D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

# DENSE URBAN DEM WITH THREE OR MORE HIGH-RESOLUTION AERIAL IMAGES

Uğur M. Leloğlu*†, Michel Roux‡, Henri Maître‡

† TÜBİTAK-BİLTEN
METU, 06531 Ankara, Turkey
Ph.: +90 312 210 46 13, Fax: +90 312 210 13 15
E-mail: lel@tbtk.metu.edu.tr

‡ Signal and Image Department, ENST
46 rue Barrault, 75013 Paris, France
Ph.: +33 1 45 81 81 28, Fax: +33 1 45 81 37 94
E-mail: mroux,maitre@ima.enst.fr

**KEY WORDS:** stereo, 3-D reconstruction, DEM, image matching, high-resolution aerial imagery

## ABSTRACT

In cartographic applications, area-based matching techniques are commonly used for stereo matching of low-resolution aerial images. However, these techniques fail in matching high-resolution aerial images of urban areas because of relatively frequent and sharp depth discontinuities and large occluded or textureless areas present in such images, as compared to low-resolution aerial images.
This paper presents a hierarchical correlation-based matching technique which fuses information from multiple image pairs and employs a support-collection mechanism in the object space as well as a relaxation algorithm to resolve ambiguities, producing accurate and dense disparity maps. The disambiguating power of the algorithm and the use of multiple pairs allow us to use very small correlation windows, so that the computational complexity is kept small and boundary overreach problem is avoided. They also allow us not to use any threshold on correlation values; as a result, very dense disparity maps can be obtained.

## 1 INTRODUCTION

Obtaining digital elevation models (DEM) from aerial images is useful for a number of cartographic applications. The use of correlation-based stereo in establishing DEMs is common and well-studied. Such techniques are known to be successful in low-resolution images or in images of non-urban areas where the depth changes smoothly and where there exists rich texture.

In this paper, we address correlation-based stereo correspondence in the domain of high-resolution aerial images of urban areas, that typically contain large textureless regions (e.g. roads and especially roofs which are of great importance), frequent sharp depth discontinuities, and, large occlusions. The images on which we develop and test the method presented here are 24-bit RGB aerial images of West European industrial or urban zones with a ground resolution of 8cm. The internal camera parameters are readily available and external parameters can easily be determined using standard calibration techniques.

This paper is organised as follows: In the following section, some related work is summarised. In section 3, a hierarchical relaxation algorithm, which calculates the disparity maps from multiple image pairs simultaneously, is described. In section 4, some experimental results are presented and, finally, the paper is concluded with a discussion on results in section 5.

## 2 RELATED WORK

One of the ways to use more than two images in stereo reconstruction is to construct epipolar image pairs to obtain disparity maps with conventional stereo techniques, and then, to merge the resulting matches in the object space. There is a rich literature towards the integration of depth data from multiple sources, not necessarily from stereo, but from shape-from-shading (Ferrie and Levine, 1987) or range images (Shum et al., 1994)(Higuchi et al., 1993). An interesting work to merge disparity maps resulting from multiple stereo pairs is that of (Fua, 1997) where small patches (“oriented particles”) are fitted to matches in 3-D object space to estimate underlying surface.

Another way of using three or more images is to employ a correlation-like similarity measure defined over all images involved. In multibaseline stereo, the pixels in each image, corresponding to a given pixel in the reference image and a given depth, can easily be found. The sum of squared differences (SSD) within a window around those pixels, which was first used by (Okutomi and Kanade, 1993) in this context, can be drawn as a function of depth. (Kang and Szeliski, 1997) use SSD in panoramic images; (Park and Inoue, 1997) use only two median of four differences obtained from five cameras to overcome the problem of occlusion; (Scharstein and Szeliski, 1996) use an adaptive support region instead of a square window; and (Canu et al., 1995) use sum of normalised correlations instead of SSD.

A third way is to project all possible matches from multiple pairs to 3-D, and then, to choose the true matches in object space. (Zitnick and Webb, 1996) project matches from multiple cameras with respect to a reference camera to 3-D and eliminate some of the false matches by tracking each match, in all pairs, in increasing order of baseline distance. So, a point can be matched only when it can be seen from all cameras. The remaining 3-D points are grouped into continuous surfaces, considering their depth differences in 3-D and their pixel distance in 2-D. The most numerous groups are assumed to correspond to true surfaces.

## 3 DESCRIPTION OF THE ALGORITHM

In the case of merging disparity maps from multiple stereo image pairs, one does not benefit from the information in three or more images during the matching process. But, some false matches could be eliminated or more matches could be obtained in that early phase. The use of correlation-like measures defined on three or more images are more powerful in that sense, however, a match which is very clear in one pair of images (i.e., a very sharp and large peak in the correlation signal) can be lost because of noise, occlusion or high disparity gradient. Besides, when all cameras are

* Supported by TÜBİTAK (The Scientific and Technical Research Council of Turkey) and by IMPACT: Esprit project 20243

not on the same baseline, it is difficult to calculate such similarity measures. For these reasons, we choose to calculate the correlations pairwise and mark all maxima as candidates. Then, unlike (Zitnick and Webb, 1996), through a mechanism of support collection from candidates from all pairs, true matches are selected. In our method, a point which can not be seen from all views, or surfaces which have small projections on the images can also be matched.

In our image database, the images are taken from arbitrary points of view, and therefore, it is not possible to bring three or more images into a unique epipolar geometry. So the images are aligned pairwise. All points from all image pairs which are identified as peaks with a positive value in the correlation space are accepted as candidate matches and projected into the object space where a support value for each candidate match is calculated. A confidence value in the range [-1,1] which is equal to the correlation value at the initialisation, is also kept with each match. Using a relaxation algorithm some matches are accepted as true matches and some are rejected. By this way, we never create an arbitrary dominant eye, i.e., a reference camera, throughout the stereo reconstruction process.

### 3.1 Hierarchy

The image pyramid on which we apply the hierarchical algorithm is obtained by low-pass filtering followed by sub-sampling, as usual. The matches obtained at a certain level constrain the solution in the next finer level. This coarse-to-fine approach, which is commonly used in computer vision problems, has one main disadvantage: an error at a level is spread up to all finer levels. In stereo case, usual practice is to limit the possible disparity range of a pixel, corresponding to a match from the lower level, around the disparity value of that match. In this case, if the error is very large in the lower level, we can never reach the true disparity in the finest level. To avoid this problem, the following method is used for constraining the solution in the upper level: for each match from the lower level a sphere around this match is included in our search volume. The intersection of all these spheres defines the volume in which we search for matches. Therefore, if a true match has an accepted match in his 3D-neighbourhood in the lower level, it is considered as a candidate.

### 3.2 Algorithm

In Figure 1, the algorithm is depicted schematically at an arbitrary level of the hierarchy. In principle, any number of pairs obtained from any number of images can be used. For the sake of simplicity, the figure shows only two pairs obtained from three images. First, the correlations across points from epipolar pairs, which correspond to the search space defined by the lower level and by the possible disparity range, are calculated. For the lowest level, only the disparity range is used. The correlations which are local maxima and which have a positive value are considered as candidate matches. Then, all the matches are projected to the object space so that a support value, $s_i$, can be calculated. There, by the relaxation algorithm, some of the matches are accepted and some are discarded. Each match, either candidate or accepted, is defined by a 5-tuple $P_i \overset{def}{=} (x_i, y_i, z_i, c_i, l_i)$ where $x$, $y$ and $z$ denote the 3-D coordinates of the match in the object space, $c$ and $l$ denote the confidence value and the label of the match, respectively. Confidence value is defined simply as the normalised correlation value (in the range $[-1, 1]$) for candidates and as unity for accepted matches. The label of the match is its origin, i.e., the index of the pair it is generated from.

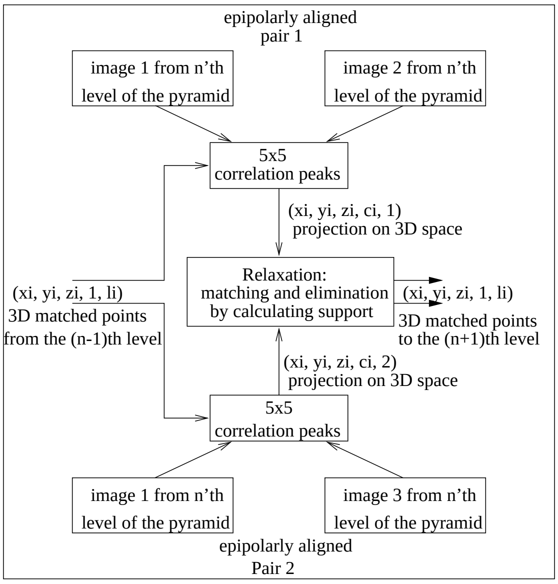
epipolarly aligned
pair 1
image 1 from n'th level of the pyramid
image 2 from n'th level of the pyramid
5x5 correlation peaks
(xi, yi, zi, ci, 1) projection on 3D space
Relaxation: matching and elimination by calculating support
(xi, yi, zi, 1, li) 3D matched points from the (n-1)th level
(xi, yi, zi, 1, li) 3D matched points to the (n+1)th level
(xi, yi, zi, ci, 2) projection on 3D space
5x5 correlation peaks
image 1 from n'th level of the pyramid
image 3 from n'th level of the pyramid
epipolarly aligned
Pair 2

Figure 1: The Algorithm in an Intermediate Level.

### 3.3 Relaxation

The relaxation algorithm to choose true matches among $P_i$, $i = 1, \ldots, N$ at any level is as follows:

```
set $t_{high}$ to a very high value
set $t_{low}$ to a very low value
do
    for i=1 to $N$
        calculate $s_i$, support for $P_i$
        If $s_i > t_{high}$
          label the match as accepted
          $c_i := 1$
          reject any candidate which is from the same
                epipolar image pair and
                which violates the ordering constraint
        If $s_i < t_{low}$
          reject $P_i$ from the pool of matches
    decrease $t_{high}$
    increase $t_{low}$
while ( $t_{high} > t_{low}$ )
end
```

During this relaxation, ambiguities are expected to be resolved in several ways:

- Since accepted matches vote more strongly for their neighbours, the support values of good candidates are likely to increase.
- Accepted matches cause some of the false candidates to be discarded via the ordering constraint. False matches with very small support are also rejected. As a result, the support value of remaining false matches have a tendency to decrease.

### 3.4 Correlation

In correlation-based methods, the size of the correlation window is the result of an important trade-off. When the disparity gradient is small and there is no depth discontinuity within the window, large windows perform well with the price of high computational complexity. But, in our case, there exist sharp discontinuities which cause boundary overreach and we may end up with areas (e.g. roofs) significantly larger than their real size. To avoid boundary overreach and other problems caused by depth discontinuities and to reduce the computational complexity considerably, we have

D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

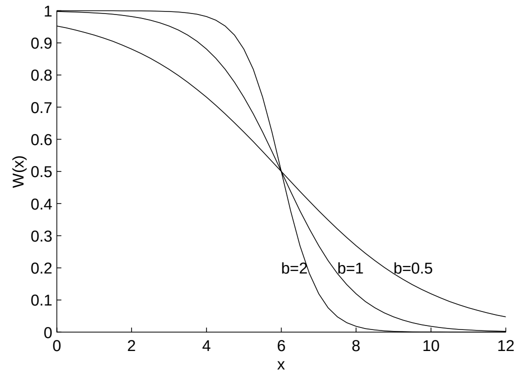

Figure 2: $W(x)$ for three values of $b$ where $r = 6$.

used normalised correlations of size 5×5. Using windows of such small size, in turn, makes the process sensitive to noise, but our algorithm can eliminate false matches caused by noise. We also benefit from the colour information in 3-band images by using the median of red, green and blue band correlations for each pixel as suggested by (Roux et al., 1997).

### 3.5 Support Function

The use of a support function is known to be powerful in resolving ambiguities in stereo matching. All cooperative algorithms use a function of the neighbourhood to encourage or discourage a match. A support function is defined so as to allow only positive contributions from the neighbours. (Prazdny, 1985) states that dissimilar matches should not inhibit each other “because they potentially carry information about different surfaces”. Since we want to cope with surfaces at different heights, we have chosen the same approach. Prazdny uses Gaussian distribution function to calculate the support to a match from the neighbourhood. This support function implicitly implies a disparity gradient limit and favours frontoparallel surfaces. Our support function is defined in a different manner due to the nature of the problem.

In the choice of the support function one should consider the fact that the points in 3-D are viewed from almost arbitrary angles in each image and in the urban area there exist surfaces which are vertical as well as horizontal. Therefore, instead of a function which favours horizontal surfaces, we have chosen a spherically symmetrical support function such that the support to $i$th candidate is

$$s_i = \sum_{k=1}^{N} c_k W\Big(\sqrt{(x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2}\Big)$$

where $N$ is the total number of matches (either candidate or accepted) and $W(\rho)$ is a monotonically decreasing function of $\rho$ which determines the radius of the effective neighbourhood and defined by the sigmoid function

$$W(\rho) = \frac{1}{1 + e^{b(\rho - r)}}.$$

Here, $r$ is the radius of the effective neighbourhood which votes for the match and $b$ determines how fast the transition is (see Figure 2).

### 3.6 Simplifying Assumptions

In practice, following simplifying assumptions are made because of the limitations on memory and computation time:

*Limit on Number of Peaks:* We do not keep all the peaks in the correlation space but only the two greatest peaks in the range of possible disparities for each pixel in each image. This is a reasonable assumption, because, since the disparity search space is constrained by the lower level, there exist hardly two peaks in that range. If the true disparity is, although very unlikely, at the third or later peak, we miss it.

*3-D Distance Approximation:* Keeping the real-world coordinates of each candidate match consumes too much memory. Instead, we try to estimate the Euclidean distance between candidates using directly their image coordinates and disparities. Let two matches from the same pair be $(i_1, j_1, d_1)$ and $(i_2, j_2, d_2)$ where $i$ and $j$ are row and column numbers for the first image of the pair, respectively, and $d$ is the disparity. We want to express the distance between these two matches in terms of $(i_1 - i_2)$, $(j_1 - j_2)$ and $(d_1 - d_2)$ and keep this function in a look-up table of three variables. After a series of geometric calculations and assumptions (see appendix A), we reach the following simple formula:

$$d_{1,2} = s_h\sqrt{(i_2 - i_1)^2 + (j_2 - j_1)^2 + (j_2 - j_1)(d_2 - d_1) + a^2(d_2 - d_1)^2}$$

where $s_h$ and $a^2$ are determined by the geometry. Note that the function is symmetric with respect to images in a pair, i.e., the distance between the corresponding matches $(i_1, j_1 + d_1, -d_1)$ and $(i_2, j_2 + d_2, -d_2)$ in the other image of the pair is the same as that of the first one. Thus, a dominant eye is not created. Although the algorithm is not sensitive to small variations in the distance function, the formula is quite accurate (See figure 3). Besides, by changing the parameter $a$, one can favour vertical or horizontal surfaces. In order to calculate the distance between matches from different pairs, one of the matches is transformed to the coordinate system of the other, since internal and external camera parameters and the transformations for the epipolar alignment are all known.

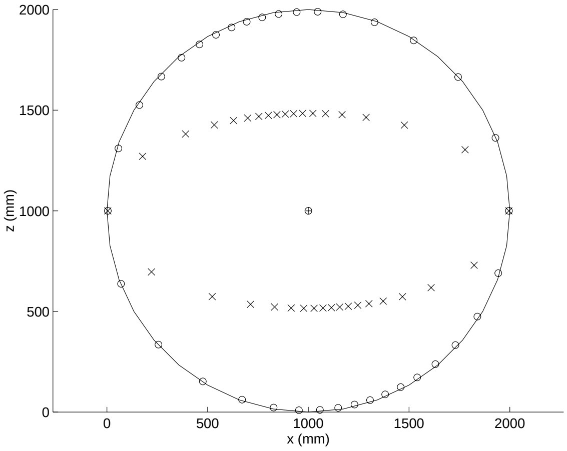

Figure 3: *Solid line:* A circle of diameter 1000 mm in the object space. *Circles:* A group of points which are at 1000 mm distance to the centre according to the approximate formula. *Crosses:* The same points when the parameter $a$ is multiplied by $2$. The parameters are those calculated for the second image pair for which the results are presented in section 4. The central point corresponds to an arbitrary match.

*Support from only 2-D Neighbourhood:* To avoid the computationally expensive process of calculating the support from all $N$ matches, the support calculation is constrained to the matches at the pixel locations within a rectangular 2-D neighbourhood of the matching pixel large enough to cover all the matches in the effective neighbourhood.

### 3.7 Computational Complexity

Since the number of candidate matches at any pixel is limited to two, $N$ is proportional to the size of the image. At each iteration, the support value is calculated $N$ times. The number of matches in the search neighbourhood is independent from $N$, so the computational complexity of the relaxation algorithm is proportional to $N$, hence, to the image size.

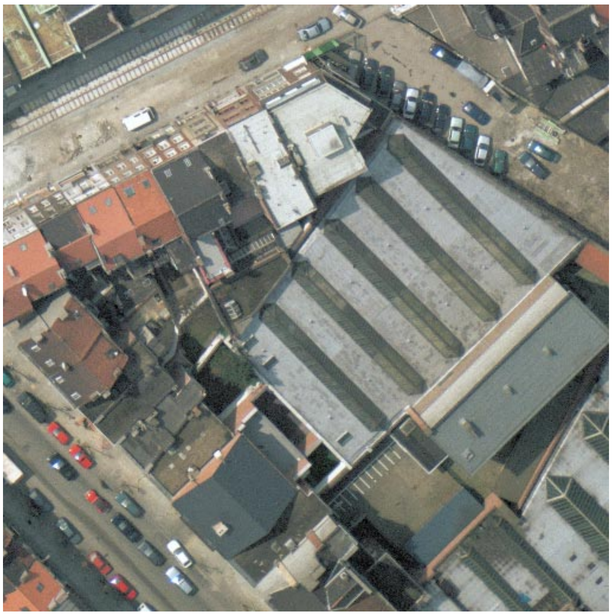

Figure 4: One of the three views of the same area (1024×1024). Courtesy of Eurosense.

## 4 RESULTS

The algorithm described is tested on high-resolution (8cm×8cm) aerial colour images. Figure 4 depicts one of the three images of the same area on which we test our algorithm and Figure 5 shows the result of 9×9 correlation followed by thresholding on one of the stereo pairs, in a simple three level hierarchical architecture.

Figure 6 displays the final result of the new algorithm with three levels and six iterations over the main loop where two pairs were used and Figure 7 is the 3-D rendering of this result. Table 1 and Table 2 present the schedule used for relaxation and the number of candidates before and after relaxation, respectively. The radius of sphere, $r$, equals $500mm$ and the sigmoid parameter, $b$, equals $(1/40)mm^{-1}$ in the finest level. The same schedule is used at all levels.

| iteration | $t_{max}$ (%) | $t_{min}$ (%) |
|---|---|---|
| 1 | 55 | 4 |
| 2 | 45 | 5 |
| 3 | 35 | 6 |
| 4 | 25 | 8 |
| 5 | 15 | 10 |
| 6 | 12 | 12 |

Table 1: The schedule for relaxation. The threshold values, which are chosen such that they converge at the last iteration, are given as percentages of the maximum possible support value.

| level | size | total number of candidates | number of accepted candidates | % |
|---|---|---|---|---|
| 1 | 256x256 | 125038 | 48582 | 74.1 |
| 2 | 512x512 | 486014 | 177109 | 67.6 |
| 3 | 1024x1024 | 1916831 | 775984 | 74.0 |

Table 2: The number of candidates before and after relaxation belonging to the second pair, and ratio of matched pixels to the total number of pixels, as a function of the level in the pyramid.

To demonstrate the disambiguating power of relaxation, we show, in Figure 9, the results of the algorithm without relaxation step, i.e., the results obtained by thresholding the support values resulting from two pairs. Note that the uniqueness and ordering constraints are sometimes violated in this disparity map.

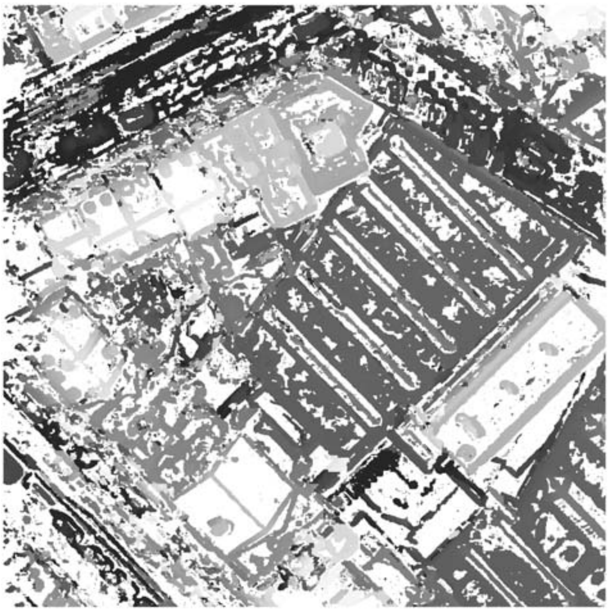

Figure 5: Disparity map corresponding to Figure 4 obtained by 9×9 correlation with three levels of hierarchy and with a threshold of $0.45$ using two views only (i.e., one pair).

Figure 8 is the disparity map obtained by using the proposed algorithm with only one pair of images. Increasing the number of pairs results in denser and more accurate maps.

## 5 CONCLUSION

In this work, we have proposed an algorithm which combines the information from multiple image pairs and which eliminates false matches in correlation-based stereo matching. We have demonstrated the efficiency of the algorithm qualitatively on real images. Since no threshold is used on the correlation values and since a match appearing in any pair has a chance to survive, very dense disparity maps are obtained. By means of support calculation from multiple pairs in object space, the fusion is realised in an intuitive way. Combining support collection with relaxation resulted in a robust algorithm producing accurate matches. The algorithm can be extended in several directions: an estimation of the surface direction can be used to modify the shape of the support function, the colour of the matching pixels may contribute to the calculation of the support, or matches between any kind of features (e.g. corners, segments) can be projected on the same space and can collect support from the matches.

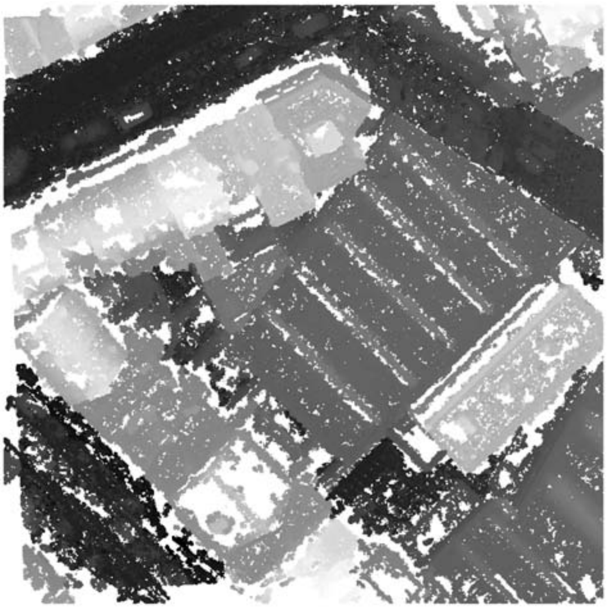

Figure 6: Final result of the proposed algorithm (74% of the pixels are assigned a disparity value).

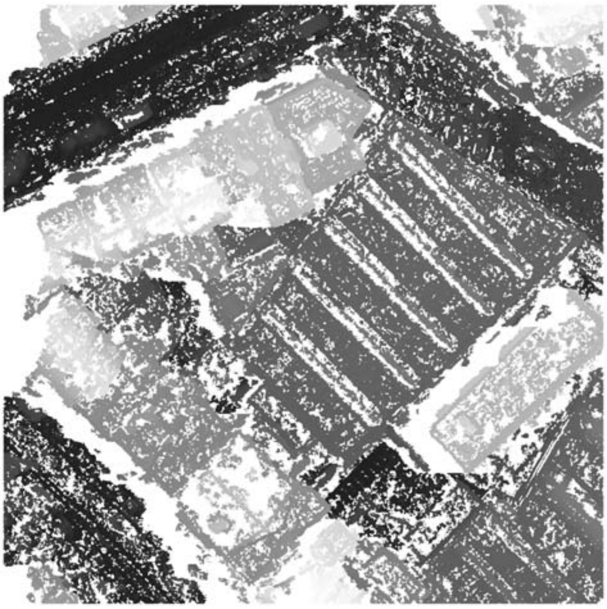

Figure 8: Disparity map with two views only.

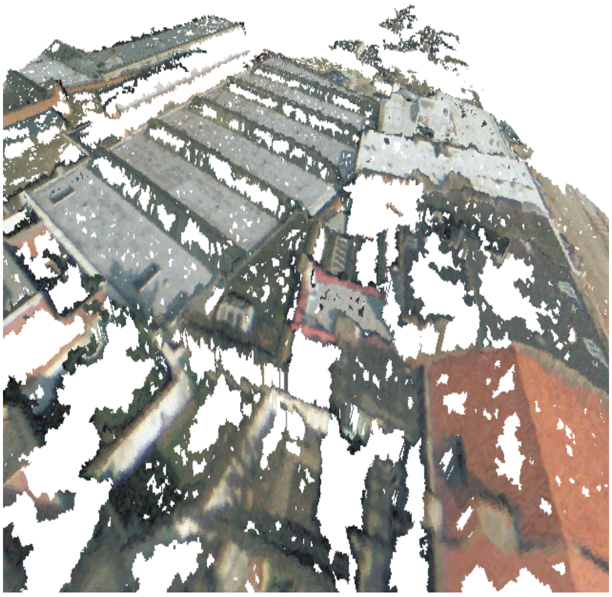

Figure 7: 3D rendering of the final result.

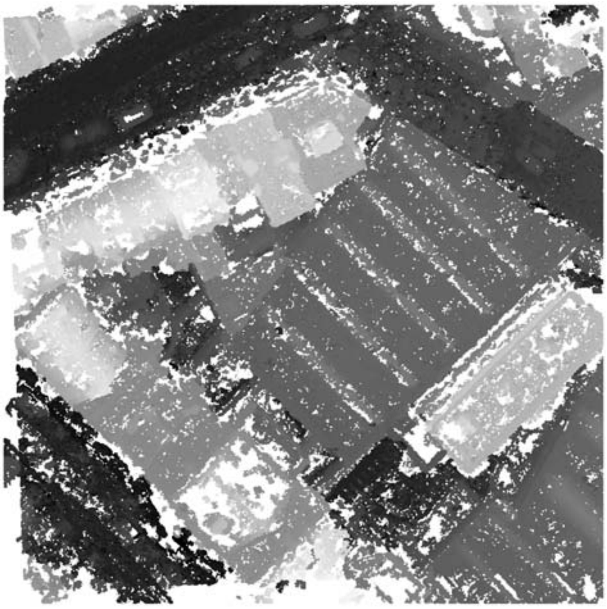

Figure 9: Disparity map without relaxation, but with two pairs.

## REFERENCES

Canu, D., Ayache, N. and Sirat, J. A., 1995. Accurate and robust stereovision with a number of aerial images. In: SPIE European Symposium on Satellite Remote Sensing II, Paris, France, pp. 1–9.

Ferrie, F. P. and Levine, M. D., 1987. Integrating information from multiple views. In: Workshop on Computer Vision, Representation and Control, IEEE Computer Society, Miami Beach, FL, USA, pp. 117–122.

Fua, P., 1997. From multiple stereo views to multiple 3-d surfaces. International Journal of Computer Vision 24(1), pp. 19–35.

Higuchi, K., Hebert, M. and Ikeuchi, K., 1993. Building 3-d models from unregistered range images. Technical Report CMU-CS-93-214, Carnegie Mellon University, Pittsburgh, PA, USA.

Kang, S. B. and Szeliski, R., 1997. 3-d scene data recovery using omnidirectional multibaseline stereo. International Journal of Computer Vision 25(2), pp. 167–183.

Okutomi, M. and Kanade, T., 1993. A multiple-baseline stereo. IEEE PAMI 15(4), pp. 353–363.

Park, J. and Inoue, S., 1997. Hierarchical depth mapping from multiple cameras. In: ICIAP97, Vol. 1, University of Florence, Florence, Italy, pp. 685–692.

Prazdny, K., 1985. Detection of binocular disparities. Biological Cybernetics 52, pp. 93–99.

Roux, M., Maître, H. and Girard, S., 1997. A step towards stereo reconstruction of urban aerial images. In: 3D Reconstruction and Modelling of Topographic Objects, ISPRS, Stuttgart, Germany, pp. 107–114.

Scharstein, D. and Szeliski, R., 1996. Stereo matching with nonlinear diffusion. In: CVPR'96, IEEE Computer Society, San Francisco, CA, USA, pp. 343–350.

Shum, H. Y., Ikeuchi, K. and Reddy, R., 1994. Principal component analysis with missing data and its application to object modelling. In: CVPR'94, IEEE Computer Society, Seattle, Washington, USA, pp. 560–565.

Zitnick, C. L. and Webb, J. A., 1996. Multi-baseline stereo using surface extraction. Technical Report CMU-CS-96-196, Carnegie Mellon University, Pittsburgh, PA, USA.

## A 3D DISTANCE APPROXIMATION

Consider the epipolar camera geometry shown in Figure 10.

The coordinates of object point, in terms of $(X_{left}, Z_{left})$, are:

$$x = -\frac{kx_l}{(x_r - x_l)} \qquad z = \frac{kf}{(x_r - x_l)}.$$

Since $x_l = aj$ and $x_r = a(j + d)$ (where $j$ is the column number in the left image, $a$ is the pixel size and $d$ is the disparity),

$$x(d,j) = -\frac{kj}{d} \qquad z(d,j) = \frac{kf}{ad}.$$

We assume that $d \in [d_0 - r, d_0 + r]$, i.e., all objects are in a certain depth range. We open $x(d,j)$ to a Taylor series around $(d_0, j_0)$ where $j_0 = -d_0/2$ (or equivalently, $x = k/2$) and ignore 2nd or higher order terms.

$$\begin{aligned} x(d,j) \simeq x(d_0,j_0) \quad &+ (d - d_0)\tfrac{\partial x}{\partial d}(d,j)\Big|_{\substack{d = d_0 \\ j = j_0}} \\ &+ (j - j_0)\tfrac{\partial x}{\partial j}(d,j)\Big|_{\substack{d = d_0 \\ j = j_0}} \end{aligned}$$

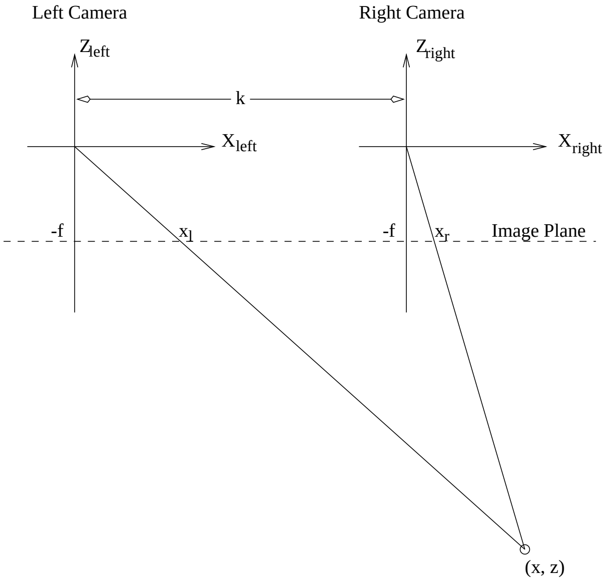

Figure 10: The Camera Geometry.

or

$$x(d,j) = \frac{k}{2} - (d - d_0)\frac{k}{2d_0} - (j - j_0)\frac{k}{d_0}.$$

So, the distance in x-coordinate between the two points in object space, determined by $(j_1, d_1)$ and $(j_2, d_2)$ is

$$x_2 - x_1 = -(d_2 - d_1)\frac{k}{2d_0} - (j_2 - j_1)\frac{k}{d_0}.$$

Similarly,

$$z(d,j) = \frac{kf}{ad_0} - (d - d_0)\frac{kf}{ad_0^2}$$

and so, the distance in z-coordinate is

$$z_2 - z_1 = -(d_2 - d_1)\frac{kf}{ad_0^2}.$$

Then, the distance between the points is

$$dist = s_h\sqrt{a^2(d_2 - d_1)^2 + (j_2 - j_1)^2 + (d_2 - d_1)(j_2 - j_1)}$$

where

$$s_h = \frac{k}{d_0}$$

and

$$a^2 = \frac{1}{4} + \frac{f^2}{a^2 d_0^2}.$$