3D BUILDING FACADE RECONSTRUCTION UNDER MESH FORM FROM MULTIPLE WIDE ANGLE VIEWS

Lionel Pénard, Nicolas Paparoditis, Marc Pierrot-Deseilligny
Institut Géographique National/MATIS, 2-4 avenue Pasteur, 94165 Saint-Mandé, France -
(lionel.penard, nicolas.paparoditis, marc.pierrot-deseilligny)@ign.fr

KEY WORDS: Surface Reconstruction, Digital Facade Models, Global Optimization, Multi-Image Matching, Terrestrial, Urban, Photogrammetry

ABSTRACT:

In this paper, we present an algorithm, which automatically generates textured meshes of building facades from a set of multi overlapping calibrated and oriented images. We are in a context of massive data production and aim at high geometric accuracy. The central idea is to create a 3D point cloud. After comparison between object-space and image-space techniques, we choose the latter and compute the depths of the pixels in each image using a correlation-based method. A post-processing step is necessary to filter points according to a confidence index and remove the sparse speckle noise. We then perform a global optimization to find a regularized surface. Finally, the 3D point cloud is triangulated. The resulting mesh is an accurate representation of the facade surface from each image view point. We obtain promising results, with a correct texture projection on the reconstructed model.

1. INTRODUCTION

Recently, the automation of high scale 3D city model reconstruction has been a field of great interest. Satellite imagery, aerial images and airborne laser scanners have been used to retrieve the geometry of buildings e.g. (Baillard and al., 1999; Fischer and al., 1998), but their resolution is not sufficient to provide accurate geometry on building facades. This task requires terrestrial data acquisition. Terrestrial reconstruction techniques mostly depend on the aim of the work and the acquisition system employed i.e. the available data. Many authors address the reconstruction of remarkable buildings or small sets of buildings (e.g (Coorg and Teller, 1999; Stamos and Allen, 2000; Fitzgibbon and Zisserman, 1998)). In this case, much time can be devoted to data acquisition (several range laser scans, images from many viewpoints...). But in our case, we want to acquire geometric information on the whole city instead of one single and remarkable building. Steady terrestrial range imagery can not thus be used, because of time and cost. Large 3D city models imply using a mobile acquisition device, onboard a vehicle which moves along the streets. Many authors have already investigated such mobile mapping devices (Frueh and al., 2005; Zhao and Shibasaki, 2001). In these papers, the geometry is determined using laser scans from the vehicle. Our data acquisition system is only image-based, the geometry has to be extracted from the images which are calibrated and oriented (Bentrah and al., 2004b).

Retrieving 3D geometry from a set of images is a classical problem. Semi-automatic methods (e.g. (El-Hakim, 2002)) enable to model the scene but require interaction with a human operator. We want our technique to be fully automatic since we are only interested in surface topography. Many attempts of automatic methods have been undertaken with good results: space carving (Kutulakos and Seitz, 2000), voxel coloring (Seitz and Dyer, 1999) or level sets methods (Faugeras and Keriven, 1998). But we want the algorithm to be very simple and not too time consuming because of the amount of input data involved. In (Wang and al., 2002), the authors propose to recover facade geometry by computing depth difference with the facade main plane within rectangular periodic patches. This technique is well suited for windows, but other objects may be present on facades. We opted for a technique based on 3D point extraction and triangulation.

This article is organized as follows. Section 2. gives the general strategy we adopted, section 3. concerns 3D point extraction, and section 4. briefly presents surface reconstruction. Then, results and future work are discussed in section 5..

2. THE STRATEGY

2.1 Available Data

The Stereopolis system (Bentrah and al., 2004a) consists of six cameras on top of a vehicle: the first two form a vertical stereo rig towards the left, two others form a vertical stereo rig towards the right, and the last two form a horizontal stereo rig pointing to the rear of the vehicle. The cameras are built and precisely calibrated by the IGN instrumentation lab, we thus assume that all the intrinsic parameters (focal length and distortion) are perfectly known. We also suppose that the pose parameters are known. Work (Bentrah and al., 2004b) is currently being undertaken to automate this preliminary step, but is far beyond the scope of this paper. Images at full resolution are 12 bits, 4096×4096, with a signal to noise ratio of 300. The camera have wide angle field of view with a 28 mm focal. Pixel size on buildings is a few millimeters. The overlapping is sufficient to ensure that every point of a facade is at least present on 2 images, and mostly on 4 or 6 images (see Figure 1).

We also assume that a 2D map of the city is available. It can be derived from a former coarse 3D model generated with aerial images or from a high scale cadastral ground map.

2.2 Algorithm Overview

The process can be summarized as follows:

- Select a building in the ground map
- Segment it into main facades
- Select oriented images viewing parts of this facade (section 2.3)
- Extract a 3D point cloud from each viewpoint (section 3.)
  - Calculate a correlation volume (sections 3.1 and 3.2)
  - Optimize the surface (section 3.3)
  - Select reliable 3D points (section 3.4)
- Mesh triangulate this cloud (section 4.)

Each step described in this paper is fully automatic.

2.3 Image Selection

The adopted strategy consists in selecting a building on the 2D map. Then its contour is segmented into several main edges representing the main facade planes of the building. Finally, all the images representing at least a part of this facade are kept for the following stages. Let note I this set of images. Figure 1 shows the set I corresponding to the selected facade.

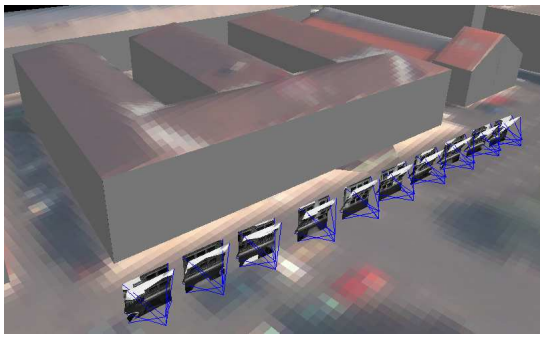


Figure 1: 3D view of the selected images in front of the coarse model.

3. OBJECT-SPACE VS IMAGE-SPACE APPROACH

We aim at the reconstruction of the facade surface, given the extracted set of calibrated images. Two types of techniques have been developed. In the first one, a master image is considered, from which features (e.g. points of interest) are extracted. The 'best' corresponding 3D feature position hypothesis is evaluated by reprojection in the other images, and computation of a matching score. These techniques are referred to as image-space based techniques. The other type is the object-space based techniques. The object space (the real 3D world) is discretized in voxels in the neighborhood of the facade's estimated position. Each voxel is projected in the images, then matching scores are calculated, and finally the 'best' 3D point hypothesis is selected.

3.1 Object-Space Based Point Matching

In our application, the object-space seems to be more natural than the image-space, since a facade is considered as a whole, instead of dealing with successive master images. It is also possible to keep perfect symmetry in the handling of the images.

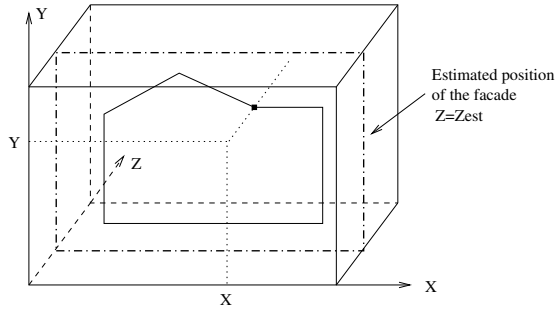


Figure 2: Object space discretization

We implemented the object-space technique as follows. We consider a coordinate system (X, Y, Z) chosen so that the Z axis is perpendicular to the estimated facade plane, and that the Y axis is vertical (see Figure 2).

The position of the facade, estimated from the 2D ground map, is thus the plane $Z = Z_{est}$, with estimated boundaries $X_{min} < X < X_{max}$ and $Y_{min} < Y < Y_{max}$. Since the facade position is not precisely known, and since facades are not strictly planar, let us delimit a search space $Z_{min} < Z < Z_{max}$ including $Z = Z_{est}$. The volume is discretized with a planimetric pitch $X_{pitch} = Y_{pitch}$ and a depth pitch Z_{pitch}. For each X and Y, we browse the Z axis between Z_{min} and Z_{max}. Figure 3 shows a top view.

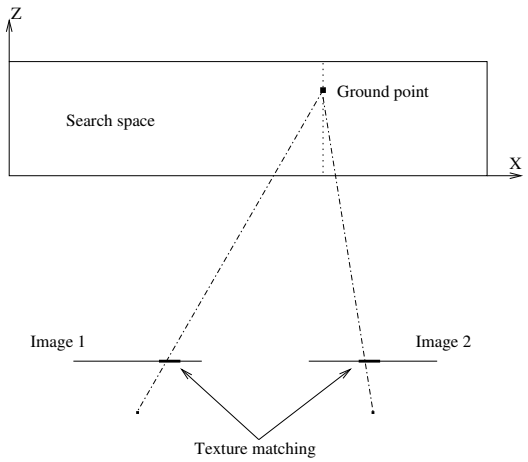


Figure 3: Reprojection of a ground point in the images. Top view.

When there is no self-occlusion of the building, a point $M(X, Y, Z)$ is theoretically visible in all images belonging to a subset I_M of I that can be calculated *a priori*. Let us denote m_i the projection of M in image i where $i \in I_M$. The coordinates (c_i, l_i) of image point m_i are the exact (subpixellar) values of the projected coordinates. We compute the multi-correlation coefficient of windows centered on the (c_i, l_i) which requires to resample the images. The multi-correlation score used is the one given by (Paparoditis and al., 2000).

The grey values v_{ij} of the correlation window centered on (c_i, l_i) are stored in the vector $\mathbf{V_i}$:

$$\mathbf{V_i} = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{in} \end{pmatrix} \quad (1)$$

and the multi-correlation coefficient is given by:

$$cor(X, Y, Z) = \frac{var(\sum_{i \in I_M} \mathbf{V_i})}{\sum_{i \in I_M} var(\mathbf{V_i})} \quad (2)$$

where $var(V_i)$ denotes the variance of the vector components.

Given X and Y, the set I_M depends on Z. To avoid considering different sets I_M when Z varies, let us choose

$$I_{X,Y} = \bigcap_{Z_{min} < Z < Z_{max}} I_M \quad (3)$$

and replace I_M by $I_{X,Y}$ in equation 2.

Finally, the result is $\hat{Z}_{X,Y}$ such that:

$$\hat{Z}_{X,Y} = \underset{Z_{min} < Z < Z_{max}}{\operatorname{argmax}} \; cor(X, Y, Z) \quad (4)$$

Unfortunately, this algorithm produces very poor results with our data. Figure 4 shows the digital facade model (DFM) obtained,

(a) Right part of the facade (b) Zoom on a window

Figure 4: Digital facade model, right of the facade. Planimetric pitch: 2cm. Correlation window size: 5 pixels. Dark pixels correspond to high depths, light pixels correspond to the nearest areas

with a planimetric pitch of 2cm, and a correlation window size of 5×5 pixels.

The depth of the window glass is totally wrong due to the absence of texture and to specular effects in this area. But far more concerning are the numerous errors in textured areas, even when no self-occlusion occurs. Our explanation of this phenomenon relies on two facts: on the one hand, many structures of the building are recurrent (e.g. stone edges, window posts,...) and thus false matchings can produce high correlation scores. On the other hand, our maximum search technique (X and Y are given, we look for the best Z) can lead to such errors. Figure 5 illustrates the fact that a wrong 3D point hypothesis (i.e. a 3D point with a false Z) can get a higher score. Correlation favours contrasted areas, and in the case of Figure 5, the correlation score is better for the red point because its reprojection corresponds to similar contours in all images.

Significant errors are present over continuous large areas. This is the reason why the optimization step performed in the image-space case would not be suitable here. In order to avoid these drawbacks, an image-space approach was adopted.

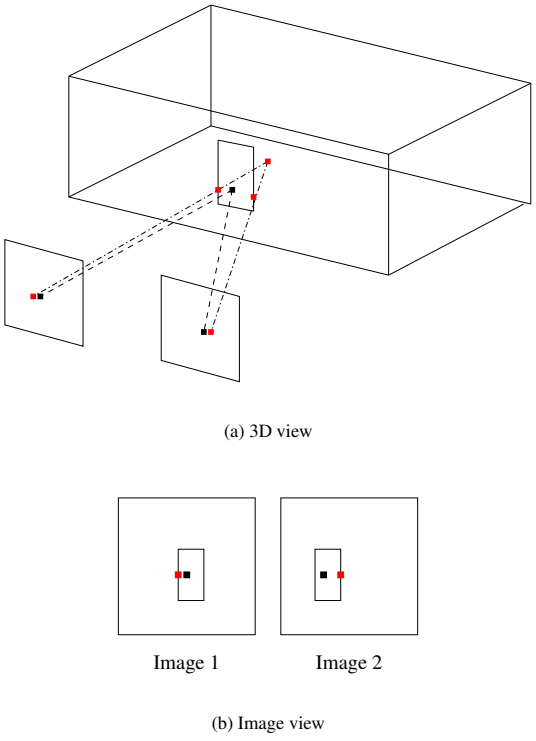
3.2 Image-Space Matching Strategy

For each pixel of a master image, we want to retrieve the depth of the corresponding 3D object. Each image represents only a part of a facade, and from a given view point, self-occlusion of the building always occurs, except in the case of strictly planar facades which are easy to model. It is thus necessary to add a post-processing step, which would consist in merging the geometrical information extracted from each view. This step is not described in this paper.

Let us select a master image. Now the 3D space is discretized according to this image geometry, i.e. a 3D point (X, Y, Z) is determined by its integer coordinates in the master image and by its depth.

For each pixel of the master image, for each depth between $depth_{min}$ and $depth_{max}$, the subpixellar position of the point is calculated in all images (see Figure 6). Then, we compute the associated correlation scores. The previous correlation coefficient is not adapted anymore because in this case a master image is to be compared to the others. Thus, the symmetric coefficient has to be replaced by a dissymmetric correlation value; we have used the mean value of the classical cross correlation coefficient between the master image and all the other images.

$$cor_i(c, l, depth) = \frac{1}{N-1} \sum_{j \neq i} CCS(\mathbf{V_i}, \mathbf{V_j}) \qquad (5)$$



(a) 3D view

(b) Image view

Figure 5: Wrong matches can get higher scores when the correct position is little textured. Black dots: correct position. Red dots: detected position

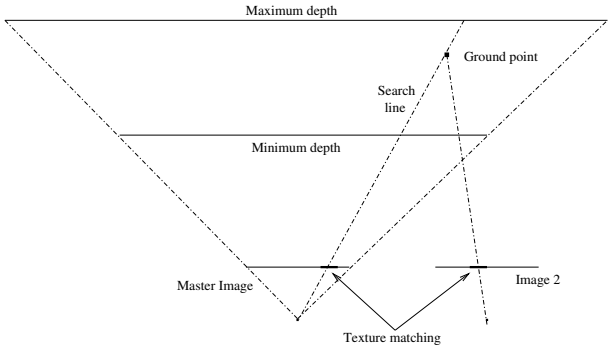


Figure 6: Search line from a given pixel in the master image. Reprojection of a ground point hypothesis in the other images. Top view.

where i denotes the master image, N the total number of images viewing the corresponding 3D point, $\mathbf{V_i}$ the texture vector of the correlation window centered on the projection of $(c, l, depth)$ and $CCS(\mathbf{V_i}, \mathbf{V_j})$ the classical correlation score.

Results are far better than in the object-space case, as it can be seen on Figure 7.

The depth of the maximum score corresponds to the real depth for nearly all pixels. Yet, some pixels have erroneous depth values, mainly along lines which represent other image limits. This phenomenon is due to the change of the normalization factor $N - 1$ in equation 5 when these lines are crossed.

(a) Original image

(b) Depth corresponding to correlation maximum

(c) Depth corresponding to correlation maximum: zoom on a window

Figure 7: Image-space depth extraction. Correlation window size: 3 pixels, depth pitch: 2cm.

3.3 Global Matching Optimization

In order to improve the results, we perform an optimization step rather than choosing only the higher score. The goal is to remove these small perturbations and the sparse remaining noise while taking into account that the facade surface should be continuous. We use the optimization process developed by (Roy and Cox, 1998). For every pixel and every depth, correlation scores are stored in a 3D matrix: this will be the data term. The regularization parameter is determined experimentally. A low value corresponding to a low data fit error is chosen to preserve discontinuities. The result is shown in image 8(a). The surface is smoothed, but real discontinuities (for instance, around the windows) are well preserved.

3.4 Point Filtering

We expect a depth value at each pixel. But we know that results are irrelevant in some parts of the scene such as sky, background buildings whose distance is too high and whose correlation scores are low within the search space $depth_{min} < depth < depth_{max}$. Window glass is another example where depths are

(a) Depth map after optimization

(b) Depth map after filtering

Figure 8: Image-space post-processing. Original image can be seen in Figure 7(a)

false, because the lack of texture makes the correlation scores very sensitive to noise. We thus perform a filtering step, which divides into 2 parts. The first one takes place before the optimization step. In order to avoid false data which may alter the results' quality, the 3D matrix of correlation scores is modified. If one of the two following conditions are fulfilled:

$$\begin{cases} \max_{depth_{min} < depth < depth_{max}} cor_i(c, l, depth) < t_c \\ var(\mathbf{V_i}) < t_v \end{cases} \quad (6)$$

where t_c is a correlation threshold, and t_v is a variance threshold, the matrix entries corresponding to the pixel (c, l) are set to 0. The first condition removes pixels for which no good match has been found in the other images. It can correspond to occluded pixels in the master image or to background objects. The second condition removes pixels with very low texture.

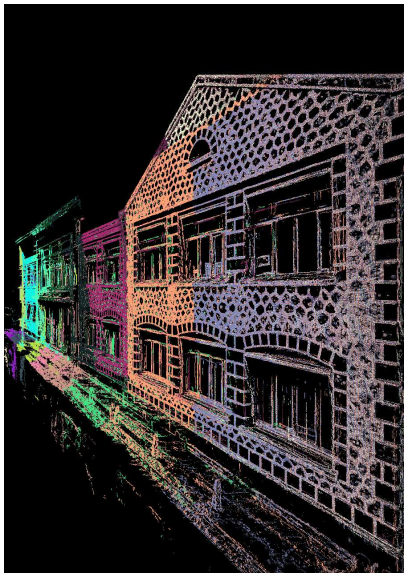
These two conditions describe pixels for which we have very little confidence to give good depth estimations. Then, after the optimization step, pixels that fulfil one of the conditions 6 are removed. The result can be seen in Figure 8(b)

4. SURFACE RECONSTRUCTION

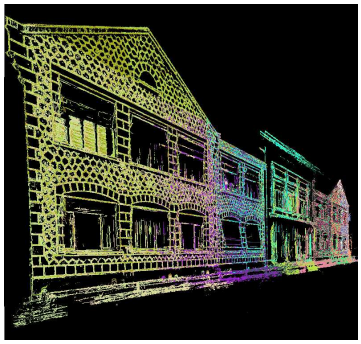
In section 3., for each image of the set I, we obtain a set of pixels whose depth is known. These are considered as 3D points. Figure 9 shows the point cloud corresponding to the whole facade surface. Colors correspond to the label of the master image that produced the point. The relative position and orientation of all the point clouds is excellent.

Since many points were removed after the filtering step, the depth map is not dense anymore. The point cloud is then triangulated, using a Delaunay triangulation, driven by the image geometry, and textured. When every pixel is removed over a large area because of the low texture, triangulation leads to planar approximation. In fact, we observed that untextured regions often correspond to planar areas such as window glass, uniformly painted materials. Figure 10 shows the result of the triangulation on a window.

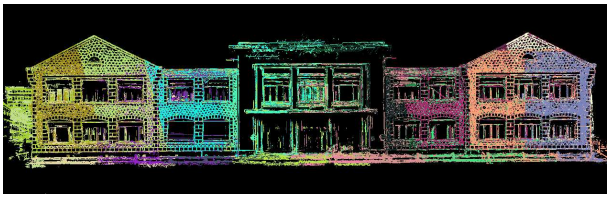
If a planar region is highly textured all the points will be kept. This generates numerous coplanar vertices. We perform the mesh simplification algorithm QSlim (Garland and Heckbert, 1997) to reduce the number of triangles without any significant loss in accuracy.



(a) Side view from left



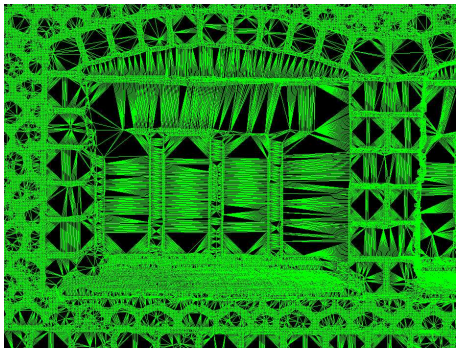
(b) Side view from right



(c) Front view

Figure 9: 3D point cloud. Points extracted from different master images have different colors

(a) Original image



(b) Corresponding triangles

Figure 10: Triangulation of the 3D points located on a window.

5. RESULTS AND FURTHER WORK

5.1 Results

Our algorithm has lead to the production of several textured meshes per facade. Each model corresponds to one image viewing point. Figure 11 shows different views of one of these models. Large planar regions are well reconstructed. The position of windows limits are very accurate. Untextured areas such as window glass are reconstructed as planar regions between the posts. False triangles appear between window posts and walls (see Figure 11(a)). This is due to the lack of high confidence points. At the bottom of the building, the large white untextured area is responsible for the lower reconstruction quality: again, confidence index on 3D points is too low. Figure 12 shows another example. Reconstruction is correct on the main part, but errors can be observed on windows and at the bottom of the building.

Table 1 shows reconstruction error on 9 ground control points. These points were used to calculate the pose parameters of the view shown figure 7(a) thus errors are only due to the reconstruction process and not to georeferencing. The good reconstruction quality on these points (maximum error: 17.7 mm with a depth discretization pitch of 10 mm) can be partly explained by the fact that they are points of interest in the image and our algorithm is more likely to work good on such points. Nevertheless, we can affirm that no important bias was introduced by our technique.

Point nb	Error (mm)
81	2.5
82	0.5
83	3.2
91	4.4
92	4.8
93	2.7
101	17.7
102	5.9
103	5.6

Table 1: Reconstruction error on 9 control points in image of figure 7(a).

5.2 Future Work

No constraints were applied to the triangulation. For instance, the numerous segments of the scene are not exploited. They are common objects on architecture scenes, and they should in the future be detectable in the images, to be matched to produce 3D segments. It would help the results to be more photo-realistic with straight edges. Another limitation of our current approach is that we obtain 3D meshes for each view point. We intend to provide complete facade models by merging all the views, and next to provide complete building models by merging all the main facades.

6. CONCLUSION

In this paper, we have presented a simple method to produce 3D facade models using only images as the source of geometric and radiometric information. Images were calibrated and oriented in a former stage. After focusing on a facade and extracting the required images, the first step of our algorithm consists in 3D point extraction, which is carried out with an image-space technique. Better results are indeed observed than when using an object-space method. Then a global optimization and a filtering are performed to remove errors on noisy points and points with a low confidence index. The obtained 3D cloud is triangulated and textured to get the final surface reconstruction. The results show good geometric precision for our data. Our method is robust, accurate and produces dense point clouds compared to mobile laser scanners. This technique has been applied to images of facades but it can be applied to any kind of landscape since no assumption was made.

REFERENCES

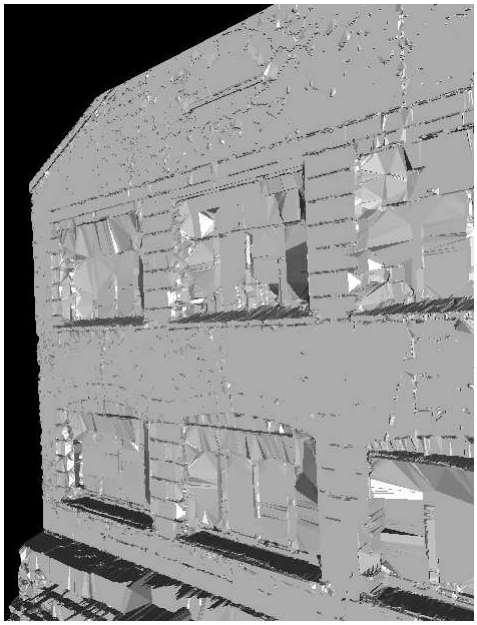
Baillard, C., Schmid, C., Zisserman, A., and Fitzgibbon, A., 1999. Automatic line matching and 3D reconstruction of buildings from multiple views. In *Proc of ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery, IAPRS*, volume 32, pages 69–80.

Bentrah, O., Paparoditis, N., and Pierrot-Deseilligny, M., 2004a. Stereopolis: An image based urban environments modeling system. In *MMT 2004. The 4th International Symposium on Mobile Mapping Technology, Kunming, China.*

Bentrah, O., Paparoditis, N., Pierrot-Deseilligny, M., and Horaud, R., 2004b. Estimating sensor pose from images of a stereo rig. In *ISPRS, Istanbul.*

Coorg, S. and Teller, S., 1999. Extracting textured vertical facades from controlled close-range imagery. In *CVPR'99*, pages 625–632, Fort Collins.

El-Hakim, S., 2002. Semi-automatic 3d reconstruction of occluded and unmarked surfaces from widely separated views. In



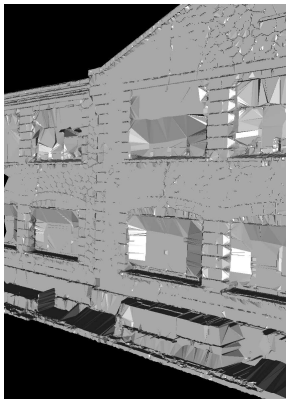
(a) Side view untextured

(b) Side view textured

Figure 11: Reconstructed surface from one image view point

(a) View 1 untextured

(b) View 1 textured



(c) View 2 untextured

(d) View 2 textured

(e) View 3 textured

Figure 12: Reconstructed surface from another image view point

ISPRS Commission V Symposium on Close Range Visualization Techniques, pages 143–148.

Faugeras, O. and Keriven, R., 1998. Complete dense stereovision using level set methods. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I*, pages 379–393. Springer-Verlag.

Fischer, A., Kolbe, T., Lang, F., Cremers, A., Förstner, W., Plümer, L., and Steinhage, V., 1998. Extracting buildings from aerial images using hierarchical aggregation in 2D and 3D. *CVIU*, 72(2):163–185.

Fitzgibbon, A. W. and Zisserman, A., 1998. Automatic 3D model acquisition and generation of new images from video sequences. In *Proceedings of European Signal Processing Conference (EUSIPCO '98), Rhodes, Greece*, pages 1261–1269.

Frueh, C., Jain, S., and Zakhor, A., 2005. Data processing algorithms for generating textured 3d building facade meshes from laser scans and camera images. *Int. J. Comput. Vision*, 61(2):159–184.

Garland, M. and Heckbert, P. S., 1997. Surface simplification using quadric error metrics. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216. ACM Press/Addison-Wesley Publishing Co.

Kutulakos, K. N. and Seitz, S. M., 2000. A theory of shape by space carving. *Int. J. Comput. Vision*, 38(3):199–218.

Paparoditis, N., Thom, C., and Jibrini, H., 2000. Surface reconstruction in urban areas from multiple views of aerial digital frames. In *Proceedings of the XIXth ISPRS Congress*, volume 33 of *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Amsterdam. ISPRS.

Roy, S. and Cox, I. J., 1998. A maximum-flow formulation of the n-camera stereo correspondence problem. In *International Conference on Computer Vision*, page 142, Bombay.

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

Stamos, I. and Allen, P., 2000. 3d model construction using range and image data. In *Computer Vision and Pattern Recognition*, pages 531–536.

Wang, X., Totaro, S., Taillandier, F., Hanson, A., and Teller, S., 2002. Recovering facade texture and microstructure from real-world images. In *Proceedings of ISPRS Commission III Symposium on Photogrammetric Computer Vision*, pages 381–386, Graz.

Zhao, H. and Shibasaki, R., 2001. Reconstructing textured cad model of urban environment using vehicle-borne laser range scanners and line cameras. In *ICVS '01: Proceedings of the Second International Workshop on Computer Vision Systems*, pages 284–297. Springer-Verlag.