A hybrid concept for 3D building acquisition

A. Brunn, E. Gülch, F. Lang, W. Förstner

Institut für Photogrammetrie, Universität Bonn
Nußallee 15, D 53115 Bonn, Germany

This paper presents a hybrid concept of interaction between scene and sensors for image interpretation. We present a strategy for 3D building acquisition which combines different approaches based on different levels of description and different sensors: the detection of regions of interest, and the automatic and semiautomatic reconstruction of object parts and complete buildings.

Key words: interpretation model, sensor fusion, object detection, digital surface models, 3D grouping

1 Introduction

Cities are the living place for more than 50% of the world population. Up to date 3D information on buildings appears to be urgently necessary for urban management and all types of the planning referring to sound emissions, air pollution, microclimatology or transmitter planning. Current maps mostly contain only the groundplan of buildings together with reference to various registers. Acquiring 3D information on buildings still is costly, only automatic or at least semi-automatic methods appear feasible in the long run. This paper discusses strategies for 3D building acquisition from various sensor data, mainly, however, from aerial images.

Building acquisition has to face a number of problems, which altogether appear as a challenge for automatic data interpretation. Some of these problems are:

[⋆] This research is supported in part by BMBF/ DLR under Grant 01 M 3018 F/6, by BMBF/ DARA GmbH under Grant 50 TT 9536, and by DFG under Grant 180-1/2. The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, expressed or implied of those agencies.

- The *notion* "building" has many facets, economical, legal, architectural, social, technical ones. Sensor data only provide access to the geometry and the physics of a building in principle. Even with this restriction there is no commonly accepted and enough formalized model of a building which can be used in automatic interpretation. In Braun et al. (1995) we argued that a volumetric representation may be the right level of representation in order to facilitate a link to the meaning of the individual parts.
- There are various *sensor* yielded information on buildings. Images, especially aerial images, certainly are a main data source. Laser scanners are on the verge of becoming an economical alternative if only geometry is of interest. But also maps or geoinformation systems can be regarded as sensors providing links to all types of registers and the corresponding semantics.

  It is, however, by no means clear, which of these sensor scales and resolutions has which use in building acquisition. This has conceptual consequences on which aggregation level a building actually can be observed, is thus intimately linked with the complex aggregation hierarchy of buildings and their parts.
- *Sensor data* are projections, thus only show a specific portion of the scene. This projection leads to various defects, starting with the missing depth in images or GIS, occlusions both in images and laser data. The problem of automatically transferring 3D information about buildings to 2D-constraints has not yet been solved in a general manner, only specific solutions are knows, e.g. on invariants.
- Finally, strategies for inverting the imaging process, inverting the projection are lacking. Many tools are available,which need to be adapted or evaluated with respect to their use in building extraction. However, this does not solve the problem really, as the impact of each of these tools on the solution remains unclear. General tools such as optimization, constraint satisfaction or heuristic search only provide a shell which needs to be filled.

The goal of this paper is to discuss the aspect of processing within automatic interpretation more in depth, clarifying the role of algorithms as a link between sensors and scene and in this way earning the description of certain strategies. A more detailed description can be found in Brunn et al. (1996).

2 Layers, models and strategy

2.1 Levels of description for data and models

We assume image analysis to be a task driven process deriving an application dependent description of the scene based on available sensor data. Both, scene and sensor data have a complex structure. Scenes are composed of objects, giving rise to a containment hierarchy for composite objects, to a specialization hierarchy for groups of objects and associations between objects for describing their mutual relations. Sensor data are complex, as we treat any type of derived sensor information also as sensor data, thus subsuming intelligent sensor systems under the notion sensor.[1] For discussing image analysis tasks we unify scene and sensor model as both can contain known and unknown parts. This allows to describe tasks, like detection, reconstruction, localization and interpretation on the one hand, and orientation and calibration on the other hand, in a uniform manner (Förstner, 1989).

An image analysis algorithm provides a link between a scene and a sensor (cf. fig. 1). The scene provides the algorithm with the necessary and available knowledge, e. g. the class of the object to be detected, the structure of the object to be reconstructed, the representation in which the location is to be given. The sensor provides the algorithm with an iconic or symbolic description of the image on the level of aggregation adequate for the algorithm, and possibly information on its orientation, or other characteristics. The result of the algorithm generally is both, the desired completion of the scene description as well as of the sensor description.

We now want to break down scene and sensor models. In the most simple case the scene may consist of objects of different type. Each of it may be described by its containment and aggregation hierarchy. The scene related task and the type of the unknown object task and type are lists, whereas the internal structure is a tree or a more complicated representation structure. Any task, type of object and internal structure may be combined leading to a specific scene related task with specific preknowledge fixing a point on the vertical axis. The task specifies which of the objects or structures are unknown.

The situation is a bit more complicated for the sensors. Besides the list of the tasks, the list of the physical sensors and the internal structure of the possibly aggregated sensor information we need to distinguish between 2D and 3D, or even 4D, information derived from sensor data of lower dimensionality by matching or tracking algorithms. Also here, any combination of task, aggregation level, dimension or type of physical sensor may lead to a specific sensor related task with specific sensor information fixing a point on the horizontal axis.

The extremely high number of possible image analysis algorithms results from the pure fact, that in principle all combinations between scene and sensor related tasks, types, structure or dimensionality may be reasonable. The rectangle spanned by scenes and sensors (cf. fig. 1) thus qualitatively represents the space of possible vision algorithms. This assumes *single* algorithms for

[1] At this point we integrate the model for the physical sensor and the derived 2D or 3D image, thus subsuming the image model under the sensor model.

each task, and does not take into account multiple goals or sensor fusion techniques. Therefore strategies for selecting appropriate algorithms and especially sequences of algorithms reveal to be necessary.

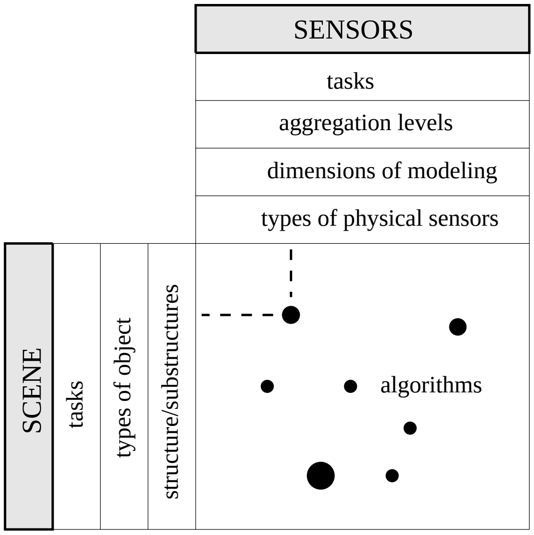

Fig. 1. Tasks and structure of scene and sensor information. The points symbolize some algorithms. The different sizes indicate different gain and usability of the combinations.

Data driven and goal driven processes generally need to be intimately linked. Aggregation, grouping or hypothesis generation tasks which necessarily need to be initiated and controled by high level knowledge, task or scene specific.

The problem of control therefore has to solve several subtasks.

(i) Choose the best sensor for a given task.
(ii) Choose the best sensor combination for a given task (sensor fusion).
(iii) Choose the best hypothesis for a substructure of the scene for a given sensor (hypothesis generation).
(iv) Choose the best hypothesis for a substructure of a scene for a given scene interpretation (internal reasoning).
(v) Decide on the completion of the task or detect failures, and give control to the operator (algorithm internal decision).

The proposed scheme can be interpreted as a structured blackboard having different layers:

a) the layered structure of the scene referring to different levels of aggregation or abstraction
b) the layered structure of the sensors referring to different levels of aggregation
c) the layered structure of the algorithms referring to the different layers of aggregation.

Obviously any vision algorithm which is meant to be used by the system needs documentation with respect to its use.

This general scheme needs to be realized in steps. The problem of building acquisition seems to be of a complex enough nature to test the validity of the setup. We therefore will specialize and give some examples in the following.

2.2 Selected strategy for building acquisition

We describe a special strategy for 3D building acquisition that is based on our experiences with the applicability of different data sources and different algorithms for this specific task. Because reconstruction using aerial image data can not be solved in one step we split the task of 3D building acquisition into different subtasks using different types of sensor information. The subtasks are

- detection of regions of interest,
- 3D reconstruction by object parts, and
- 3D reconstruction of complete buildings.

This corresponds to a multiple usage of different paths in fig. 1. The output of one subtask will be the input to the next subtask. Up to now the strategy is fixed in order to exploit their usefulness. Other strategies need to be evaluated in order to be able to adaptively choose an optimal sequence depending on sensor data, quality of algorithm and specific task to be solved.

In the first step we use pixel based 2.5D sensor data from automatically generated Digital Surface Models or laser scanners for extracting those regions which contain the desired objects (buildings) with a high probability. In this subtask buildings are searched in a high generalization level represented by low level pixel attributes (e.g. height). For these extracted building segments approximate structures of buildings can be derived depending on the resolution of the available 2.5D sensor data. The resulting segments of the elevation model are used as input for the following 3D reconstruction by object parts. This second step additionally uses a feature extraction of local structures in multiple monochrome or multichannel images. The automatically derived 3D structures are the input for the 3D reconstruction of complete buildings. This last step can be formulated on a lower aggregation level because of the strong impact of the integrated human operator even with only one stereo pair. During the third step prior results can be verified, corrected and improved. This step will be required for quite some time as a fully automatic acquisition of complex buildings can not be achieved due to the high variability of buildings. The amount of human interaction depends on the required detail of acquisition, the complexity of the buildings, and the availability and quality of sensor data. Partial solution neglecting one or more steps of our proposed procedure may be sufficient for special requirements. The algorithms of these three subtasks are described in the next chapters.

3 Detection

Both methods described in the following sections need a priori focusing on a region of interest (ROI) to reduce the search space. We apply a hierarchical focusing techniques for grid based data. In addition to the presentation of a realization for focusing in this chapter we want to show that statistical influence diagrams as a specialization of Bayesian networks can be used for low level, pixel based image interpretation.

Several focusing techniques have already been presented in the literature on image processing and interpretation, each using special data sources. In contrast to those methods using image pyramids (Ackermann and Hahn, 1991), and others using a scale space algorithm (Ravela et al., 1996), we follow a strategy based on statistical classification to find ROIs. Statistical methods support the fusion of different data sources and integrate uncertainty and errorness of observed data. Especially Bayesian networks supply a close connection to pyramidal image interpretation as will be shown below. Hierarchical approaches for segmentation tasks reduce the complexity of search (Terzopoulos, 1986). The algorithm we use differs from complex approaches (e.g. Bouman and Liu (1991)) because we use a simpler, more straight forward propagation to reduce the computational effort.

3.1 General strategy

The focusing procedure consists of two steps: the generation of a feature pyramid and the goal-oriented evaluation of the branches of a deduced pyramid of random variables. Let a feature vector field $\boldsymbol{f}(r,c), r \in \{0, \dots R-1\}, c \in \{0, \dots C-1\}$ with n observations in each vector be given, that should hold information about the interest of the pixels, which is coded in the assigned feature vectors. During the preprocessing we build up an image pyramid using techniques from multichannel image processing (e.g. Lee et al. (1992)). Let be $\boldsymbol{f}_0(r,c) := \boldsymbol{f}(r,c)$ then follows

$$\boldsymbol{f}_{l+1}(r,c) = \boldsymbol{M} \odot \boldsymbol{f}_l(2r,2c) \quad \text{with} \quad l \in \{0, \dots, L-1\} \tag{1}$$

when L is the number of layers. The amount of grid points decreases by the factor $\frac{1}{4}$ in each reduction step. $\boldsymbol{M}$ is an information preserving multidimensional filter. We have tested several morphological and statistical filters. Which filter should be applied depends on the specific problem, i.e. the type and the amount of noise in the data.

After creating the feature pyramid we start the evaluation of the feature vectors. We use every feature vector to determine whether the grid position is of interest for the specific task. For this purpose we interpret the feature vectors as observations of a measuring process and assume that – due to the central limit theorem– they are normally distributed. Because it is not possible to estimate any statistical parameter from one observation, we take the observation as representative of the mean value and use a covariance matrix given a priori.

$$\underline{\boldsymbol{f}}(r,c) \sim N(\boldsymbol{f}(r,c), \boldsymbol{C}) \tag{2}$$

At this point we start to build a dynamical generated pyramid of random variables $v_l(r,c)$. Each random variable can hold the two statements “the grid point is member of a ROI” and “the grid point is not member of a ROI” for each pixel in each layer. The **Pr**obability that a grid point is member of a region **O**f **I**nterest (**PrOI**) will be denoted with $p(v_l(r,c)=t)$ where 't' means true, $p(v_l(r,c)=f)$ resp. It is calculated by the probability that a feature vector belongs to a ROI Θ in the object space a priori defined.

$$p(v_l(r,c)=t|\Theta) = \frac{1}{\sqrt{2\pi}\det \boldsymbol{C}} \int\limits_{\Theta} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{f}_{l(r,c)})\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{f}_{l(r,c)})} \quad d\boldsymbol{\theta} \tag{3}$$

These boundary information of Θ can be obtained by estimation from a training set. We use a multidimensional open box with the edges $\boldsymbol{f}_b = (f_{bi}), i \in \{0 \ldots n-1\}$ and infinity as subspace Θ. That means that high values of the feature vector are more interesting than low ones. If all observations in the feature vector are uncorrelated, the integral can be calculated as a product from the normal distributions of each observation. The probability of the complementary outcome is calculated from eq. (3) as the complementary probability.

Starting from the highest pyramid level we perform a depth-first search to obtain the ROI. We evaluate all grid points on the L-th level and sort them in descending order $p(v_l(r,c)_{max}=t), \quad \ldots, \quad p(v_l(r,c)_{min}=t)$. In the next level we start with those pixels that are linked to the most promising element of the highest level. This is performed for all lower levels to come. Except in the highest level, we use two pieces of information, both, the actual information at the current level and the information aggregated in the previous level, which takes the conditional probabilities into account (cf. fig. 2). We neglect the correlations between the two sources due to the feature pyramid generation. Therefore, we have to distinguish between the two PrOIs, the first based on

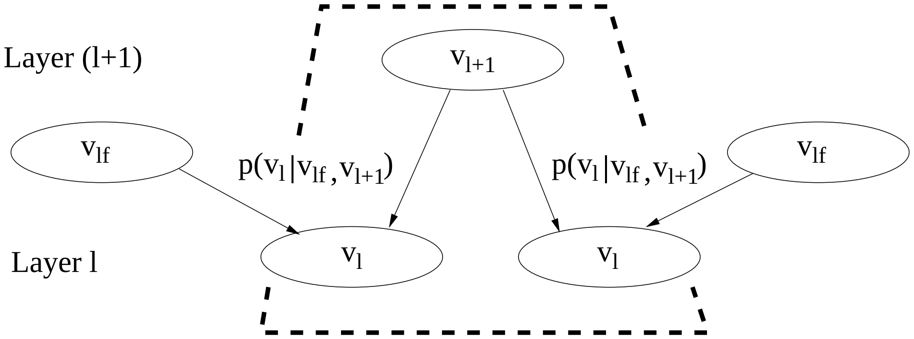

Fig. 2. Propagation of the probabilities of interest in the pyramid.

the information provided only by the feature vector $p(v_{lf}(r,c))$ and the second on the posteriori PrOI $p(v_{l+1}(r,c))$ at each level of the pyramid. In this refined notation holds

$$p(v_l(r,c)=t) \propto \sum_{v_{lf}(r,c),v_{l+1}([r/2],[c/2])} p(v_l(r,c)|v_{l+1}([r/2],[c/2]),v_{lf}(r,c)) \\ p(v_{l+1}([r/2],[c/2]))p(v_{lf}(r,c)) \quad (4)$$

when $[x]$ means the highest integer that is smaller or equal x. We do not introduce special application dependent knowledge in the conditional probabilities: when both information sources support each other, we strengthen that content, otherwise we delay the decision.

This procedure is done for each level of the pyramid up to a given pixel resolution depending on the specific task. The first pixel extracted by this depth-first search defines a first element of the ROIs and can be used as a focus for the next step of interpretation (cf. sec. 4). By backtracking in the pyramid other elements of ROIs are found. Up to this moment we have not made experiences, where to stop backtracking because classification against a ROI cannot be done at higher levels. Fine structures might appear in lower levels depending on the filter generating the feature pyramid. Only when using a maximum filter this decision can be done on a higher level, assuming the noise to be negligible.

This algorithm for focusing onto the relevant parts of data, in particular the evaluation step, can be interpreted as a dynamical Bayesian network (Pearl, 1988). Each pixel with its probability of interest represents a random variable in the Bayesian network and the conditional probabilities are associated with the edges of the graph. This configuration makes two extensions obvious: on the one hand grid points that are far apart should influence one another by incorporating more model knowledge about neighborhood relations between ROIs. On the other hand, if one region is a priori known it should influence the PrOI of related regions. Thus solving data restoration and data interpretation in a common algorithm becomes possible. That would mean bottom-up

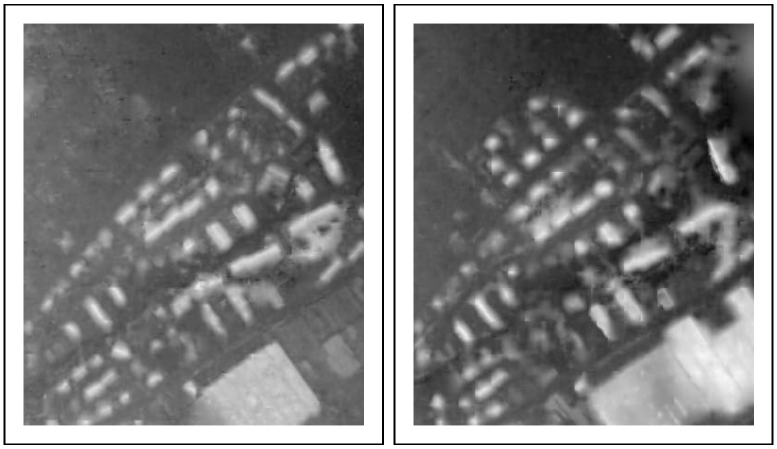

Fig. 3. a) DSM of 1982; b) DSM of 1993; For visualization the two DSMs are spread on 256 grey values. (Bright/Dark pixel mean large/low height.)

propagation in the notion of Bayesian networks. Bayesian networks seem to be a feasible tool for hierarchical image interpretation, not only for grouping as Sarkar and Boyer (1993) and Koch (1995) showed.

3.2 Using Digital Surface Models (DSMs) for focusing

The aggregation of the dataset for focusing should be very low cost or fully automatically to enlarge its advantage. So we use DSMs for focusing either directly scanned with an airborne laser scanner or automatically generated using correlation or matching techniques.

For focussing done by **change detection** we used two DSMs from an area in Alfter-Oedekoven. Both were derived by image matching. One dated back to 1982, the other to 1993. The ground resolution was $2m \times 2m$ (cf. fig. 3). The images[2] with an image scale 1:12000 do not allow the generation with a finer grid with more surface information. We use the height difference as the feature for determining the ROIs to update a map from the earlier datum. By experiences made previously, we define a height variance $\sigma_h = 2m$ for the statistical distribution. The ROI in the feature space for the height differences is defined as $f_b = 4m$ which is motivated from the usual height of a building storey. In fig. 4 the results are shown. Dark regions are classified as regions of change. For the purpose of visualization we show all pixels of the presented layers. Because of the coarse DSM grid it is not possible to stop at a higher level than the lowest one. The result shows some blurring due to effects from the automatic generation of the DSMs and to the changed vegetation of eleven years. To improve the detection, further data have to be integrated, especially

[2] The images were kindly provided by the LVA NRW (Bonn, Bad Godesberg, Germany).

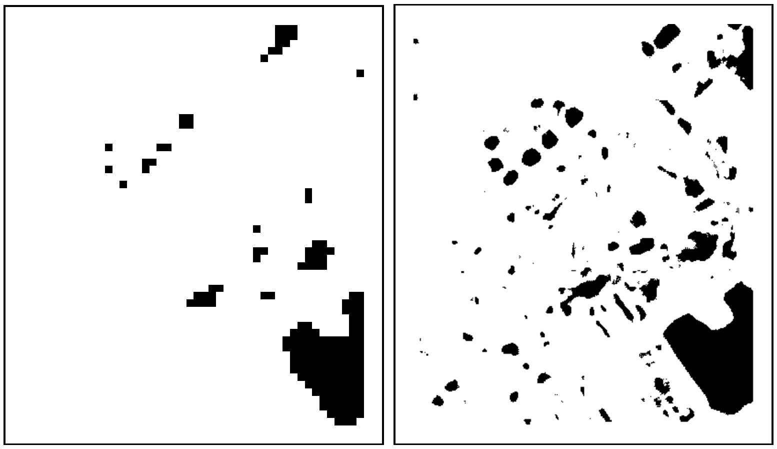

Fig. 4. a) Third layer and b) lowest layer of the classification pyramid. Small disturbances due to changed vegetation and due to the automatic DSM generating process exist in the resulting classification image.

information on color. In contrast to a simple thresholding of the difference DSM, this method reduces random noise of the difference DSM, because of the used neighborhood information from previous layers. Thus it reduces splitting of segments. Furthermore the depth-first search leads the following 3D acquisition to the ROI in descending order of interest; dominant buildings can be reconstructed at first. At the moment those extracted regions are classified by the building extraction technique of Weidner and Förstner (1995). In future work we want to improve the detection in context of building extraction by extending the feature vector and introducing color and texture observations in the focusing step.

4 3D Reconstruction by Object Parts

In this chapter we present a procedure for fully-automatic 3D reconstruction which infers local 3D structures in space from its observed local 2D structures in multiple images. The result of the low-level pixel based image interpretation is used to restrict the aerial image to small regions of interest, which serve as input for the subsequent process of 3D reconstruction. The orientation data of the images are assumed to be known, leading to geometric restriction during the matching procedure.

4.1 General aspects

The following aspects are decisive for characterizing our procedure :

Feature extraction and feature aggregation: Based on a polymorphic feature extraction (cf. (Förstner, 1994)) we derive a rich symbolic image description consisting of attributed features F^{2D}, namely points, lines and regions together with their mutual relations, that are contained in a feature adjacency graph (FAG). Analyzing the relational image description, we are able to derive sets of basic aggregates A^{2D} by a bottom up process. These are point, line and region induced structures namely vertices V^{2D}, wings W^{2D}, cells Z^{2D} containing all neighbors and possibly indirect neighbouring features These basic aggregates serve as starting point for our reconstruction.

Object Modeling: For introducing scene knowledge into the reconstruction process, the 3D object model must be placed at the same representation level like the image aggregates A^{2D} which we propose for reconstruction. Transferring the 2D aggregates A^{2D} into 3D aggregates A^{3D} results in a local boundary representation of object parts P of three different semantic classes, namely corners C, edges E and faces F. They are components of the hierarchically structured 3D building model which we proposed in (Braun et al., 1995). For reconstruction we especially use object parts of class corner. Corners are specialized into different subclasses. Each corner is described by the corner point, several lines and planar faces. The partitioning of corners is described by their line attribution, giving by the semantic labels (h), (v) or (o). This description is further refined by distinguishing vertical and oblique lines due to their slope into (v+), (v-), (o+) and (o-). Finally, different geometric constraints between the corner components like e. g. symmetry and orthogonality are considered. We also model corner pairs which eases the connection and grouping of corners to a complete surface. All these semantically attributed subclasses are collected in the set Ω_C.

4.2 Procedure and Results

As for vertex structures V^{2D}, being the projections of 3D vertices V^{3D} into the images, the imaging geometry gives strong restrictions during the correspondence analysis, we focus on the 3D reconstruction of corners C^{3D}, with corners being interpreted vertices. The reconstruction is done by following the *hypothesize and verify* paradigm:

Construction of Corner Hypotheses: Starting point for the analysis are the 2D vertices V^{2D}. In the first step we generate 3D vertices: Selected tuples of corresponding vertex structures form the basis for the transition to 3D-vertices V^{3D}. The correspondence analysis is performed by heuristically selecting a suitable sequence of vertices using the epipolar geometry and evaluating the structural similarity of matching candidates. Relational matching of the features F^{2D}, which describe the vertex structures of the correspondence tuple leads to a 3D vertex.

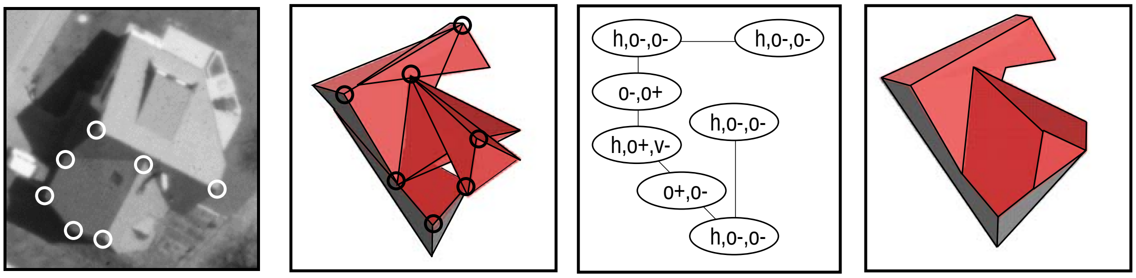

Fig. 5. shows the result of the automatic 3D reconstruction: **a.** original image; **b.** partially reconstructed roof in 3D composed by different corners, which are denoted by circles; **c.** connections between corner pairs given as a graph with the nodes being classified corners; **d.** result of a 3D grouping process, which connects the reconstructed 3D corners based on the corner graph.

Establishing corner hypotheses uses the corner model in the following way. We interpret the generated 3D vertices V^{3D} by a first classification, leading to admissible corner subclasses Ω_C. First, the line attributation is used to obtain a subset Ω'_C of Ω_C. For each element of Ω'_C the geometric constraints, associated by the model, yield a further reduced $\Omega''_\mathsf{C} \subset \Omega'_\mathsf{C}$, each element in Ω''_C being an admissible interpretation of the 3D vertex. This way we obtain several corner hypotheses c^{3D}. This principle of domain reduction of possible interpretations is also applied for the generation of corner pairs (c_i^{3D}, c_j^{3D}) leading to a corner graph.

Verification of Corner Hypotheses: As the corner classification for each 3D vertex generally will be ambiguous, we perform a second rigorous classification by statistical analysis. This is an optimization problem for finding the best interpretation of the data by a maximum likelihood parameter estimation. The functional model describes the relation between observations and the unknown parameters for the unconstrained corner. Additionally hard and soft constraints, given by the corner class ω_C, are introduced. The result of the estimations are evaluated corner interpretations including their geometric constraints, which are fundamental for a geometrically improved and verified corner reconstruction.

Figure 5 shows an example of the fully automatic 3D reconstruction. Further details of the corner reconstruction approach can be found in (Lang and Förstner, 1996b).

Connection to overall strategy: The corner reconstruction conceptually works in parallel on all vertices found in the images and guarantees local consistency. Nevertheless, the information achieved so far may be incomplete, as some corners or mutual relations between corners may not be found. Performing a subsequent aggregation, using the part-of hierarchy of our object model (cf. (Fischer et al., 1997)), we are able to link to the modeling tools from CAD, which e. g. forms the basis for our 3D reconstruction of complete objects. Finally we can reach global consistency of the result by a parameter estimation of volumetric object primitives as presented in (Lang and Förstner, 1996a).

5 3D Reconstruction of Complete Buildings

The human operator is on the top level of our strategy and is required to finalize, correct or validate the reconstruction of complete buildings as the methods above may not produce a sufficient level of detail or may produce incomplete results. To ensure reliability and completeness we propose *semi-automatic procedures*, related to early work done at SRI (Quam and Strat, 1991). The system developed by (Lang and Schickler, 1993) has been extended in cooperation with the Inst. of Computer Science III, University Bonn (Englert and Gülch, 1996), using the Constructive Solid Geometry principle (Hoffmann, 1989) for the 3D modeling of complex buildings. The system does not require stereo-viewing and is such suitable also for non-photogrammetrists.

5.1 Semi-automatic system and connection to overall strategy

In the semi-automatic approach the operator is providing the interpretation step, supported by various automated modules given the orientation data of the images.

The operator may zoom down into one aerial image and *focus* his interest on a particular building. This step can be replaced by the automated detection procedure described above.

The *modeling phase* is performed by a semi-automatic form and pose adaptation of 3D models in *one* image, i.e. of volumetric primitives following the CSG principle. By boolean operations of this primitives very complex building structures can be reconstructed, if required. During the adaptation process a 3D visualization of the building is used to check for completeness and correctness. The in-heritage of parameters allows rapid acquisition in areas with similar building types. A precise docking of primitives is enabled by matching and gluing lines and faces to already instantiated primitives.

One homologous point has to be measured in the images in order to compute 3D world coordinates. The measurement of homologous points for *single* primitives can be replaced by an automated, robust model to image(s) matching procedure (Läbe and Ellenbeck, 1996) or by a successful reconstruction of parts.

The output of the whole process are composites of 3D volumetric primitives with 3D world coordinates. In addition to the geometric acquisition a module for the extraction of texture from one or more images provides fully automatic photo-realistic texture mapping on the 3D object surfaces for visualization and animation.

5.2 Experimental results with the semi-automatic system

Three projects have been performed with standard and non-standard situations. The given average gross acquisition times contain the time for modeling, the local and global navigation and organization.

Two projects have been performed with stereo pairs of digital aerial photographs. In the *sub-urban area Oedekoven* 5499 volumetric primitives (90 secs per primitive) have been extracted. The gross time is about 25% shorter compared to an earlier version of the system. The higher performance is combined with a more detailed building acquisition and with texture extraction. 371 primitives in the *center of Rostock*, have been acquired by a non-photogrammetrist after four hours training only. The time per primitive is only 150 secs and such not even factor 2 higher as for trained users.

In *downtown Frankfurt/Main* 549 primitives (112 secs per primitive) where extracted from *one* tilted aerial photograph and known ground heights only, which demonstrates the flexibility of this system. From the aerial image, texture was extracted automatically and mapped to the 3D faces. A subset of the model is given in figure 6 which demonstrates the animation potential of photo-realistic 3D data.

The *accuracy* of the system depends on many factors, like image scale, selection and measurement of models or the generalization level. All results so far indicate a performance which corresponds to analytical photogrammetric methods. An extended description of the empirical tests can be found in (Gülch, 1996).

6 Discussion and Conclusions

The paper wants to discuss strategies for 3D-building acquisitions by giving a frame of reference for describing the role of algorithms and for making control tasks explicit. Three examples illustrate what type of problems we encounter when automating image interpretation for building acquisition.

It seems useful to discuss the role of the three examples within the complete - though by no means completed - process.

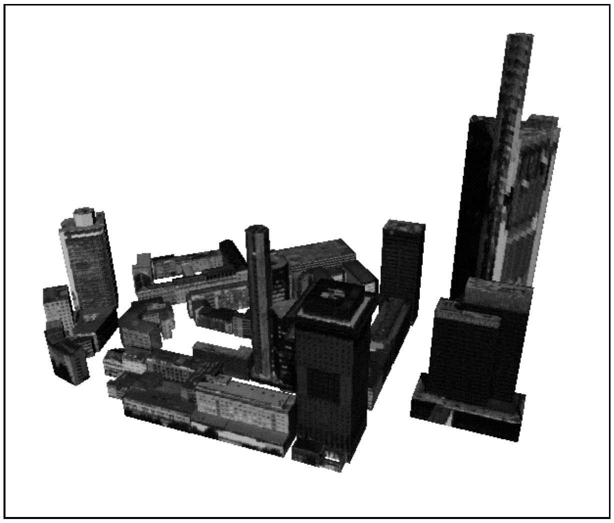

Fig. 6. Part of 3D building model FRANKFURT with automatically provided texture

(i) *Focussing* on areas of interest necessarily requires low level, i.e. raster models of the objects to be found. Thus the appearance of buildings in the raster data needs to be available. The use of Bayesian nets for modelling the context, stored in the higher levels of a feature pyramid, seems to be promising. Focussing thus is interpreted as generating hypotheses on regions for which a likelihood is given, which can be used for controlling the next steps.

(ii) *3D-reconstruction* is shown to be feasible on the mid- and the high level, as geometric and semantic models are integrated into the matching process. Again the result are hypotheses on corners of buildings. Their likelihood can be used in the grouping process. Obviously the verification step which uses a classification procedure integrates geometric knowledge, information on defects of the feature extraction and semantic knowledge and requires an optimal geometric reconstruction as a prerequisite.

(iii) Semiautomatic procedures will be necessary as long as the variety of buildings is larger than covered by the models used in the automatic procedures. The building models used in the two previous examples are too coarse and not general enough. Semiautomation always faces the problem to reasonably divide the work between computer and operator, here by performing the measurement tasks to the computer and leaving the decision steps to the operator. The CSG-modelling has shown to be quite effective. The achieved acquisition times always need to be seen as reference for automatic procedures, which claim to be operational.

The chosen strategy in all three cases was an interpretation of bottom-up with top-down processes: grouping hypotheses were generated data driven, these hypotheses were then verified using higher level models, where the final verification was left to the operator.

This report on ongoing work leaves enough open problems. Without going into the details of the described algorithms only two should be mentioned:

- The current exploration of finding good algorithms for investing the observability of building features by various sensor data should be continued. Their performance should be clearly documented in order to make these algorithm available to other research groups.
- Only after this stage different strategies, i.e. sequences of processes/algorithms can be compared in case these sequences are not meant to be fixed. The comparison then may refer to both, the theoretical performance, i.e. the performance prediction, and the empirical performance. Here standard tests are helpful.

As a consequence of these open problems we can expect progress to be slow but steady. This avoids unnecessary hurry and allows time to discuss the real interesting questions, why algorithms actually work.

References

Ackermann, F. and M. Hahn (1991). Image pyramids for digital photogrammetry. In H. Ebner, Fritsch (Ed.), *Digital Photogrammetric Systems.* Wichmann-Verlag.

Bouman, C. and B. Liu (1991, February). Multiple resolution segmentation of textured images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-13*(2), 99–113.

Braun, C., T. Kolbe, F. Lang, W. Schickler, V. Steinhage, A. Cremers, W. Förstner, and L. Plümer (1995). Models for photogrammetric building reconstruction. *Computer & Graphics 19*(1).

Brunn, A., E. Gülch, F. Lang, and W. Förstner (1996). A multi-layer strategy for 3d building acquisition. In *Proceedings IAPR/TC-7 Workshop.*

Englert, R. and E. Gülch (1996). One-eye stereo system for the acquisition of complex 3D building descriptions. *GIS 9*(4).

Fischer, A., T. Kolbe, and F. Lang (1997). Integration of 2d and 3d reasoning for building reconstruction using a generic hierarchical model. In *submitted to: SMATI '97, Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps.*

Förstner, W. (1989). Image analysis techniques for digital photogrammetry. In *Proc. of 42nd Photogrammetric Week, Stuttgart, Sept. 1989.* Schriftenreihe des Instituts für Photogrammetrie, Heft 13.

Förstner, W. (1994). A framework for low level feature extraction. In J.-O. Eklundh (Ed.), *Computer Vision - ECCV '94, Vol. II*, pp. 383–394. Lecture Notes in Computer Science, 801, Springer-Verlag.

Gülch, E. (1996). Extraction of 3d objects from aerial photographs. In *Proceedings COST UCE ACTION C4 Workshop 'Information systems and* *processes for urban civil engineering applications', Rome, Italy, November 21-22.*

Hoffmann, C. (1989). *Geometric and Solid Modeling.* Morgan Kaufmann, Palo Alto, CA, USA.

Koch, K. R. (1995, Juni). Bildinterpretation mit Hilfe eines Bayes-Netzes. *Zeitschrift für Vermessungswesen 120*(6), 277–285.

Läbe, T. and K.-H. Ellenbeck (1996). 3D-wireframe models as ground control points for the automatic exterior orientation. In *Proceedings ISPRS Congress, Comm. II, Vienna, IAP Vol. XXXI.*

Lang, F. and W. Förstner (1996a). 3D-city modelling with a digital one-eye-stereo system. In *ISPRS Congress, Comm. IV, Vienna.*

Lang, F. and W. Förstner (1996b). Surface reconstruction of man-made objects using polymorphic mid-level features and generic scene knowledge. *Zeitschrift für Photogrammetrie und Fernerkundung 6*, 193–201.

Lang, F. and W. Schickler (1993). Semiautomatische 3D-Gebäudeerfassung aus digitalen Bildern. *Zeitschrift für Photogrammetrie und Fernerkundung 5*, 193–200.

Lee, T. S., D. Mumford, and A. Yuille (1992). Texture segmentation by minimizing vector-valued energy functionals: The coupled membrane model. In *Computer Vision – ECCV '92, Proceedings*, pp. 165–173. Springer.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers.

Quam, L. and T. Strat (1991). SRI image understanding research in cartographic feature extraction. In H. Ebner, D. Fritsch, and C. Heipke (Eds.), *Digital Photogrammetric Systems*, pp. 111–121. Wichmann, Karlsruhe.

Ravela, S., R. Manmatha, and E. M. Riseman (1996). Image retrieval using scale-space matching. In B. Buxton and R. Cipolla (Eds.), *Computer Vision-ECCV'96*, pp. 273–282.

Sarkar, S. and K. Boyer (1993, Mar.). Integration, Inference, and Management of Spatial Information Using Bayesian Networks: Perceptual Organization. *IEEE Transactions on Pattern Analysis & Machine Intelligence 15*(3), 256–274.

Terzopoulos, D. (1986). Image analysis using multigrid relaxation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*(2), 129–139.

Weidner, U. and W. Förstner (1995). Towards automatic building extraction from high resolution digital elevation models. *ISPRS Journal 50*(4), 38–49.