On the choice of the correlation term for multi-baseline stereo-vision

Martial Sanfourche, Guy Le Besnerais and Frédéric Champagnat
ONERA, DTIM/IED, BP-72, 92322 CHATILLON Cedex
{sanfour, lebesner, fchamp}@onera.fr

Abstract

We address DSM reconstruction from calibrated limited-angle aerial side-looking image sequences. We use a regularised approach which combines a multi-view pixel-wise similarity criterion and a $L1$-norm regularisation term. Although it gives quite good results, it has two main drawbacks: occlusions are not dealt with and the reconstruction improvement brought by addition of a new view becomes negligible beyond a certain number of views. We argue that these problems result from lacks of goodness-of-fit term. We propose to modify this term in order to obtain high resolution DSMs with occlusions handling. First developed on synthetic sequence, these modifications are evaluated here on a real sequence with good results.

1 INTRODUCTION

Recovering 3D structure from many calibrated views was initially taken up in [4]. Since, we can discern works leading to improve either matching criterion and aggregation method [2,5,8,11,13] or optimisation process [6,9,10] of multi-base stereo methods. All these works show that aggregation over many views performs better than classical 2-views stereo algorithms (see for instance [2, 3, 8]). However, some flaws are also encountered. Taking occlusions into account becomes much more complex than in two-view approach. Moreover, the reconstruction improvement brought by addition of a new view becomes negligible beyond a certain number of views. We argue that these problems can be partly solved by a better account of data. More precisely, in a regularised multi-baseline correlation approach, simple modifications of correlation term can lead to substantial improvement of 3D structure recovery. It should be emphasised that this is obtained without major modification of the algorithm nor dramatic increase of the computational cost.

The paper is organised as follows. Section 2 describes the framework of our study and the multi-view matching method used. Section 3 focuses on the occlusion handling by the "goodness-of-fit" term, while we show in section 4 that a smart use of the sequence permit to lead to an higher resolution result. Results on a synthetic sequence are given during the description of the method and results on a real sequence are collected in section 5.

BMVC 2004 doi:10.5244/C.18.72

2 PROBLEM STATEMENT

2.1 Geometric setting

We study scene reconstruction from aerial side looking image sequence whose acquisition configuration is shown in Figure 1. Note that angular gap between first and last image is low compared to the distance to the scene, *i.e.* the ratio between baseline and ground distance is small. The camera orientation is controlled throughout acquisition so as to maintain a particular scene point in the middle of field. In this way, overlapping is maximised.

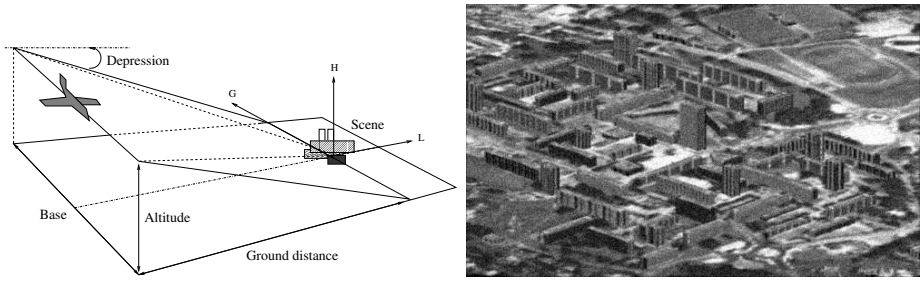

Figure 1: Left: Geometrical configuration of acquisition. Right: Middle view of synthetic sequence (baseline: 6 km and distance to scene center: 14.7 km)

In such small baseline settings, it seems better to reconstruct a DSM (digital surface model) in the geometry of one reference image (for practical point of view, central view) rather than in a planimetric reference coordinate system [14]. More precisely, we reconstruct height relative to an horizontal plane deduced from reference camera orientation. Indeed, in oblique viewing, height representation is better than depth representation because estimation precision is slowly variable in field of view (see [11] for more on this subject).

2.2 A multi-view regularised reconstruction method

The matching process is guided from world space by a plane sweep strategy [2] and based on calculation of a radiometric criterion under two strong assumptions: no occlusion occurs and we assume conservation of intensity among different frames (see [3, 8] for similar approaches).

As calibration of camera is assumed accurate (see [11] for calibration issues), for reference pixel $\boldsymbol{x}$ and height hypothesis h, corresponding pixel in view k, $\boldsymbol{x}_k$, can be obtained by a planar homography. The projected pixel generally does not fall on image grid and the gray level $I_k(\boldsymbol{x}_k)$ is retrieved by bilinear interpolation. Considering K views, we construct a radiometric vector $\boldsymbol{V}_1^K(\boldsymbol{x}, h) = \{I_k(\boldsymbol{x}_k(\boldsymbol{x}, h)\}_{1 \leq k \leq K}$. The likelihood of height hypothesis, $C(\boldsymbol{x}, h)$, is given by the standard deviation of this vector, eq.(1). For a valid height hypothesis $C(\boldsymbol{x}, h)$ should be minimal.

$$C(\boldsymbol{x}, h) = \hat{\sigma}(\boldsymbol{V}_1^K(\boldsymbol{x}, h)). \tag{1}$$

Various authors have previously noted that spatial regularisation is required to reduce matching noise. It can be achieved by use of window correlation [8] or in a regularised framework. For an oblique point of view, fronto-parallel hypothesis on scene patch implied by correlation window techniques is unrealistic. Therefore we rely on a pixel-wise criterion with a penalised regularisation to enforce spatial homogeneity of results.

Thus we minimise penalised similarity criterion

$$J_1(h) = \sum_{\boldsymbol{x}} C(\boldsymbol{x}, h(\boldsymbol{x})) + \lambda \sum_{\boldsymbol{x}' \in \mathcal{N}_4(\boldsymbol{x})} |h(\boldsymbol{x}) - h(\boldsymbol{x}')| \tag{2}$$

where $\mathcal{N}_4(\boldsymbol{x})$ is the 4-connexity neighbourhood of pixel $\boldsymbol{x}$ and λ is a regularisation parameter. The choice of $L1$-norm regularisation avoids excessive penalisation of large height discontinuities (as it occurs with more usual L2 norm), however, such a regularisation is rather crude: it constraints solution to be piecewise flat even on smooth surface. Note that minimisation is achieved by an efficient algorithm based on graph cuts [10].

2.3 Results and discussion

We present here results on a synthetic 61 frames sequence, obtained by covering digital elevation model (DEM) with real texture (the reference view is shown on Figure 1). We use two quantities to evaluate criteria performance. On one hand the accuracy on good matches is appraised by statistics on best 90% error samples. On other hand we quantify amount of false matches due to poor textured areas or occlusions by percentage of pixels with an error > 10 m.

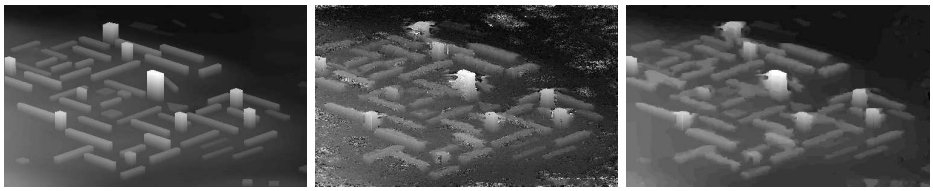

Figure 2: Results of regularised multi-baseline method on synthetic example: true height map (left), estimated height map without regularisation (middle) and estimated regularised height map (right). Gray-scale is linear in height range [−30 m, 50 m].

On this example, regularisation brings a dramatic improvement over non-regularised matching as shown on Figure 2 and by two first lines in Table 1. We observe a good compromise between noise smoothing and discontinuity preservation, except on neighbourhood of tallest buildings because similarity criterion of eq.(1) is not robust to occlusions.

Table 1 shows how reconstruction error evolves according to number of frames used for a constant baseline. Note that for each setting the regularisation parameter λ is chosen in order to minimise quadratic error to true DSM. Beyond 20

Number of frames	bias	RMS	$L1$-norm	outliers
61 views without $L1$-norm reg.	-0.43	3.64	2.57	11.6%
61 views and $L1$-norm reg.	-0.23	1.0	0.78	3.4%
5 views and $L1$-norm reg.	-0.30	1.30	0.98	3.7%
11 views and $L1$-norm reg.	-0.26	1.12	0.87	3.3%
21 views and $L1$-norm reg.	-0.24	1.08	0.83	3.3%

Table 1: Performance of criterion of eq.(1) with respect to number of frames (in meters), compared with a non regularised approach.

views no gain is brought by a new frame, as predicted by theoretical evaluation of triangulation process (see [3]).

In our opinion, these flaws are related to a sub-optimal use of the image sequence. We propose to reduce them by modifying data term, without any modification of rest of method (algorithm and regularisation). In this line, section 3 will discuss how to deal with occlusions while the section 4 will describe how data can be properly integrated to lead to a higher resolution DSM.

3 DEALING WITH OCCLUSIONS

In two-views stereo-vision, occlusion detection is commonly handled thanks to dynamic programming algorithms. However multi-frame situation is much more complex, as there are $(K-1)$ binary visibility values for each reference view pixel. Previous approaches fall into two categories which are recalled in sequel.

3.1 Geometric vs. visibility patterns dictionary

In geometric approach, relationship between scene relief and visibility is made explicit. Visibility flags, indicating if a pixel $\boldsymbol{x}$ is visible in view k, is written as function of height map and motion parameters. This dependency prevents direct DSM reconstruction. This joint estimation problem is addressed by an iterative process, alternating partial DSM reconstruction and visibility computation. Voxel coloring approaches [13] handle occlusions by ordered-depth plane exploration: hard decisions about voxel occupancy are taken and corresponding pixels are marked when a voxel is kept. Note that this framework don't lent itself to regularised approach : Kang et al. in [5] show that convergence of such iterative process in a regularised framework can be erratic.

In [7], Nakamura et al. propose to use a dictionary of "visibility patterns", *i.e.* a set of binary vectors indicating valid views. For each part of the scene visibility pattern is selected thanks to a decision rule, for instance minimum rule $\min_P\{C_P(\boldsymbol{x}, h)\}$, leading to a global criterion $C(\boldsymbol{x}, h)$. Of course, dictionary choice is crucial. On one hand, full set of visibility patterns heads back to geometric approach, on other hand, too crude a modelisation yields poor results.

Nakamura et al. identify their dictionary by a statistical analysis of occlusion phenomenon from synthetic data with ground truth. They use a biased minimum

rule to compensate for semi-occluded visibility pattern advantage. For these configurations the similarity criteria are multiplied by a weight w_P greater than 1, depending on number of valid views.

Using notations from section 2, we write Nakamura's criterion as

$$C^{\text{Nakamura}}(\boldsymbol{x}, h) = \min\{\hat{\sigma}(\boldsymbol{V}_1^K(\boldsymbol{x}, h)), w_P \hat{\sigma}(\boldsymbol{V}_P(\boldsymbol{x}, h)) \ \forall P\}. \tag{3}$$

More recently, Kang et al. [5] proposed a simple two-patterns dictionary for side-looking sequences. They assume that a point visible in reference view will be occluded either in the "left part" of sequence, *i.e.* views preceding reference, or in the "right part" of sequence. As two patterns have same length, no weight is needed in decision rule. We write Kang's criterion as

$$C^{\text{Kang}}(\boldsymbol{x}, h) = \min\{\hat{\sigma}(\boldsymbol{V}_1^r(\boldsymbol{x}, h)), \hat{\sigma}(\boldsymbol{V}_r^K(\boldsymbol{x}, h))\}. \tag{4}$$

This simple adjustment improves noticeably reconstruction around depth discontinuities [5], located in our examples near tall buildings. Note however that all pixels are considered as "half-occluded" points, although most of them are actually visible in whole sequence. As a result, reconstruction with criterion (4) shows a higher RMS score (see Table 2).

3.2 An alternative decision rule

We adopt dictionary strategy and we propose a new decision rule. A first decision is made about pixel half-occlusion and a second about left-occluded case or right-occluded case. The first decision is made by comparison between standard deviations of $\boldsymbol{V}_1^r(\boldsymbol{x}, h)$ and $\boldsymbol{V}_r^K(\boldsymbol{x}, h)$: for a good candidate height, these values should be both approximately equal to the noise level. However, if point is half-occluded, one of these radiometric vectors is augmented by difference of intensity around point and intensity of region which occludes it. Therefore we threshold standard deviation difference following

$$D_s = |\hat{\sigma}(\boldsymbol{V}_1^r(\boldsymbol{x}, h)) - \hat{\sigma}(\boldsymbol{V}_r^K(\boldsymbol{x}, h))|$$

and obtain "mixed" criterion

$$C^{\text{Mixed}}(\boldsymbol{x}, h) = \begin{cases} C^{\text{Kang}}(\boldsymbol{x}, h) \text{ if } D_s > t \\ C(\boldsymbol{x}, h) \text{ otherwise} \end{cases} \tag{5}$$

similarity	bias	RMS	$L1$-norm	outliers
C	-0.23	1.0	0.78	3.4%
C^{Nakamura}	-0.18	0.88	0.72	1.8%
C^{Kang}	-0.21	1.16	0.92	1.45%
C^{Mixed}	-0.15	0.85	0.70	1.85%

Table 2: Performance of various similarity criteria with $L1$-norm regularisation.

Using an *ad hoc* threshold of $t = 8$ for a 8-bit image sequence, we obtain good results both around discontinuities and in low areas, as shown by Figure 3

compared with Figure 2. Table 2 shows a quantitative comparison between four criteria C, C^{Kang}, C^{Nakamura} and C^{Mixed} on the synthetic sequence. Kang's criterion decreases outliers percentage at price of a degraded precision. Our strategy gives interesting results. Error over best 90 percent of samples are as good as criterion C and, in a same time, number of outliers is reduced. Note that we obtain similar results with Nakamura's criterion.

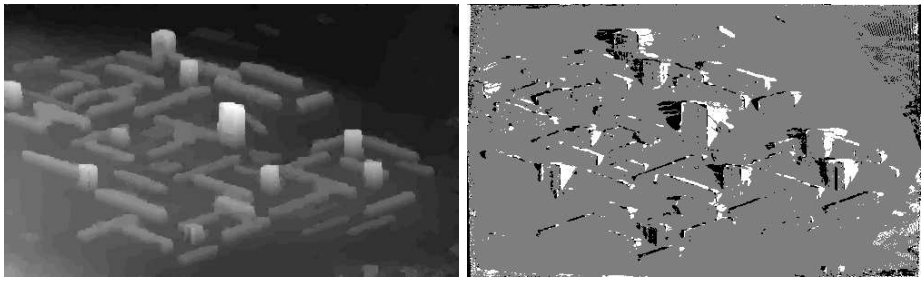

Figure 3: Synthetic data. On left: estimated height map using "Mixed" similarity (5) with threshold= 8 and $L1$-norm regularisation ($\lambda = 1.25$). On right corresponding *a posteriori* decisions about visibility.

Moreover, our method allows an *a posteriori* visualisation of decisions made by algorithm among the three status non-occluded/left-occluded/right-occluded of each pixel, shown respectively in gray, white or black color in right part of Figure 3. We observe a very good agreement between decisions and our knowledge of 3D scene. Thanks to the regularisation term occluded part are indeed compact and localised around buildings, although probably slightly over-estimated.

4 DSM RESOLUTION IMPROVEMENT

As shown on Table 1, for a short baseline and when video frame rate increases, stereo effects between 2 successive views become too small to give a relevant contribution to 3D estimation. In opposite, these small inter frame motions can be useful to compute a higher resolution version of image sequence by a super resolution method [12].

The sequence is split into "Group of Pictures" (GoP). In each GoP a reference frame is selected: entire GoP will be processed to obtain a higher resolution version of its reference frame. As stereo effects are very small inside a GoP, a parametric motion model can be used between reference view and an other view of GoP (called inspection view). In practice we use an homographic model and parameters estimation is conducted though a gradient based method (see [1] for a similar approach). The estimated motion is used to build the motion compensated subsampling matrix which gives transform between the high resolution frame and GoP frames. Quadratically regularised inversion of these data is formulated as a Bayesian estimation problem and achieved by a conjugate gradient algorithm. We use our own implementation of the method described in [12]. In practice, each GoP contains 5 images and we compute twice as better resolved version of GoP

reference view.

High resolution sequence is computed before matching process which does not need to be modified. Our approach can be seen as a two step temporal integration process. First, a local integration within each GoP to obtain higher resolution images. Then aggregation of all super resolved frames in regularised 3D reconstruction.

To assess interest of this approach, we test it on a half-resolution version of synthetic sequence (in contrast, the original synthetic sequence is called "standard resolution sequence" in sequel). Then we compare height maps computed from:

- standard resolution sequence
- a super resolved sequence obtained from low resolution data

A third height map is computed on low resolution sequence but using four times as many pixel in reference view. Such a process is equivalent to reconstruction from a bilinearly interpolated sequence.

Results are joined in Table 3 and in the Figure 4.

Sequence type	bias	RMS	$L1$-norm	outliers
Standard Resolution	-0.16	1.02	0.82	2.3%
Super-resolution	0.38	1.10	0.90	2.2%
Bilinear interpolation	-0.23	1.25	0.95	3.0%

Table 3: Performance of C^{Mixed} criterion with $L1$-norm regularisation for three 15 frames sequences.

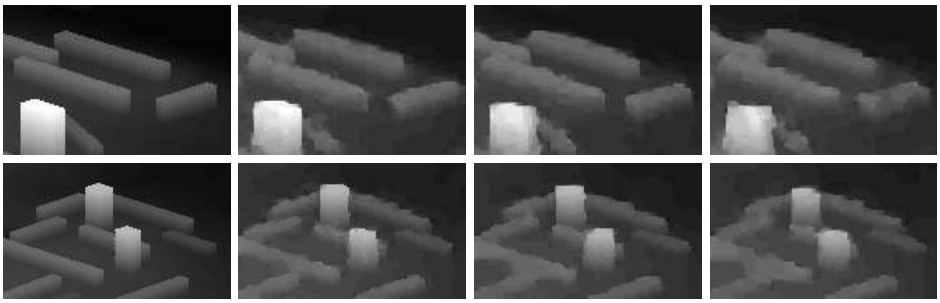

Figure 4: Comparison of 2 DSM improvement methods (criterion: C^{Mixed}) illustrated by two local areas of DSM. From left to right: ground truth, standard resolution, super resolution and bilinear interpolation.

Table 3 shows that super resolution approach gives a solution a bit closer to standard resolution reconstruction than a simple bilinear interpolation approach.

Partial views on Figure 4 show a good localisation of buildings even if some errors occur in form of spurious junctions between buildings, which are less flagrant with the standard resolution sequence. Building contours obtained by the super-resolution approach are more regular than with a bilinear interpolation approach. Note that this result comes from reduction of aliasing in super resolution techniques.

These results show that one can obtain DSM with higher resolution than original sequence on condition that frame rate is high enough to neglect stereo effects between neighbouring frames.

5 RESULTS ON REAL AERIAL SEQUENCE

5.1 Real sequence

We present results on a real aerial sequence ("Town1"). Distance to scene is approximately 3 kilometers, average camera altitude is 500 m. This sequence, taken on a . 900 m baseline, includes 101 8-bits images. Reference image is 51^{th}, shown in Figure 5. As raw calibration data are not accurate enough, a refinement is necessary beforehand (see [11] for more on this topic).

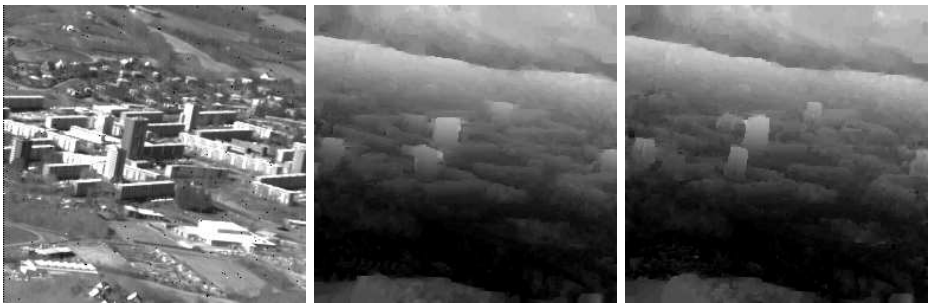

Figure 5: "Town 1". Left: reference view. Middle and Right part, DSM after $L1$-norm regularisation for, respectively, criterion of eq.(1) and C^{Mixed} criterion.

5.2 Results

Figure 5 shows estimated height maps with and without dealing with occlusions. Standard criterion of eq.(1) induces smooth map but a bad building localisation which seem wider than real. Criteria dealing with occlusions correct this bias. Note that these different results show a compromise between good segmentation of tallest buildings and effective detection of smaller ones. C^{Kang} of eq.(4) gives a very good segmentation of the tallest elements but on right part of town, it is difficult to recognise buildings. Results obtained with C^{Mixed} and C^{Nakamura} are close together and perform better than C^{Kang} in regions where buildings are small and close to each other (see Figure 5). Finally, in upper part of reconstruction, where planimetric resolution is low, hill ridge is well detected. Farther, we estimate only large scale relief.

As resolution is rather low for sequence "Town1", resulting DSM is not easily exploitable, for instance to segment building and ground. The proposed super resolution approach allows to obtain more effective results as shown in Figure 6. Buildings are better resolved, even in areas where buildings are close to each other.

There is more residual matching noise in the bilinear approach. Once again aliasing reduction in super-resolved images improves DSM estimation.

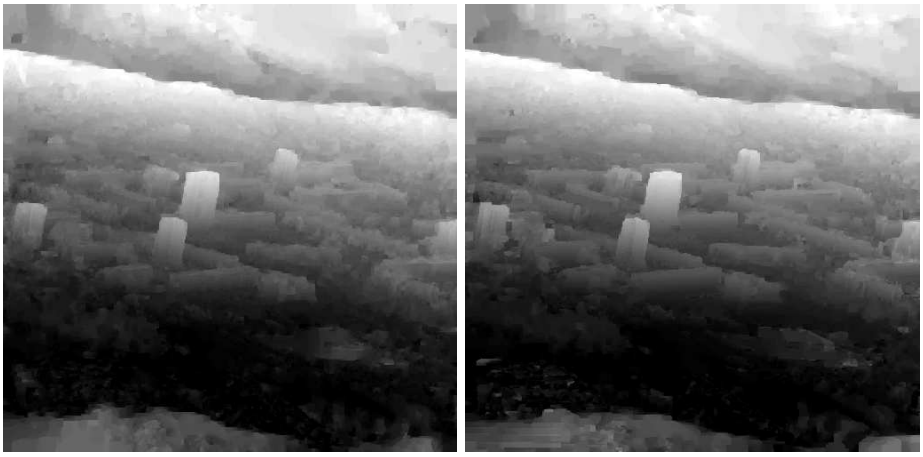

Figure 6: High Resolution DSM on "Town1" sequence (criterion: C^{Mixed}). Left: DSM obtained by bilinear interpolation of reference view. Right: DSM obtained by super-resolution method.

6 CONCLUSION

In this paper we have proposed some modification of the goodness-of-fit term leading to multi-view DSM estimation method improvements. We showed that data term modifications in order to deal with occlusions as proposed by [5, 7] are useful and proposed a new modification without any supplementary computational cost well adapted to side looking sequences. We are currently investigating some automatic methods to avoid *ad hoc* choices of threshold or weighting coefficient. We have also demonstrated DSM resolution improvement by exploiting in a novel way sequence frames when inter-frame motion is small. We believe that such an improvement should be reachable with many video sequences.

References

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, March 2004.

[2] R. Collins. A space-sweep approach to true multi-image matching. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 358–363, San Francisco,USA, June 1996.

[3] B. Géraud, G. Le Besnerais, and G. Foulon. Determination of dense depth map from an image sequence: application to aerial imagery. In *European Sym-*

posium on Remote Sensing, Image and Signal processing for Remote sensing IV, September 1998.

[4] T. Kanade, M. Okutomi, and T. Nakahara. A multiple-baseline stereo method. In *proc. DARPA Image Understanding Workshop*, pages 409–426, San Diego, California (USA), January 1992.

[5] S. B. Kang, R. Szeliski, and Chai J. Handling occlusions in dense multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 103–110, Kauai, Hawaii (USA), December 2001. IEEE.

[6] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV'02*, May 2002.

[7] Y. Nakamura, T. Matsuura, Satoh K., and Ohta Y. Occlusion detectable stereo - occlusion patterns in camera matrix. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 371–378, San Francisco, USA, June 1996.

[8] N. Paparoditis, G. Maillet, F. Taillandier, H. Jibrini, L. Guigues, and D. Boldo. Multi-image 3d feature and dsm extraction for change detection and building reconstruction. In Baltsavias et al., editor, *Automatic Extraction of Man-Made Objects from Aerial and Space Images (III)*, pages 217–230, 2001.

[9] S. Paris and F. Sillion. Robust acquisition of 3d informations from short image sequences. In *Pacific Graphics*. IEEE Computer Society, october 2002.

[10] S. Roy and I. Cox. A maximum-flow formulation of the n-camera correspondence problem. In *International Conference on Computer Vision (ICCV)*, pages 498–499, Bombay, India, January 1998. IEEE.

[11] M. Sanfourche, G. Le Besnerais, and S. Philipp-Foliguet. Height estimation using aerial side looking image sequences. In *Photogrammetric image analysis*, volume 34, pages 33–38, Munich, Germany, September 2003. ISPRS.

[12] R.R. Schultz and R.L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE tr. on Image Processing*, 5(6):996–1011, June 1996.

[13] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *Int. J. Computer Vision*, 35(2):151–173, 1999.

[14] R. Szeliski. Scene reconstruction from multiple cameras. In *ICIP'00*, pages 13–16, September 2000.