DETAILED RELIEF MODELING OF BUILDING FACADES FROM VIDEO SEQUENCES

Wolfgang v. Hansen[a], **Ulrich Thönnessen**[a], **Uwe Stilla**[b]

[a]FGAN-FOM Research Institute of Optronics and Pattern Recognition
Gutleuthausstr. 1, 76275 Ettlingen, Germany
wvhansen@fom.fgan.de
[b]Photogrammetry and Remote Sensing, Technische Universitaet Muenchen,
Arcisstr. 21, 80280 Muenchen, Germany

Working Group III/2

KEY WORDS: Building, Surface, Reconstruction, Photo-realism, Texture, Urban, Close Range, Video

ABSTRACT

Three dimensional building models have become important during the past years for various applications like urban planning, enhanced navigation or visualization of touristic or historic objects. Altough for some applications geometric data alone is sufficient, for visualization purposes a more realistic representation with textured surfaces is necessary. The associated textures from buildings are extracted either from airborne imagery or, especially for facades, from images taken by ground based cameras. For very high demands on photorealistic quality, textures mapped on simple geometric models like polyhedra or regular surfaces are not sufficient because relief structure is not preserved. This leads to an unrealistic and artificial impression on close-up views. Relief structures are beneficial for large scale models that are closely inspected within a limited area in a virtual world. In this paper the extraction of reliefs to improve existing planar surfaces of wire frame models of buildings is described. Given several uncalibrated views onto a surface of the polyhedral model, a depth map is estimated by correlation. The underlying plane is used to guide the correlation in order to detect outliers and to fill in homogeneous areas.

1 INTRODUCTION

Three dimensional building models have become important during the past years for various applications like urban planning, enhanced navigation or visualization of touristic or historic objects (Brenner et al., 2001). Building models are typically acquired by a (semi-) automatic processing of laser scanner elevation data or aerial imagery (Baillard et al., 1999). Although for some applications geometric data alone is sufficient, for visualization purposes a more realistic representation with textured surfaces is necessary.

The associated textures from buildings are extracted either from airborne imagery or, especially for facades, from images taken by ground based cameras (Teller, 1998). For very high demands on photorealistic quality, textures mapped on simple geometric models like polyhedra or regular surfaces are not sufficient because relief structure is not preserved. This leads to an unrealistic and artificial impression on close-up views. Relief structures are beneficial for large scale models that are closely inspected within a limited area in a virtual world. For architectural and touristical applications it is advantageous to dispose of enhanced facade models with relief information in order to improve visualization of door and window openings as well as ornaments.

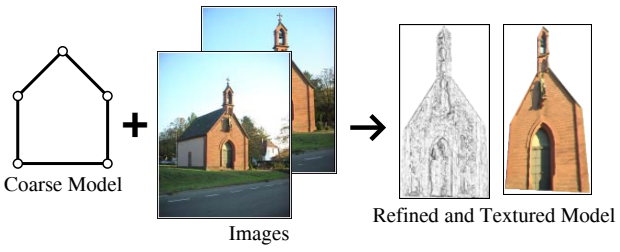

Figure 1: Refinement of a coarse model by images.

Facade relief information can be extracted by measurement of depth information using a terrestrial laser scanner or by photogrammetric analysis of the same images also acquired for texture mapping. Nowadays, video cameras are widely available as an inexpensive source of data. Data acquisition often is not planned as thoroughly as for a true photogrammetric campaign and therefore provides more challenging data. For some historic buildings that have been destroyed or otherwise changed, images taken by tourists or non-professionals might be the only source of information available (Grün et al., 2003). Because the resolution of images of a video camera is low compared to that of a photographic camera one of the important questions is the amount of geometric accuracy and level of detail that can be expected from such acquisitions.

In this paper the extraction of reliefs to improve existing planar surfaces of wire frame models of buildings is described. Given several uncalibrated views onto a surface of the polyhedral model, a depth map is estimated by correlation. The underlying plane is used to guide the correlation in order to detect outliers and to fill in homogeneous areas.

Basically, two different methods to reconstruct 3D objects from images for visualization in a virtual environment can be identified. One is to generate surface models directly by dense stereo matching (Pollefeys, 1999). No explicit knowledge of the exact shape of the objects is required which makes the approach simple in the sense that only

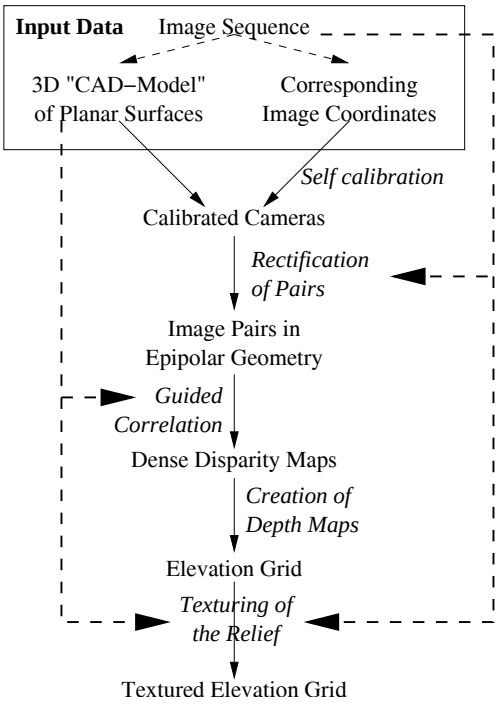

Figure 2: Overview of the processing chain from a coarse to a refined model.

very generic structural models (e. g. several smooth planes at different depth levels) are required. On the other hand, different objects or parts of them are not separated by the model and can not be treated individually by the visualization process.

The other approach is to use models of objects which are expected to be present in the scene like e. g. buildings in an urban environment. This typically results in scene objects represented by polyhedral models made from plane surfaces. For visualization, areas of the original images are extracted and mapped as flat textures on each of the surfaces. The advantage is that the objects are already hierarchically structured entities that are easy to handle. But relief structures within each surface are lost if they are – and this is the general case – not contained in the model.

Both methods produce models which are less realistic on close inspection and very oblique views because no relief structure is visible. In our approach, a hybrid method that combines the advantages of both reconstruction techniques in a hierarchical manner is proposed (Fig. 1). Polygonal boundaries of planar surfaces are taken as input and constrain the reconstruction of the finer details which are retrieved via dense matching. The result is a dense elevation grid that can be used to replace the flat texture. Because the elevation grid is restricted geometrically to the polyhedral model at its borders, it fits to other surfaces without any error.

2 PROCESSING CHAIN

In this section the processing chain from input data to the final product is described. Refer to Fig. 2 for an overview.

2.1 Input Data

The inputs of the processing chain are an initial coarse geometric model of the object and a set of associated images mapping the building. The coarse model is a wire frame model of the building, in which planar surfaces are described by 3D outline polygons. The images are needed for both relief reconstruction and texturing. They are linked to the polygons through known image coordinates of the vertices. The process needs the determination of the camera parameters by using the 3D coordiates. Therefore, it must be assured that the 3D points are not coplanar in space so that camera parameters can be computed for the 3D reconstruction. This is not the case for e. g. a single facade, but can be circumvented by inclusion of some additional points outside the plane. Here are two possible methods to obtain suitable input models:

2.1.1 Retrieve model from images For a site where only images exist, i. e. where no model has been generated or made available, it is possible to retrieve the wire frame model from the images. Corresponding points in different images generally provide enough information to calibrate the cameras and to reconstruct the 3D coordinates of the imaged points. Next, either the 3D bounding lines of the surfaces are extracted and intersected or the corners of these surfaces are connected in order to create the polygons.

2.1.2 Coregistrate existing model to images We have assumed that there already exists a wire frame model. Here, the 3D points are already given and only have to be marked in the images so that the corresponding image coordinates are available. No knowledge about the camera parameters is needed as input, but for later self calibration it should be known which of the parameters (e. g. focal length) can be considered constant.

2.2 Self calibration

The estimation of a depth map requires knowledge about the pose of the cameras as well as their calibration parameters. If these are not known they have to be computed from the given point assignments. Such a task – simultaneous computation of inner and outer camera parameters when no initial values are known – is commonly referred to as auto or self calibration (Hartley and Zisserman, 2000). In the case that corresponding 2D and 3D points are available, the following two step strategy could be applied:

(a) Linear Resection Resection computes the homogeneous 3×4 projection matrix $\mathbf{P}$ from corresponding image points $\mathbf{x}_i$ and world points $\mathbf{X}_i$ which are related by

$$\mathbf{x}_i = \mathbf{P}\mathbf{X}_i \tag{1}$$

Using the vector cross product

$$\mathbf{x}_i \times \mathbf{P}\mathbf{X}_i = 0 \tag{2}$$

Eq. 1 can be transformed into an equivalent equation

$$\mathbf{A}\mathbf{p} = 0 \tag{3}$$

Figure 3: Three different views onto the west facade of Alexiuskapelle. Superimposed on the middle image is the given outline polygon of the facade.

where $\mathbf{A}$ is composed from known $\mathbf{x}_i$ and $\mathbf{X}_i$, and $\mathbf{p}$ are the coefficients of the unknown matrix $\mathbf{P}$ written as a vector. The solution $\mathbf{p}$ is the right null vector of $\mathbf{A}$. Refer to (Hartley and Zisserman, 2000) for details.

An initial solution of $\mathbf{P}$ is computed seperately for each image. There are some disadvantages to the set of camera parameters gained this way. First, parameters that usually are constant for all images, like e. g. the principal point, are calculated individually and will have slightly different results due to noise in the coordinates. Second, as this is a linear solution and therefore only linear parameters can be estimated. Nonlinear parameters such as radial lens distortion can not be computed. Finally, only the algebraic error $||\mathbf{A}\mathbf{p}||$ is minimized which in general is not geometrically meaningful (Hartley and Zisserman, 2000).

(b) Bundle adjustment To overcome all disadvantages, bundle adjustment is a feasible choice. It is based on an equation system of nonlinear versions of Eq. 1 where a term for radial distortion has been added. Parameters that are constant for all images have been included only once. The equation system is solved simultaneously for all unknown parameters. The geometric error minimized is the sum of the squared distances between projected and measured points in the images.

The result of these two steps are a set of camera parameters – camera calibration as well as pose – that define the relationship between the vertices of the coarse model and the image points best.

2.3 Rectification of Image Pairs

Rectification of image pairs is to transform both images such that all epipolar lines are parallel to image scan lines and that corresponding epipolar lines have the same y-coordinate (Koch, 1997). This is an important preprocessing step for dense stereo matching because it simplifies the search for corresponding pixels along the (normally slanted) epipolar lines to a search along the scan lines.

First the epipoles are projected to infinity. Then the images are rotated such that the epipoles lie in the direction of the x-axis. Finally, one of the images is shifted in the y-direction so that corresponding lines of the pair coincide. There are still two degrees of freedom left to improve the rectification without destroying its properties: a shear along the x-axis and scale factors for both axes. These have been exploited such that the coordinates in one of the images are closest possible to the original coordinates.

The rectifying 2D homographies are left multiplied to the camera matrices which causes the actual camera calibrations to change. This must be taken into account when computing depth from pixel disparities.

2.4 Guided Correlation

Given images in epipolar geometry, the main task is to compute a dense disparity map, i. e. a map containing the relative distances between corresponding pixels, that describes the surface relief. In a later step the disparity map will be upgraded to a depth map containing metric units instead of pixel units. Corresponding pixel locations are found via cross correlation – a local maximum hints at a possible match, but repetitive patterns will also deliver false hints and in homogeneous regions there are no clear maxima. A blind search for maximum correlation only does not give satisfactory results.

2.4.1 Dynamic Programming A dynamic programming scheme that allows to incorporate some constraints to guide the matching has been used (Falkenhagen, 1997). The general idea is to define a cost function plus some constraints and to find its global minimum. This is feasible whenever all decisions can be broken up into a sequential scheme.

For each scan line, we can put up a two dimensional cost matrix where columns and rows correspond to pixel positions and disparity values respectively (see Fig. 7). The sequential scheme is implemented by filling in the costs

Figure 4: The first two images of Fig. 3 in epipolar geometry.

columnwise from left to right, i. e. by moving along the scan line. Each column with known costs represents the set of all partial solutions up to that point. The next column will be filled in such that the best possible solution found so far is extended. Pointers to the preceding solution are stored in the matrix so that one can trace back to recover the complete path of position/disparity pairs for the final solution (Fig. 8).

For each cell (i, j), its costs $C_{i,j}$ are defined by

$$C_{i,j} := \min_{i' \in P_{i,j}} C_{i',j-1} + C^0_{i,j} \tag{4}$$

where $P_{i,j}$ is the set of allowed predecessor rows of (i, j) and $C^0_{i,j}$ are the costs to include (i, j) into the solution. I. e. the best solution found so far among all valid predecessors will be extended by the current position.

The costs $C^0_{i,j}$ have been set to

$$C^0_{i,j} := 1 - \mathrm{r}_{j,j+i} \tag{5}$$

where $\mathrm{r}_{j,j+i}$ is the cross correlation coefficient between both image locations (i is the disparity, i. e. an offset to the one-dimensional position j).

2.4.2 Constraints Three constraints have been applied to guide the matching process such that recovered depth at the borderline of the polygon matches the borders and that in the interior a smooth surface will be generated.

(a) Start and end cell In a general setup, one would have all cells of the first column as start cells. After cost propagation through to the last column every row is a valid end cell with known total costs. The one with the least costs would be chosen and backtracking reveals the corresponding start cell.

Here, the disparities at both ends can be computed from the known polygons. This allows to give only one start cell on which all intermediate solutions will be based. The end cell will no longer be chosen based on its costs, but according to the disparity at the other end. At first glance, it seems counterintuitive not to choose the globally optimal solution, but it is reasonable because the path that links predefined start and end cell still is the optimal path between these two.

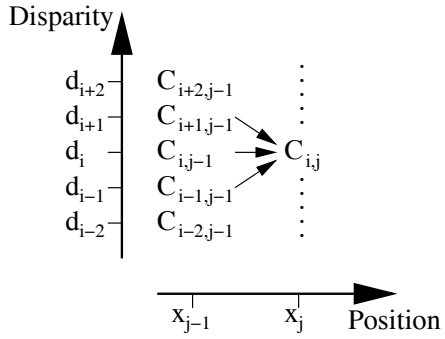

Figure 7: Detail of the cost matrix.

Figure 8: Path through the cost matrix. The lower left triangle is missing because of the given start cell and the ordering constraint.

(b) Ordering constraint The most important constraint to get a feasible path is the ordering constraint (Koch, 1997). It is obvious that the order of object points given in one image can not be reversed in the other image because in this case the view would be blocked. This limits the change of disparity between neighbouring pixels to no more than ± 1. Thus we define the set of valid predecessors as

$$P_{i,j} := \{i' : |i' - i| \leq 1\} \tag{6}$$

(c) Neighbouring scan lines One of the problems with a scan line oriented approach is how to deal with neighbouring scan lines. If they are processed completely unrelated, it happens in borderline cases that a wrong path is chosen and large jumps in disparity occur between consecutive lines. A penalty $C^p_{i,j}$ had been introduced to keep disparity values on one line close to those of the preceding line which extends the previous definition of $C^0_{i,j}$:

$$C^0_{i_0,j} := 1 - \mathrm{r}_{j,j+i_0} + C^p_{i_0,j} \tag{7}$$

$$C^p_{i_0,j} := c\,|i_0 - i_{-1}| \tag{8}$$

where i_0 referes to the disparity of the current line and i_{-1} to that already found for the preceding line. c is a small constant factor to affect the influence between neighbouring lines. Since there exists a closed polygon of initial values, there are always reference disparities available.

2.5 Creation of the Depth Map

After guided correlation, there exists a disparity map for each combination of input images. For the three input images shown in Fig. 3, three disparity maps are calculated. First, each of them is converted to a depth map by triangulation using the recovered camera parameters.

Additionally, the 3D coordinates can be reduced to 2.5D (two spatial dimensions plus depth) by subtraction of the given surface plane of the wire frame model. This results

Figure 5: Two artificial views onto the west facade of Alexiuskapelle with flat surface model. See text for explanation of details: (a) sky, (b) statue, (c) shadow.

Figure 6: The same two views as in Fig. 5 but this time with recovered relief taken into account.

in a set of elevation grids which can be combined into one final elevation grid by means of planar homographies. Depths that belong to the same xy-position will be overlayed by these transforms. Some robustness is gained in this way, because each depth has been calculated independently – with respect to errors of the guided matching – and the final depth is computed as an average or median depth.

2.6 Texturing the Relief

Similar to the depth, the appropriate color is taken from one input image from the respective location. In principle, color could be taken from all images but then it has to be solved which pixels are adequate. The final result is a regular 2D raster where each cell contains a depth value and an associated surface color.

3 EXPERIMENTS

Test Data Tests have been carried out on test data of a chapel (west facade of Alexiuskapelle, Ettlingen). A video sequence had been taken using a SONY DCR VX-2000 handheld video camera while walking along the pavement across from the chapel. The camera had been pointed towards the facade so that the images are all slightly convergent. The image format is standard video resolution of 720×576 with the camera held sideways and the zoom set to the minimum focal length of about 6 mm (1/3″ CCD) in order to fit best to the dimensions of the facade. From the whole sequence, three images – each about 100 frames apart – have been selected for processing (see Fig. 3).

Camera Calibration Because the camera is a standard consumer product originally not intended for measurement purposes, its inner parameters were unknown and had to be determined by means of self-calibration. Control points have been chosen manually to eliminate possible outliers that could reduce the quality of the reference frame. The results of the bundle adjustment for the inner parameters are given in Tab. 1.

Along with the inner parameters of the camera, a metric reference frame for all points has been computed. An absolute scale has been introduced by setting the width of the fassade to 6 m. The 3D-coordinates of the control points have got a standard deviation of 14 cm.

Rectification and relief recovery An example for a rectified image pair is given in Fig. 4. One can verify that corresponding points indeed lie on the same image row. Two different models have been generated according to the processing chain as described in the previous sections. The one in Fig. 5 assumes a flat surface whereas the other is shown with the recovered relief (Fig. 6). In order to demonstrate the differences, both models are shown from the same two artificial viewpoints.

4 DISCUSSION

On first impression, the quality of the flat model looks better because straight lines look much smoother than in the relief model. This is due to geometrical errors of the relief. Because the resolution of the video camera is low copared

Parameter	Unit	Value
focal length f	pix	947
aspect ratio a	1	0.94
skew s	1	0
principal point x_0	pix	259
principal point y_0	pix	384
radial distortion k	pix^{-2}	$-1.7 \cdot 10^{-7}$

Table 1: Inner camera parameters.

to photographic cameras, one pixel on the facade has a size of approx. 3 cm, whereas a disparity change of one pixel changes the depth by about 5 cm even though camera setup could be considered wide base stereo. A noise of 1 pixel in the disparity map therefore results in a noise of 5 cm in the depth map which makes the surface pretty coarse. This could be improved with an enhanced dynamic programming scheme that allows subpixel disparity values. Since this example has been computed with only three images out of over 100, better results might be obtained if more images are used in order to take advantage of averaging.

To verify the influence of the relief on the visual impression of the results, three positions have been marked in Fig. 5. For the flat model it can be seen that the relative position of the marked features does not change with respect to their surroundings. Both views are related by a planar homography.

For the relief model, different angles of view lead to different visibility of details that are below or above the facade. Bell and sky (a) can only be seen from the right view. The statue (b) which is embossed on the facade changes its relative position to features directly on the facade. The door frame blocks off the view onto the leftmost part of the door in the left image, making the sunlit part (c) smaller than compared to Fig. 5.

There are some gross errors in reconstruction of the facade. Fig. 9 shows the left bottom part of the model where errors during guided matching lead to a cavity. This and other arrors occur probably because the epipolar lines are parallel to the horizontal structures on the facade due to horizontal movement. In areas with little or even irritating texture this will mislead the matching. One possibility to circumvent this effect would be to use images taken from a different height so that the epipolar lines run diagonally or vertically across the facade. Errors like these however do not influence the outline of the reconstructed model because depth is forced to given values at the borders.

5 CONCLUSIONS

A hybrid model that refines a coarse wire frame model by detailed relief recovered from images has been proposed. The approach is suitable for rapid prototyping, because the required model can easily be constructed by manual interaction whereas the relief will be generated automatically. Results including a comparison with a flat model are given for a set of images taken by a hand held video camera.

Figure 9: Gross error in reconstruction. This figure shows the lower left part of the facade.

They show that relief structure can be retrieved and that relief information can upgrade visual impression. The overall quality is limited by the resolution of the camera but improvements are expected if the number of images is increased. Additionally, the dynamic programming scheme could be enhanced to allow a disparity estimation with subpixel accuracy.

REFERENCES

Baillard, C., Schmid, C., Zisserman, A. and A.Fitzgibbon, 1999. Automatic line matching and 3d reconstruction from multiple views. In: ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery, Vol. 32.

Brenner, C., Haala, N. and Fritsch, D., 2001. Towards fully automated 3d city model generation. In: E. Baltsavias, A. Grün and L. van Gool (eds), Proc. 3rd Int. Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images.

Falkenhagen, L., 1997. Hierarchical block-based disparity estimation considering neighbourhood constraints. In: Proc. Int. Workshop on SNHC and 3D Imaging, Rhodes, Greece.

Grün, A., Remondino, F. and Zhang, L., 2003. Automated modelling of the great buddha statue in bamiyan, afghanistan. In: Photogrammetric Image Analysis, IAPR, Vol. XXXIV-3/W8.

Hartley, R. and Zisserman, A., 2000. Multiple view geometry in computer vision. Cambridge University Press.

Koch, R., 1997. Automatische Oberflächenmodellierung starrer dreidimensionaler Objekte aus Stereoskopischen Rundumansichten. PhD thesis, Universität Hannover.

Pollefeys, M., 1999. Self-calibration and Metric 3D Reconstruction from Uncalibrated Image Sequences. PhD thesis, Katholieke Universiteit Leuven.

Teller, S., 1998. Automated urban model acquisition: Project rationale and status. In: DARPA Image Understanding Workshop.