HEIGHT ESTIMATION USING AERIAL SIDE LOOKING IMAGE SEQUENCES

Martial Sanfourche, Guy Le Besnerais and Sylvie Philipp-Foliguet

ONERA, DTIM/IED, BP-72, 92322 CHATILLON Cedex
sanfour@onera.fr, lebesner@onera.fr

Equipe Traitement des Images et du Signal (CNRS UMR 8051)
University of Cergy-Pontoise/ENSEA
6, av. du Ponceau 95014 Cergy-Pontoise Cedex
philipp@ensea.fr

Commission III, WG III/2

KEY WORDS: Surface reconstruction, Images sequence analysis, Occlusions, Bundle adjustment

ABSTRACT

We are concerned with DSM reconstruction from limited-angle aerial side-looking image sequences. Despite this suboptimal configuration, we have devised a complete process starting from a partially calibrated sequence and leading to an estimated height map. The calibration step is fully automatic: it is made with respect to a reference view by means of interest points tracking and bundle adjustment. We propose a novel dense matching process combining a multi-views pixelwise similarity criterion, a L1-norm regularization term and a final step of height map refinement. Occlusions are accounted for without supplementary computational cost by a modification of the similarity criterion. First developed on synthetic sequences, this method is evaluated here on a real sequence with promising results.

RÉSUMÉ

Nous étudions l'estimation de MNS à partir de séquences d'images aériennes en visée latérale présentant un angle stéréoscopique faible. Malgré cette configuration difficile, nous avons développé une méthode d'exploitation de séquences d'images aériennes partiellement calibrées aboutissant à une carte de hauteurs estimées sur la zone considérée. La calibration est automatique et utilise un ajustement de faisceaux sur des points d'intérêt suivis sur la séquence. Le procédé de mise en correspondance dense combine un critère de similarité sur toutes les vues et un terme de régularisation en norme L1. Une étape finale d'affinage de la carte des hauteurs est effectuée avec un critère L2-L1. Cette méthode a été développée sur des séquences synthétiques et fait l'objet ici d'une évaluation sur une séquence réelle : elle fournit des résultats convaincants.

1 INTRODUCTION

We study the estimation of digital surface model (DSM) from aerial image sequences over urban areas. High precision (i.e. submetric) DSM are useful for many industrial and military tasks and require frequent updates to account for the fast changes which arise in urban areas. DSM are nowadays estimated using multiple views taken in various orientations, in order to obtain high precision and density (Paparoditis et al., 2001). However, it is sometimes necessary to handle degraded observational conditions. We focus here on the use of a single side-looking image sequence taken from a lateral viewpoint with limited angular exploration. Such a situation arises when the aircraft cannot fly upon some areas of interest and when it should stay at a low altitude. In these situations, the estimation of a partial and (perhaps) degraded DSM is useful, for instance to update an older DSM, and detect changes in high structures like buildings over the considered area. There are few references which deal with this difficult problem. Our work is in the line of the multi-baseline approach of (Le Besnerais and Duplaquet, 2002, Géraud et al., 1998) (see also the space-sweep methods described by (Collins, 1996)). Another approach is based on optical flow estimation (Mandelbaum et al., 1999) but yields, according to our experience, a lower precision.

We describe a complete process to derive a partial DSM from a partially calibrated side-looking sequence of images. More precisely, the obtained DSM consists in the estimated heights of every structure visible in a reference image, which is generally chosen in the middle of the sequence. The main features of the process are: (1) use of a global pixel-wise similarity score computed with all the images of the sequence, which leads to a temporal integration of the information; (2) spatial regularization with a discontinuity-preserving $L1$-norm penalization term whose optimization is achieved thanks to an efficient graph-based algorithm; (3) explicit account for occlusion effects in the criterion and estimation of occluded parts. Results are provided for synthetic as well as real sequences.

The paper is organized as follow. Section 2 describes in more details the experimental configuration, gives the necessary definitions and an overview of the method. Section 3 focuses on aerotriangulation. The dense matching process of the images is presented in section 4, while section 5 emphasizes on how to handle occlusions. Results on a synthetic sequence are given during the description of the method and results on a real sequence are collected in section 6.

2 PROBLEM STATEMENT

2.1 Geometric setting

The generic experimental configuration that we consider is shown in Figure 1. We name this kind of configuration "spotlight sequence" because the camera orientation is controlled throughout the acquisition to maintain a particular point of the scene in the

middle of the field, so as to maximize the recovering part between the images. The main geometrical parameters are the baseline B (ie. the nearly horizontal translation between the first and last views), the ground distance to the scene D_0 and the flight altitude H_0, which defines the mean depression angle θ_0 of the camera during the acquisition.

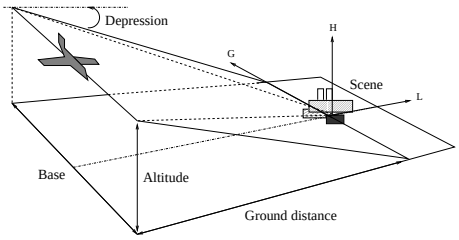

Figure 1: Geometrical configuration of the acquisition

The *a priori* evaluation of the stereoscopic precision in this configuration is more difficult than the usual close to nadir stereoscopy. Essentially, with respect to the central camera, which we take as a reference, the precision of the estimated depth of a 3D point is limited by the ratio between its distance to the central view and the baseline. However, for a relatively flat area, this precision is highly variable in the field of view, because of the low depression angle. For instance, denoting θ the angular depression of a particular pixel (see Figure 2.1), the depth precision is governed by the derivative $\delta Z/\delta p = 2H/B(\sin\theta)^2$, where δp is the angular size of a pixel. We choose to reconstruct the height H: its stereoscopic precision is also variable in the field of view ($\delta H/\delta p = 2H/B\sin\theta$), but its range is correct (from $1.1:1$ to $2:1$) in the cases we consider (small field of view).

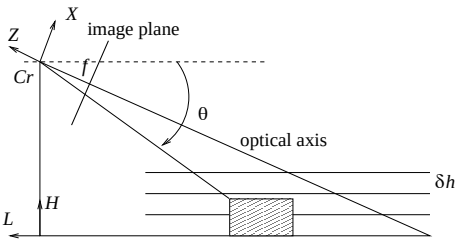

Figure 2: Reconstruction space and reference coordinate system (C_r, X, Y, Z)

From a practical point of view, the central view coordinate system is chosen as reference, and the horizontal plane is simply deduced from the reference camera orientation parameters. The reconstruction space is then discretized in horizontal planes. We choose the step δh which yields an image coordinate variation of the order of matching errors. The reconstruction is made in the reference camera geometry, which means that we estimate the height of the 3D point associated with each pixel of the reference image. This choice of representation is by far the most convenient in our experimental setup, although it implies that we do not reconstruct the points which are visible in other images than the reference one. But, as the viewpoint is not very different all along the sequence, this is not a severe limitation. Besides, we use a criterion symmetrical in all images (Collins, 1996).

2.2 Synthetic sequence

To illustrate our choices we use a synthetic sequence of 61 images obtained using a numeric elevation model covered with real textures and a simple sensor model (sensor transfer function and additive Gaussian noise): the reference image of the sequence is shown in Figure 3. the geometrical parameters are the following: $D_0 = 1.85$km, $H_0 = 630$m, $B = 800$m, $\theta_0 \approx 20$deg, field of view approx. 16 deg.

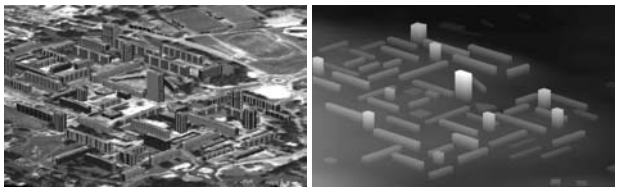

Figure 3: Central image of the synthetic sequence and the corresponding true height map

subsectionOverview of the proposed method

We recall that we intend to estimate the heights of each point visible in the reference view. For instance, in the synthetic case, we have to estimate the true height map shown on the right side of Figure 3. Our method is briefly sketched hereafter and the details are given in the following sections.

1. track interest point all over the sequence
2. refine the image motion parameters with bundle adjustment
3. use the reconstructed sparse 3D structure to deduce the world discretization range
4. compute the similarity cube: for each reference image pixel $\boldsymbol{x} = [x, y]^{\mathrm{t}}$ and each height hypothesis h compute a similarity measure using all the frames of the sequence
5. compute a regularized height surface in two steps:
   - obtain a "raw" reconstruction using L1 norm regularization
   - refine the surface via local optimization.

3 FEATURE TRACKING AND BUNDLE ADJUSTMENT

As many authors in computer vision (Schmid et al., 2000), we rely on Harris-like interest points, tracked from frame to frame using the "Kanade-Lukas-Tomasi" tracker downloaded from S. Birchfield web site (Birchfield, 1999).

Let us briefly recall that bundle adjustment (Wong, 1994) is a joint refinement of camera motion parameters and 3D position of some scene points, in order to minimize a compound criterion. The first term of the criterion is an image error term between reprojection of scene points and tracked features positions. The second term is a motion error between sought camera parameters and the parameters originating from the ego motion sensor of the plane (GPS/IMU). The image errors are weighted by the inverse of the matching error variance, which has been fixed to 0.2^2 (pixel^2) in all our (real and synthetic) examples, after an empirical evaluation of the performances of the KLT tracker.

Here bundle adjustment is conducted via a sparse Newton-like optimization (we use our own Matlab implementation) using the information available from the carrier ego motion sensor as initialization value of the motion parameters. Note that in order to deal with the possible outliers, the optimization is made twice, first with all the tracked features and second, after a simple outliers rejection step (Szeliski and Kang, 1993).

Let us emphasize that we do not use any ground calibration point: the 3D structure parameters are initially zero and there is no 3D error term in the criterion. The structure/motion ambiguity (choice of the coordinate system) is eliminated by using the reference camera coordinate system. As we do not use ground calibration point, the whole process is unsupervised.

At the end, we have a refined motion parameter set and a sparse 3D structure. The structure is used to define the range of our world discretization (the choice of the height step has been discussed in section 2), while the refined motion parameters will permit a precise computation of the matching criterion (2).

4 A MULTI-FRAME DENSE MATCHING METHOD

The matching process is based on the calculation of a radiometric criterion combining the K frames of the sequence with the strong hypothesis that 3D points are visible in any image (we will consider occlusions in section 5).

4.1 Similarity criterion

As the camera calibration is supposed to be precisely known (thanks to the bundle adjustment cf section 3), for a reference image pixel $\boldsymbol{x}$ and a height hypothesis h, the homologous pixel position $\boldsymbol{x}_k$ in frame k can be obtained by a planar homography. Generally the projected pixel does not fall on the image grid; a linear interpolation is used to retrieve a gray level, denoted $I_k(\boldsymbol{x}_k)$. Considering all the frames, we can 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 similarity criterion $C(\boldsymbol{x},h)$ is given by the standard deviation of this vector.

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

This choice is very close to (Géraud et al., 1998) and is also a pixelwise version of the multi-image similarity score of (Paparoditis et al., 2001). All these criteria rely on the conservation of the intensity among different frames. The extension to homogeneous illumination variations can be partly done by histogram equalization, but non homogeneous effects (such as specular reflection on some building window) cannot be dealt with.

As previously noted by various authors, the use of many frames allows a reduction of the traditional correlation window size (for instance (Paparoditis et al., 2001) use 3×3 or 5×5 windows) in matching algorithms. Here we simply use a pixelwise criterion and rely on regularization to increase spatial homogeneity of the result. Our results (even on real sequences, see section 6) show that such an approach allows noise smoothing and height discontinuity preservation.

4.2 L1-norm regularization

We minimize the penalized similarity criterion

$$J_1(h) = \sum_{\boldsymbol{x}} C(\boldsymbol{x},h(\boldsymbol{x})) + \lambda \sum_{\boldsymbol{x}'\in V_4(\boldsymbol{x})} |h(\boldsymbol{x}) - h(\boldsymbol{x}')| \quad (2)$$

where $V_4(\boldsymbol{x})$ is the 4-connexity neighborhood of pixel $\boldsymbol{x}$ and λ is the regularization parameter. The choice of a L1-norm regularization term avoid the excessive penalization of large height discontinuities (as it occurs with the more usual L2 norm). Note, however, that such a regularization is rather crude: it constraints the solution to be piecewise flat. We will devote the next section to a refinement of the solution.

The resulting optimization problem is huge and highly non convex. We use our own implementation of an efficient algorithm based on graph cuts (Roy and Cox, 1998). These recent algorithms have the remarkable ability to give in a polynomial time the exact optimum $\hat{h}_1$ of (2) over discretized height levels, even if the first term is non convex. Their only drawback is the large required memory space: in our examples (around $300 \times 500 \times 41$ similarity cube) approximately 270 Mo are used.

The resulting estimated height map using the synthetic sequence is shown in Figure 4 with an ad-hoc regularizing parameter λ fixed at 1.5.

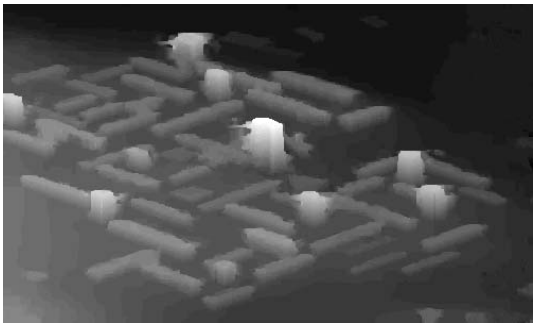

Figure 4: Synthetic data: estimated height map using similarity criterion (1) and L1 regularization ($\lambda = 1.5$)

We observe a good compromise between noise smoothing and discontinuity preservation. The quantitative improvement on height estimation (with respect to minimum similarity reconstruction) is dramatic, as shown in the third and fourth lines of Table 1. Some errors occur in the vicinity of tall buildings because the similarity measure is not robust with respect to occlusions (we will discuss this problem in section 5). In addition, some piecewise flat errors appear, see left bottom corner of Figure 4.

4.3 Local refinement

The piecewise flat errors already mentioned are a typical side effect of the L1-norm regularization over first order differences. It could be corrected using a L2-L1 regularization, ie. penalty function which are quadratic near zero and linear further (Idier, 2001). Unfortunately, the graph cut algorithm for L2-L1 regularization uses a larger workspace, which is rapidly unrealistic (Ishikawa, 2000, Paris and Sillion, 2002). An even better solution would use penalization over second order differences of the height map, but it cannot be optimized by graph cut algorithms. We are then led to suboptimal solutions. We propose a local refinement of the L1-norm optimal height map $\hat{h}_1$ whose large discontinuities do not be called into question.

We therefore search for a better solution in a neighborhood of $\hat{h}_1$ in the sense of a second order L2-L1 regularization, ie. by minimizing the criterion

$$\begin{aligned} J_2(h) = &\sum\nolimits_{\boldsymbol{x}} C(\boldsymbol{x},h(\boldsymbol{x})) + \\ &\lambda \sum\nolimits_{\boldsymbol{x}} \{\phi(h(x-1,y) - 2h(x,y) + h(x+1,y)) + \\ &\phi(h(x,y-1) - 2h(x,y) + h(x,y-1)) + \\ &\phi(h(x-1,y-1) - 2h(x,y) + h(x+1,y+1)) + \\ &\phi(h(x-1,y+1) - 2h(x,y) + h(x+1,y-1))\} \end{aligned}$$

with $\phi(t) = 2t - \log(1 + |t|/2)$ (Idier, 2001). Continuous derivative of the similarities are interpolated from a second finer similarity cube computed with a finer discretization step and for heights limited to a small range around $\hat{h}_1$. Figure 5 shows comparisons between profiles of the true map, the L1-norm estimation and the refined estimation: improvements are noticeable and are confirmed by quantitative results, see the fifth line of Table 1.

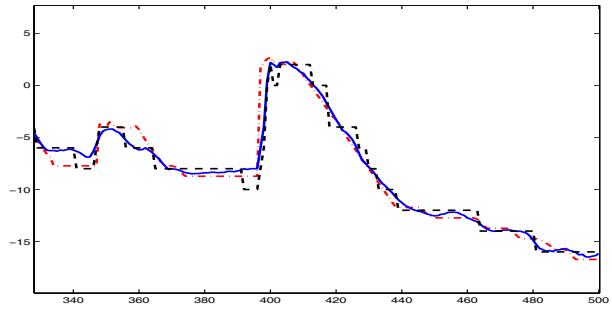

Figure 5: Synthetic data: profiles of the true height map (red and dash-datted) , the L1-norm regularized map (black and dashed) and refined L2-L1 estimation (plain blue) — $\lambda = 1.5$ and 50 steps of projected gradient descent

5 DEALING WITH OCCLUSIONS

The main problems in the reconstructed maps shown in section 4 are due to erroneous decisions in the occluded areas around the tallest buildings. Occlusions are correctly handled in the two-views situation thanks to dynamic programming algorithms. However, in the many-frames situation it is a very intricate problem. Some methods use an explicit flag to detect occluded pixels, which lead to a very difficult joint estimation problem (height and occlusion flag) which is generally conducted in a sub-optimal way. Recently, S. B. Kang (Kang et al., 2001) has proposed a simple modification ("temporal selection") of the similarity criterion to account for occlusion, which leads to a solution with the same computational cost as the original algorithm. In this line, we propose a new solution, which corrects some flaws of the approach of (Kang et al., 2001).

When using the similarity criterion of section 4.1 one consider every pixel as non occluded, as this a criterion tends to minimize the similarity among *all* views. If a pixel is occluded in some of the views, the intensity level measured in those views should not be taken into account. Therefore, Kang minimizes the similarity criterion in only half of the available views and proposes two ways of "temporal selection" among the views. In our configuration, we simplify this choice as a binary one, by postulating that a point visible in the reference view will be occluded either in the "left part" of the sequence, ie. in the views preceding the reference, or in the "right part", made of the views taken after the reference. Using notations from section 4.1, we write Kang's criterion as

$$C^{\text{Kang}}(\boldsymbol{x}, h) = \min\{\hat{\sigma}(\boldsymbol{V}_1^r(\boldsymbol{x}, h)), \hat{\sigma}(\boldsymbol{V}_r^K(\boldsymbol{x}, h))\}. \quad (3)$$

As previously noted in (Kang et al., 2001), this simple modification improves noticeably the reconstruction around depth discontinuities, located in our examples near tall buildings. However, in this setting, all pixels are considered as "half-occluded" points, but we know that most of them are visible in the whole sequence. As a result, the reconstruction with criterion (3) is more corrupted by noise than (1) in the low areas. We have implemented another strategy, where a first decision is made about whether the point is half-occluded and a second about the left-occluded case or right-occluded case. The first decision is made by comparison between the standard deviations of $\boldsymbol{V}_1^r(\boldsymbol{x}, h)$ and $\boldsymbol{V}_r^K(\boldsymbol{x}, h)$: for a good candidate height, these quantities should be both approximately equal to the noise level. However, if the point is half-occluded, one of these is augmented by the difference of the intensity around the point between the the intensity of region and the region which occludes it. Therefore we threshold the standard deviation difference

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

and obtain the "mixed" criterion

$$C^{\text{Mixed}}(\boldsymbol{x}, h) = \begin{cases} C^{\text{Kang}}(\boldsymbol{x}, h) \text{ if } D_s > t \\ C(\boldsymbol{x}, h) \text{ otherwise} \end{cases} \quad (4)$$

Using an *ad hoc* threshold of $t = 15$ for a 8-bit image sequence, we obtain good results both around discontinuities and in low areas, as shown by Figure 6 compared with Figure 4. The first lines of Table 1 shows a quantitative comparison between the three criteria C, C^{Kang} and C^{Mixed} on the synthetic sequence. Kang's criterion decreases outliers percentage at the price of a degraded precision. Our strategy makes an interesting compromise between the two others method.

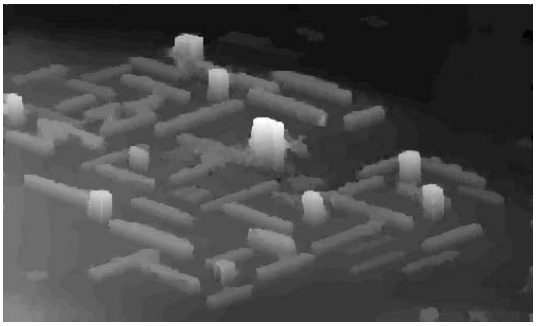

Figure 6: Synthetic data: estimated height map using "Mixed" similarity (4), L1 regularization ($\lambda = 1.5$) and refinement method described in section 4.3

similarity	bias	RMS	L1-norm	outliers
C of eq. (1)	0.14	1.08	0.73	3.35%
C^{Kang} of eq. (3)	−0.26	1.92	1.04	1.6%
C^{Mixed} of eq. (4):				
no regularization	0.15	9.09	5.66	28%
L1-norm	−0.03	1.4	0.76	2.1%
with refinement	0.00	1.27	0.57	2.05%

Table 1: Performance of various similarity criteria: statistics (in meters) on the best 90% error samples and percentage of outliers (defined as points with an error > 10m).

Moreover, our method allows an *a posteriori* visualization of the decisions made by the algorithm among the three status non-occluded, left-occluded or right-occluded of each pixel, shown in gray, white or black color in the Figure 7. We observe a very good agreement between the decisions and our knowledge of the 3D scene. Thanks to the regularization term the occluded part are indeed compact and localized around buildings, although probably slightly over-estimated.

6 EVALUATION ON A REAL AERIAL SEQUENCE

6.1 Real sequence

We present our results on a real aerial sequence ("City1") acquired during a mission around the same town which we have modeled in the synthetic sequence. The geometric parameters of the acquisition are the following. The distance to the center of the scene is approximatively 3 kilometers and the average altitude is 500m, hence the depression angle is of the order of 10 degrees.

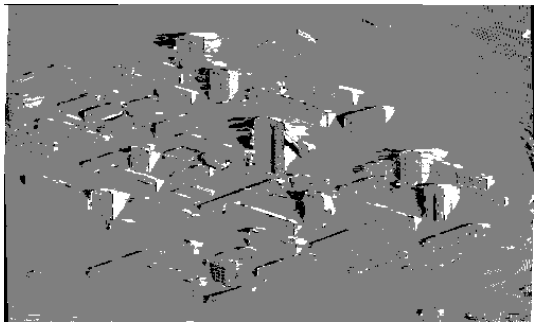

Figure 7: Synthetic data: *a posteriori* decisions about the visibility of each pixel (see text)

This low value means that masking effects and variation of precision in the field of view are important. The sequence, taken on a approx. 800m-long baseline, includes 51 8-bits images. The 26^{th} is the reference image, shown in Figure 8.

Figure 8: Reference image of sequence "City1" (see text)

6.2 Protocol of evaluation

To evaluate results on real sequences, we have used the data taken from the IGN's topographic database for the DTM and a numeric model for buildings. Some errors occurring on this model were corrected thanks to map of a part of the district. As our DSM estimation is camera dependent, we give results relating to a reference point located at the same time on topographic database and in the reference image.

6.3 Results and evaluation

Figure 9 shows the estimated height map coded in gray levels from -50m to 50m with a 2m step, above the reference plane chosen using the reference camera motion parameters. Table 2 shows quantitative comparison with IGN's topographic database. The main buildings are well located and their relative heights with respect to the ground level are correct. Their borders are precise, thanks to the L1-norm regularization. In the regions where buildings are smaller and closer to each other, the masking effects are important (recall that the depression angle is less than 10 degrees) making the segmentation of super-structures less effective. We took test points in these region and their height is correctly estimated, see Table 2. In the upper right part of the reconstruction, which corresponds to farther areas, the planimetric resolution is low and we estimate only the large scale relief: the diagonally oriented hill and valley are confirmed by the database examination. The decision map shown in Figure 10 shows many false decisions, probably due to the low depression angle. The left and right occlusions tends to be correctly located with respect to the main tall structures.

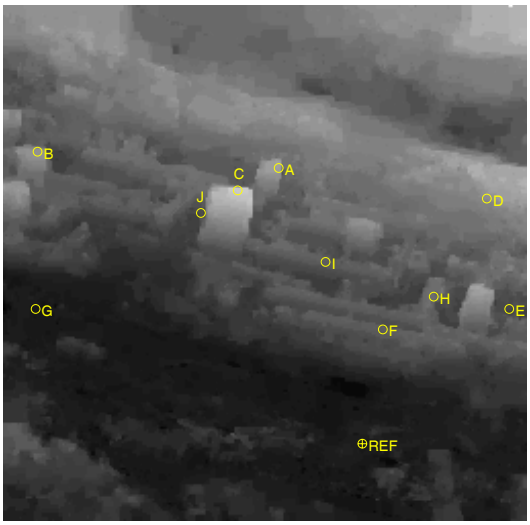

Figure 9: City1 sequence: Estimated height map. Circles represent our test points and crossed circle is the reference point (see text)

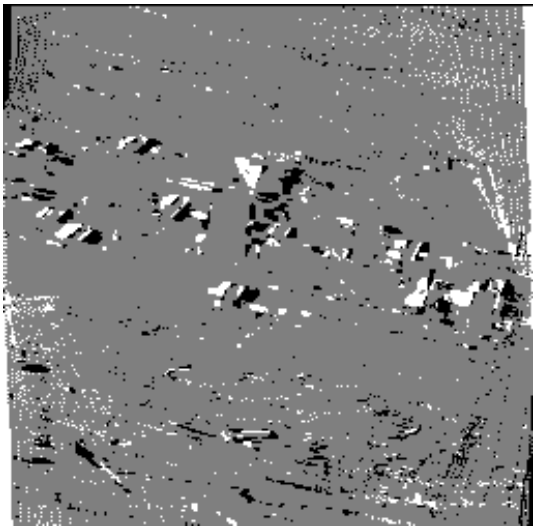

Figure 10: City1 sequence: Choice of visibility for each pixel: non occluded (gray), left-occluded (black) and right-occluded (white)

7 CONCLUDING REMARKS

We have described a novel method to reconstruct partial DSM from aerial side-looking image sequences, which has been evaluated both on synthetic and real sequences. Occlusions are accounted for in a simple way which gives interesting results without any supplementary cost. Further work will concern fusion of height map retrieved from very different viewpoints (for instance opposite).

Finally, we would like to emphasize the importance of synthetic

Point	IGN Database (m)	Estimated height (m)
A	60	48
B	45	44
C	69	66
D	45	40
E	16	14
F	28	26
G	2	-2
H	31	30
I	31	26
J	20	20

Table 2: Estimated heights on the test points shown in Fig. 9 and corresponding heights from the IGN topographic database (meters)

data which have been constantly employed throughout our work. They allow control of various geometrical/sensor parameters and make it possible to conduct precise comparison and evaluation of different methods. Of course, the final step should be evaluation on real data, which we have only began in this paper: further evaluations are under progress.

Acknowledgments

The authors wish to thank Hélène Oriot and Marie-Lise Duplaquet for fruitful discussions and suggestions.

REFERENCES

Birchfield, S., 1999. Download site for the KLT Tracker. http://vision.standford.edu/ birch/klt/.

Collins, R., 1996. A space-sweep approach to true multi-image matching. In: ARPA/IUW.

Géraud, B., Le Besnerais, G. and Foulon, G., 1998. Determination of dense depth map from an image sequence: application to aerial imagery. In: European Symposium on Remote Sensing, Image and Signal processing for Remote sensing IV.

Idier, J., 2001. Approche bayésienne pour les problèmes inverses. Traité IC2, Hermès, Paris.

Ishikawa, H., 2000. Optimization using embedded graphs. PhD thesis, New York University.

Kang, S. B., Szeliski, R. and J., C., 2001. Handling occlusions in dense multi-view stereo. In: CVPR'01, Vol. 1, pp. 103–110.

Le Besnerais, G. and Duplaquet, M.-L., 2002. Traitement de séquences d'images aériennes en visée latérale pour la reconstruction 3d. Bulletin de la SFPT (166), pp. 27–33.

Mandelbaum, R., Salgian, G. and Sawhney, H., 1999. Correlation-based estimation of ego-motion and structure from motion and stereo. In: ICCV'99, IEEE.

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

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

Roy, S. and Cox, I., 1998. A maximum-flow formulation of the n-camera correspondence problem. In: ICCV'98.

Schmid, S., Mohr, R. and Bauckahge, C., 2000. Evaluation of interest point detectors. International Journal of Computer Vision 37(2), pp. 151–172.

Szeliski and Kang, 1993. Recovering 3d shape and motion from image streams using non linear least squares. Technical report, Cambridge research laboratory, Digital equipment corporation.

Wong, K., 1994. Basic mathematics of photogrammetry. In: C. Slama (ed.), Manual of Photogrammetry, 4th edn, American Society for Photogrammetry and Remote Sensing.