.edu/stereo>, 2002.
- [13] J. Delon, *Fine comparison of images and other problems*, PhD thesis, Ecole Normale Supérieure de Cachan, 2004.
- [14] J. Delon and B. Rougé, *Le phénomène d’adhérence en stéréoscopie dépend du critère de corrélation*, GRETSI, Toulouse, France, 2001.
- [15] J. Delon and B. Rougé, *Analytic study of the stereoscopic correlation*, Technical report, CMLA, 2004, accepted for publication in JMIV.
- [16] A. Desolneux, *Événements significatifs et applications à l’analyse d’images*, PhD thesis, École Normale Supérieure de Cachan, 2000.
- [17] A. Desolneux, L. Moisan, and J.-M. Morel, *Edge detection by Helmholtz principle*, *Journal of Mathematical Imaging and Vision*, **14**(2001), 271–284.
- [18] A. Desolneux, L. Moisan, and J.-M. Morel, *Maximal meaningful events and applications to image analysis*, *Annals of Statistics*, **31**(2003), 1822–1851.
- [19] A. Desolneux, L. Moisan, and J.-M. Morel, “Gestalt Theory and Image Analysis,” *A probabilistic Approach*, 2006.
- [20] G. Facciolo, *Variational Adhesion Correction with Image Based Regularization for Digital Elevation Models*, PhD thesis, Universidad de la República Oriental del Uruguay, 2005.
- [21] G. Facciolo, A. Almansa, and A. Pardo, *Variational approach to interpolate and correct biases in stereo correlation*, GRETSI, Louvain-la-Neuve, Belgique, 2005.
- [22] G. Facciolo, F. Lecumberry, A. Almansa, A. Pardo, V. Caselles, and B. Rougé, *Constrained anisotropic diffusion and some applications*, *BMVC 2006*, Edimburg, UK, 2006.

- [23] O. Faugeras, Q.-T. Luong, and T. Papadopoulos, "The Geometry of Multiple Images : The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications," MIT Press, 2001.
- [24] A. Giros, B. Rougé, and H. Vadon, *Appariement fin d'images stéréoscopiques et instrument dédié avec un faible coefficient stéréoscopique*, patent number O4O3143, 2004.
- [25] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, 2004.
- [26] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association, **58** (1963), 13–30.
- [27] B. K. P. Horn and B. Schunk, *Determining optical flow*, AI, **17** (1981), 185–204.
- [28] P.J. Huber, "Robust Statistics," John Wiley & Sons, 1981.
- [29] L. Igual, *Image Segmentation and Compression using The Tree of Shapes of an Image. Motion Estimation*, PhD thesis, Universitat Pompeu Fabra, October 2005.
- [30] S.S. Intille and A.F. Bobick, *Incorporating intensity edges in the recovery of occlusion regions*, Proc. Intl Conf. Pattern Recognition, **1** (1994), 674–677.
- [31] G. Koepfler, C. Lopez, and J. M. Morel, *A multiscale algorithm for image segmentation by variational method*, SIAM, **31** (1994), 282–299.
- [32] V. Kolmogorov and R. Zabih, *Computing visual correspondence with occlusions via graph cuts*, IEEE International Conference on Computer Vision (ICCV01), (2001), 508–515.
- [33] Vladimir Kolmogorov and Ramin Zabih, *What energy functions can be minimized via graph cuts?* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **26**(2004), 147–159.
- [34] F. Lecumberry, *Cálculo de disparidad y segmentación de objetos en secuencias de video*, Master's thesis, Facultad de Ingeniería, Universidad de la Republica, Uruguay, 2005.
- [35] B. D. Lucas and T. Kanade, *An iterative image registration technique with an application to stereo vision*, Proc. of the 7th IJCAI, 674–679, Vancouver, Canada, 1981.
- [36] D. Marr and T. Poggio, *A computational theory of human stereo vision*, Proceedings of the Royal Society of London. Series B, Biological Sciences, **204** (1979), 301–328.
- [37] J.M. Morel and S. Solimini, "Variational Methods in Image Processing," Birkhäuser Verlag: Basel, 1994.
- [38] D. Mumford and J. Shah, *Boundary detection by minimizing functionals*, Proc. of IEEE ICASSP, (1985), 22–26.
- [39] V. Muron, *Manuel utilisateur de la chaîne de calcul de décalages entre images par l'algorithme MARC*, Technical report, CNES, 2003. Report cssi/111-1/cor-et-marc-5.
- [40] P. Musé, *On the definition and recognition of planar shapes in digital images*, PhD thesis, École Normale Supérieure de Cachan, 2004.
- [41] M. Ortner, X. Descombes, and J. Zerubia, *Building extraction from digital elevation model*, Technical report, INRIA Sophia-Antipolis, 2002.
- [42] M. Ortner, X. Descombes, and J. Zerubia, *Building outline extraction from digital elevation models using marked point processes*, International Journal of Computer Vision, (2006).
- [43] J. Preciozzi, *Dense urban elevation models from stereo images by an affine region merging approach*, Master's thesis, Facultad de Ingeniería – Universidad de la República, 11300 Montevideo, Uruguay., 2006.
- [44] B. Rougé, *Théorie de la chaîne image optique et restauration à bruit final fixé*, Mémoire en vue de l'obtention de l'habilitation à diriger des recherches. Option: Mathématiques appliquées, (1997).
- [45] S. Roy and I. J. Cox, *A maximum-flow formulation of the n-camera stereo correspondence problem*, in ICCV, (1998) 492–502.
- [46] D. Scharstein and R. Szeliski, *A taxonomy and evaluation of dense two-frame stereo correspondence algorithms*, International Journal of Computer Vision, **47** (2002), 7–42.
- [47] A.M. Tekalp, "Digital Video Processing," Prentice-Hall, 1995.
- [48] R. Grompone von Gioi, *Polygones significatifs*, Mémoire Master 2 MVA, 2006.
- [49] R. Grompone von Gioi and J. Jakubowicz, *Multisegment detection*, Personal communication.

Received for publication December 2006.

E-mail address: laura.igual@upf.edu

E-mail address: jprecio@fing.edu.uy

E-mail address: almansa@cmla.ens-cachan.fr

E-mail address: luis.garrido@upf.edu

E-mail address: vicent.caselles@upf.edu

E-mail address: Bernard.Rouge@cnes.fr