

REFINEMENT OF FACADE MODELS USING IMAGE SEQUENCES

Wolfgang v. Hansen^a, Ulrich Thönnessen^a, Uwe Stilla^b

^aFGAN-FOM Research Institute of Optronics and Pattern Recognition
Gutleuthausstr. 1, 76275 Ettlingen, Germany
wvhansen@fom.fgan.de

^bPhotogrammetry and Remote Sensing, Technische Universität München,
Arcisstr. 21, 80280 München, Germany

KEY WORDS: Building, Surface, Reconstruction, Texture

ABSTRACT

Three dimensional building models have become important during the past years for visualization of touristic or historic objects. Detailed facade models are essential for realistic impressions on a terrestrial observer. Such models require both a geometric component that describes the shape of the building or parts of it as well as a texture component that provides color information and depicts details that have not been incorporated in the geometric model.

Textures for building facades are typically extracted 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.

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. Video sequences or photographs might be the only source of information available, if facades have been destroyed or otherwise changed or if costs are a limiting factor.

In this paper the refinement of a coarse wire-frame model is described. Given several views of a building along with some outline polygons for its facades, textured relief maps will be added to the model. The depth information is recovered from the images by correlation, where the given polygons provide guidance to reduce the number of outliers.

Results are given for a set of images taken by a hand held video camera. They show that relief structure due to window holes or ornaments can be retrieved but that the overall geometric quality is limited by the resolution of the camera.

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 touristic 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.

Facade relief information can be extracted by measurement of depth information using a terrestrial laser scan-

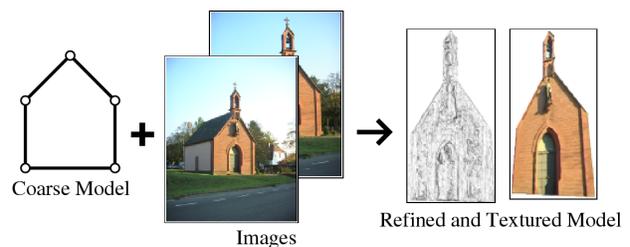


Figure 1: Refinement of a coarse model by images.

ner 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.

Basically, two different methods to reconstruct 3D objects from images for visualization in a virtual environment can be identified. One is a bottom-up approach in which surface models are generated 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 one global unstructured object model is generated. Different objects or parts of them are not separated by the model so that modeling of geometric primitives from the point cloud would require further processing. A major advantage is that texture information already is included in the point mesh so that explicit definition and mapping of texture images is not necessary.

The complementary top-down 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 with an overall simple geometry. 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. Only a few large texture images – appropriate parts either directly from original images or preprocessed versions thereof – are required. The disadvantages for this approach are that an automatic generation is more difficult compared to the bottom-up approach and that relief information will not be available inside the polygons if relief structures are – and this is the general case – not contained in the model.

Both methods have in common that they do not cross the facade level. The bottom-up approach results in a point mesh; even though quite large objects may be covered by such a mesh, each point is only linked to its direct neighbors. None of the point links carries any information whether or not a boundary between two neighboring surfaces is crossed. The top-down approach on the other hand models large objects quite accurately by breaking them into meaningful parts like e. g. different surfaces. Such a surface, however, is described only by its polygonal boundary. The interior of the polygon remains unmodelled except for the assumption that in general it is flat.

In this paper, a hybrid method that combines the advantages of both reconstruction techniques in a hierarchical manner is proposed (Fig. 1). 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. 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.

The advantages of this approach are that the geometric qualities are taken from the globally more accurate CAD model whereas the fine structures of a more detailed relief are recovered from images by dense matching. Flat textures are enhanced to textured reliefs.

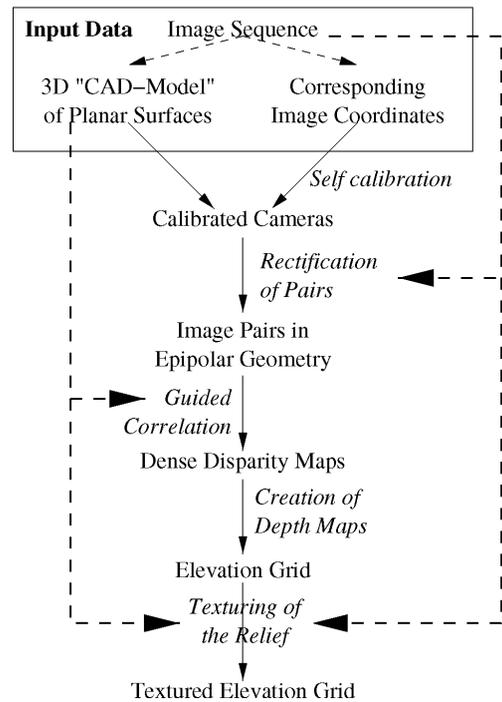


Figure 2: Overview of the processing chain from a coarse to a refined model.

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 input 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 3D coordinates may be given only as rough estimates, as their precise location can be determined via bundle adjustment through the measured image coordinates. A set of images is needed for both relief reconstruction and texturing. They are linked to the polygons through known image coordinates of the vertices. The process includes the determination of the camera parameters using the 3D coordinates. 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 such a special case 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 otherwise, 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



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

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

$$\mathbf{x}_i = \mathbf{P}\mathbf{X}_i \quad (1)$$

Using the vector cross product

$$\mathbf{x}_i \times \mathbf{P}\mathbf{X}_i = 0 \quad (2)$$

Eq. 1 can be transformed into an equivalent equation

$$\mathbf{A}\mathbf{p} = 0 \quad (3)$$

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} . This transform is known as *direct linear transform* (DLT) which yields

an initial solution without knowing any parameters beforehand. Refer to (Hartley and Zisserman, 2000) for details.

An initial solution of \mathbf{P} is computed separately 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, 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. Their initial values are set to the mean values from all retrieved initial camera matrices. The initial value for radial distortion is set to 0. 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.

For the case that the initial values of the 3D coordinates are only rough estimates, a gauge free bundle adjustment that handles both object coordinates and camera parameters as unknowns has been used. 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.

Rectification is carried out for each pair of images that shall be used for depth estimation. 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. Thus, the scale of the rectified image is close to the original and the facade is least distorted (see Fig. 4 for an example).

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. Dynamic programming 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 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) , the cumulated costs $C_{i,j}$ for the optimal solution ending in (i, j) are defined by

$$C_{i,j} := \min_{i' \in P_{i,j}} C_{i',j-1} + C_{i,j}^0 \quad (4)$$

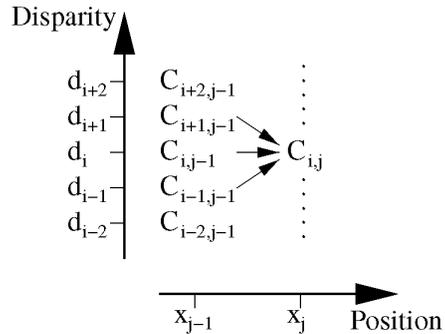


Figure 7: Detail of the cost matrix.



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

where $P_{i,j}$ is the set of allowed predecessor rows of (i, j) and $C_{i,j}^0$ 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.

2.4.2 Constraints Two 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, i. e. any disparity would be valid. 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. A globally optimal path would be found this way.

Here, the disparities at both ends are known because they can be computed from the intersection of the known polygon borders with the scan lines. This allows to define 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.

(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 neighboring pixels to no more than ± 1 . Thus we define the set of valid predecessors in Eq. 4 as

$$P_{i,j} := \{i' : |i' - i| \leq 1\} \quad (5)$$



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

2.4.3 Costs In addition to the constraints that limit the solution by exclusion of illegal results, cost function have been applied to keep the solution close to an optimum. In contrast to constraints, costs require additional parameters to define the strength of their influence. Four different cost functions have been used.

(a) Similarity The basic cost term $C_{i,j}^0$ in Eq. 4 has been set to

$$C_{i,j}^0 := 1 - r_{j,j+i} \quad (6)$$

where $r_{j,j+i}$ is the cross correlation coefficient between the two image locations (i is the disparity, i. e. an offset to the one-dimensional position j). This expresses that if two image patches are similar, they are likely to be locations of corresponding features.

(b) Image contrast If there is only low contrast in the x -direction, the correlation coefficient will not change much for a small change in disparity. In order to strengthen the influence of the correlation at strong vertical edges, $C_{i,j}^0$ is weighted proportional to the absolute value of the gradient along the scan lines. The gradient is computed via a Sobel kernel which carries out discrete smoothing with a Gaussian and differentiation:

$$C_{i,j}^{0'} := (1 - r_{j,j+i}) / |\text{grad}_x(i, j)| \quad (7)$$

(c) Neighboring scan lines One of the problems with a scan line oriented approach is how to deal with neighboring 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 $C_{i,j}^p$ had been introduced to keep disparity values on one line close to those of the preceding line which extends the previous definition of $C_{i,j}^0$:

$$C_{i_0,j}^{0''} := (1 - r_{j,j+i_0}) / \text{grad}_x(i_0, j) + C_{i_0,j}^p \quad (8)$$

$$C_{i_0,j}^p := c_1 |i_0 - i_{-1}| \quad (9)$$

where i_0 refers 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 neighboring lines. Since there exists a closed polygon of initial values, there are always reference disparities available.

Adjacent pixels in one scan line Similarly, disparity can not change arbitrarily between two neighboring pixels within one line. This has already been taken care of by the ordering constraint. But since there already exists an approximation for the facade, the disparity change can be predicted and any discrepancy penalized. Eq. 9 can be extended as

$$C_{i_0,j}^p := c_1 |i_0 - i_{-1}| + c_2 |d_{pred} - (i' - i)| \quad (10)$$

where d_{pred} is the predicted disparity and $i' - i$ the disparity value for the current cell. Note that this is independent from absolute disparity values, i. e. it is only assumed that the facade is parallel to the plane of the bounding polygon.

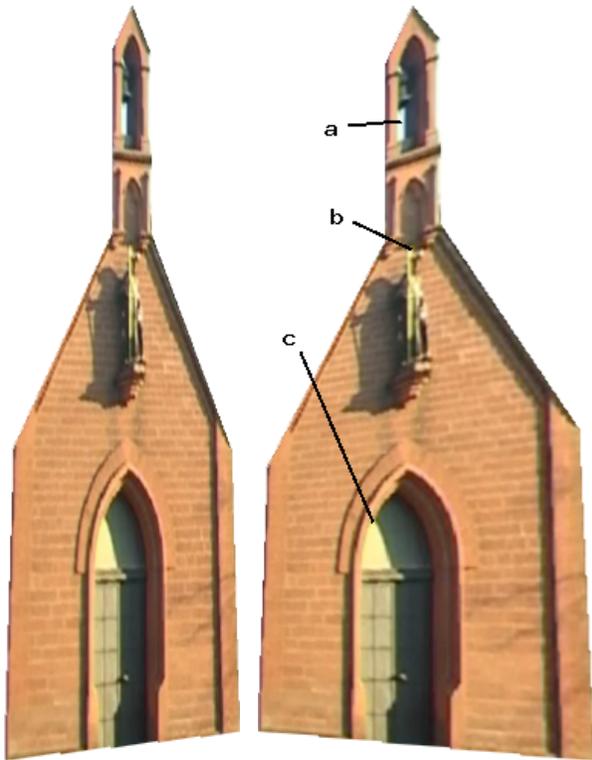


Figure 5: Two artificial views onto the west facade of Alexiuskapelle with flat surface model. See text for explanation of details: (a) sky, (b) statue, (c) shadow.

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 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 overlaid 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



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

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).

Generation of flat facade model For means of comparison, a flat model has been generated (see Fig. 5). Different views on this model are all created by projective transforms of the original image, i. e. the relative position of pixels does not change.

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 facade to 6 m. The 3D-coordinates of the control points have got a standard deviation of 14 cm.

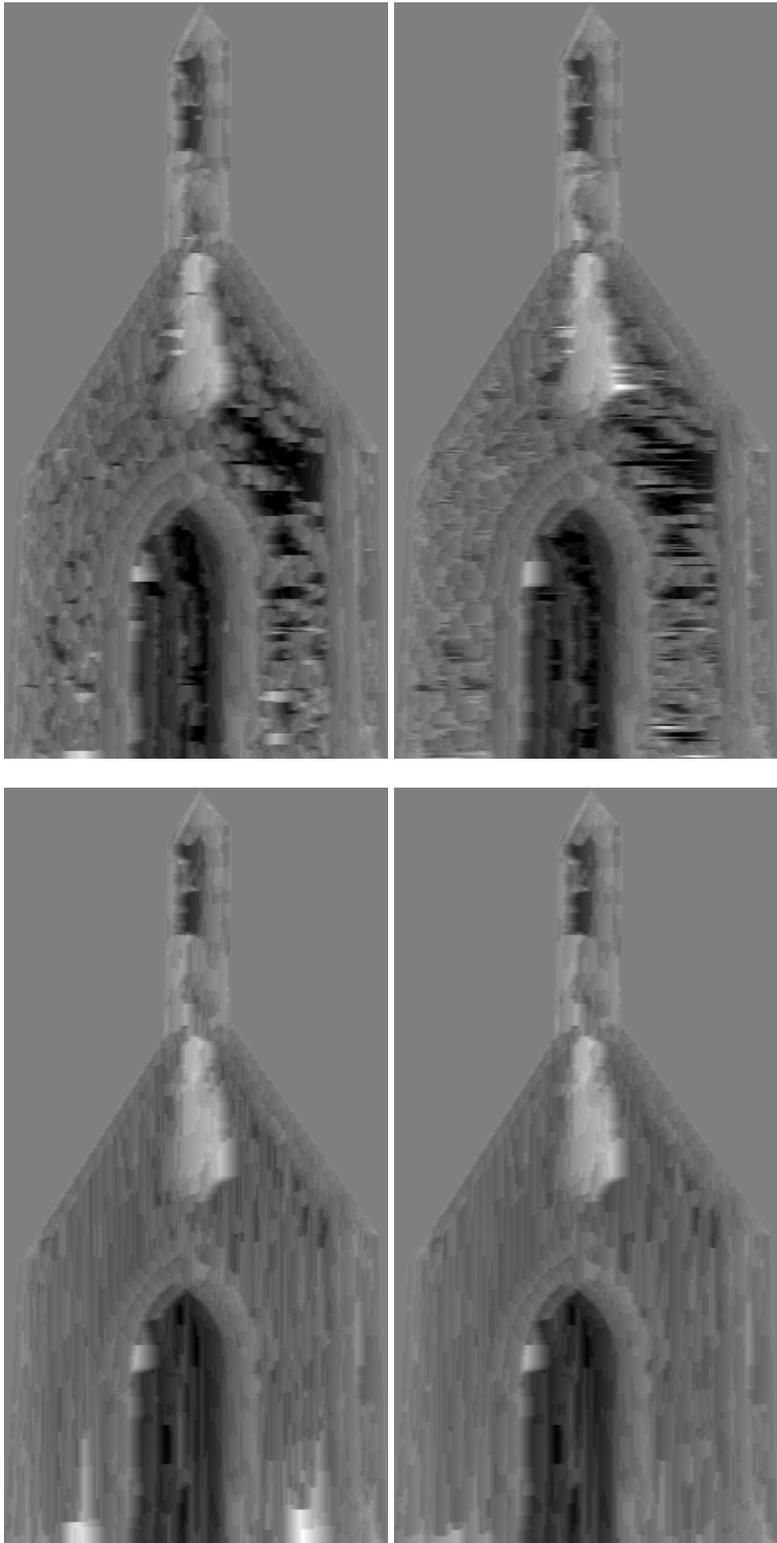


Figure 9: Influence of cost parameters. See text for explanations.

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.

Rectification 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.

Relief recovery The influence of the different parameters is shown in Fig. 9. Bright areas stand out from the facade whereas dark areas designate holes. In the upper left, no improvements have been made through any parameter. The main structures as the portal and the statue above it can be seen. However, to the upper right of the facade, there are some gross errors where a hole has been detected. There are also some locations where smaller regions stand out from the facade. In the upper right, costs for adjacent pixels in rows have been applied as well as the use of gradient information. The situation is improved only a bit. Since both corrections are applied to scan lines only, i. e. the direction in which the dynamic programming already tries to find an optimal solution, it can be concluded that the criteria used do not lead to a completely different solutions.

In the lower left, the influence of neighboring rows has been accounted for. The results are vastly improved, even though at the bottom two gross errors appear as bright regions. Obviously, the improvements stem from rows now no longer being independent of each other. A combination of all cost parameters has been used for the result in the lower right. The increase in quality compared to the unimproved version is mainly due to the linkage of neighboring scan lines, but the two cross errors could be eliminated through the other parameters. One gross error is visible in all four examples at the left part of the door.

A model with the with the recovered relief is shown in (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 compared 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.

The influence of the different parameters has been demonstrated through examples. It could be seen that the purely row based dynamic programming approach could be improved best through the linkage of neighboring scan lines. Minor errors that remain can be eliminated to some extent by the other parameters. The results are not too sensitive to the parameter setting, but a good choice has to be found by some test calculations.

Errors 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. The influence of different parameter settings had been demonstrated and showed that especially a connection of neighboring rows is important. The results 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.

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.

v. Hansen, W., Thoennessen, U. and Stilla, U., n.d. Detailed relief modeling of building facades from video sequences. In: IAPRS, Vol. XXXV, part B3, Istanbul, 2004.