A Scene Analysis System for the Generation of 3–D Models

Oliver Grau
University of Hannover, Institut fuer Theoretische Nachrichtentechnik
und Informationsverarbeitung. D–30167 Hannover (Germany), Appelstr. 9A,
E–Mail:grau@tnt.uni–hannover.de

Abstract

A scene analysis system for the 3–D modeling of objects is presented. It combines surface reconstruction techniques with object recognition for the generation of 3–D models for computer graphic applications. The system permits the insertion of highlevel constraints, like a specific angle between two house walls, in an explicit knowledge base implemented as a semantic net. The applicability of those constraints is proved by asserting and testing hypotheses in an interpretation phase. In the case of rejection a more general constraint or model is selected. The capabilities of the system were shown for the modeling of buildings using depth from stereo and contour information. The system reconstructs the surface of the scene objects using the constraints selected in the prior interpretation.

1 Introduction

For the visualization of virtual environments, for example in applications like flight and driving simulators, architecture and landscape planning, there exists a high demand for photorealistic 3–D models of real objects. The manual construction of these models with CAD systems is expensive and fails to reach photo realism. Consequently the automatic or semi–automatic generation of 3–D models from images is currently of high interest. The presented system is designed for the reconstruction of the object shape from digital images. After the restoration of the 3–D geometry the texture and color of the objects is taken from the original input images and will be stored in a texture map. With this approach photorealistic models can be obtained.

The approach described in this paper uses stereo image pairs as the input for the reconstruction of the 3–D object surface. In a first step the scene depth is estimated pointwise for these image pairs. In a second step the depth values are integrated into a 3–D surface description. Both processing steps need additional constraints to be well conditioned.

The pointwise computation of the depth employs a blockmatching–based depth estimator under the assumption of piecewise continuous surfaces [2]. The resulting depth maps are usable but noisy for the considered outdoor scenes (see fig. 3 a). Each depth value represents a 3–D point relative to the camera coordinate system.

For the integration of the 3–D points into a consistent surface description a suited surface model must be selected. A commonly used method is the interpolation of the 3–D points with a spline function. This general surface model is sufficient for natural objects like persons [3]. For objects with a more special geometry, like buildings, this surface model is not sufficient because the spline surface smoothes edges and the noise in the depth maps produces a rough surface even in object parts that are planar. For those objects it is necessary to select a more special surface model, like a planar polygon. Further the relations between the object parts, like symmetries or known intersection angels are very useful constraints for the reconstruction of the object surface.

The contribution presented here uses the scene understanding system AIDA as a framework for the knowledge representation and the basic control structure of the interpretation. AIDA is further applied to the modeling of landscapes from arial images [5][6].

The approach discussed here adaptively selects suitable constraints during an interpretation phase. The constraints are represented in a generic, exchangeable knowledge base. The system can then be easily adapted to new object classes. After a successful symbolic interpretation of the scene content the derived constraints and the data (depth values, contours) are used for the reconstruction of the 3–D surface.

The following section gives a system overview. Section 3 gives a short description of the image processing methods. Section 4 gives an overview of the interpretation and used knowledge representation scheme. Section 5 describes the data fusion in the surface reconstruction. The paper finishes with a presentation of the results and conclusions.

Accepted paper for: International Conference on Recent Advances in 3–D Digital Imaging and Modeling .
May 12–15, 1997, Ottawa, Ontario Canada.

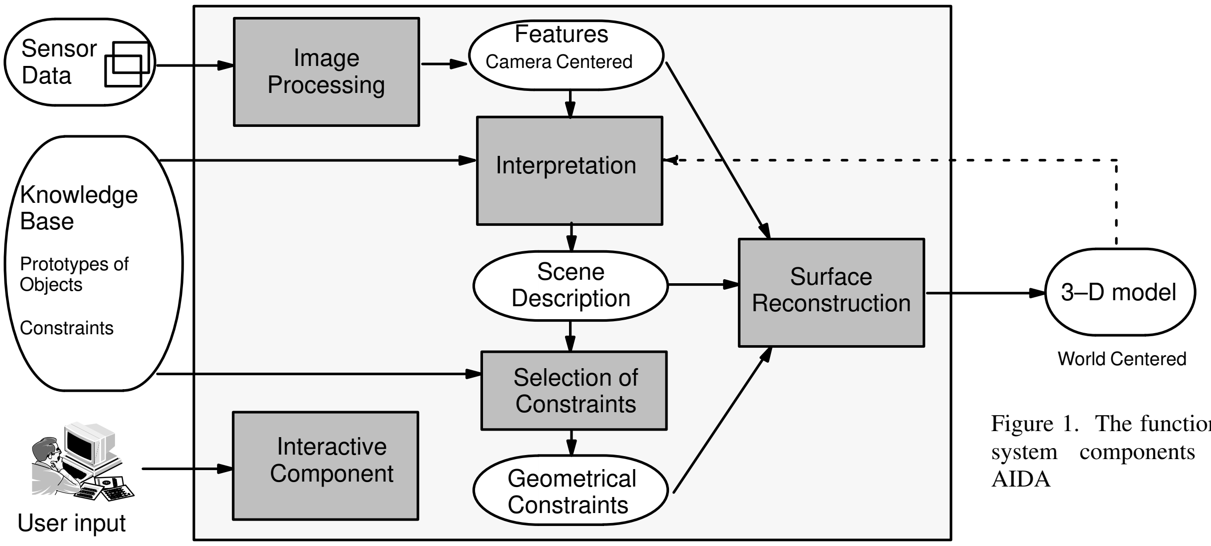


Figure 1. The functional system components of AIDA

2 System Overview

Figure 1 shows the main functional components of the presented system AIDA. Input data are sequences of stereo image pairs. The data is processed sequentially, i.e. at each cycle one stereo pair is presented at the input. The images pass through several image processing modules where various features are extracted (contours, regions, depth from stereo) and grouped to primitives (2–D edges and 2–D polygons). The features are camera centered at this stage.

Figure 2: a) left input image of stereo input image pair

The interpretation module groups and labels the extracted primitives and creates a semantic description of the scene in terms of the generic prototype description found in the knowledge base. Section 4 gives an overview about this component and the knowledge representation.

The *selection of constraints* module selects additional geometrical constraints from the knowledge base that are suitable for recovering the object surface.

The surface reconstruction described in section 5 performs several tasks: First it transforms new camera oriented primitives extracted from the image processing pipeline into world coordinates and inserts them into the 3–D model data base. The transformation requires the knowledge of the coordinates of the input stereo camera system in a common world coordinate system, which are determined through view point estimation. The surface reconstruction employs depth information from new camera views and the geometrical constraints selected from the knowledge base to further improve already present surface patches created from the prior stereo image pairs.

With the interactive component the user of the system has control over the system. The most important role of the interactive component is the development of the generic prototype model description in the knowledge base. The quality of the description determines the performance of the system's recognition facilities and can be improved by an interactive analysis of intermediate results. The interactive component visualizes the processing steps and explains the results of the recognition process.

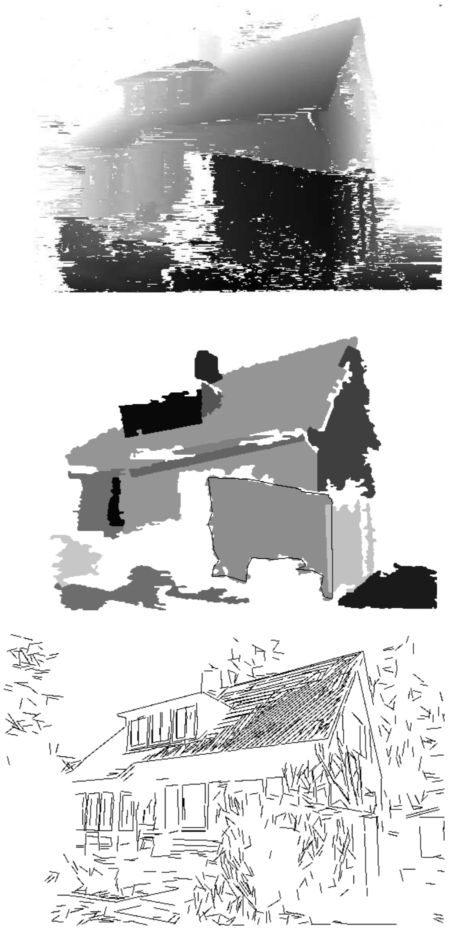
3 Image Processing

The image processing modules extract various features from the input images, including depth maps. The most important visual cue for the depth estimation is depth from binocular stereo. For the considered application this method turned out to be a reliable and robust technique to recover the depth information from camera input images [2][3]. The input images were taken with a pair of identical photo cameras mounted on a bar 1 m apart (example in figure 2).

In a first pre–processing step the stereoscopic camera system is calibrated using a regular pattern of control points. The

Accepted paper for: International Conference on Recent Advances in 3–D Digital Imaging and Modeling .
May 12–15, 1997, Ottawa, Ontario Canada.

calibration estimates the intrinsic and relative extrinsic camera parameters. The intrinsic parameters contain the radial distortion of the lenses and the focal length. The relative extrinsic parameters cover the orientation of both cameras relative to each other. With this information the image pair is rectified to achieve epipolar geometry. In the next step a disparity map is calculated using stereoscopic correspondence analysis. The camera parameters are used for the calculation of depth values from the disparity estimation (depicted in figure 3a).



a) color coded depth map

b) result of the segmentation

c) scene conturs

Figure 3: Image processing results

The depth maps are used as input for the segmentation of the scene (figure 3b). The segmentation is based on a region growing algorithm and groups 3–D points computed form the depth values that have a small distance to a plane in 3–D. The plane parameters are computed iteratively from all points in the already found region. After the update the region growing algorithm is called again until terminated by a stop condition. The basic algorithm is adapted from [4] and considers the problems of the noisy and discrete depth maps computed with the used approach.

In addition to the estimation of the depth maps and regions, contours are extracted from the luminance images using an edge detector and a line segment approximator (figure 3c). All features or primitives are camera centered at this processing stage.

4 Scene Interpretation

A lot of scene interpretation systems have been developed in the past. They can be classified by several criteria:

a) The *knowledge representation*: Rules are found in SPAM [8], production rules and semantic nets in MESSIE [11][12], semantic nets in ERNEST [13].

b) The *application*: Many systems have been developed for the interpretation of aerial images [8][9][11]. These systems handle rather 2–D data than 3–D structures. Knowledge based systems for close range 3–D problems have different requirements. First approaches are discussed in: [1][12].

c) The *structure* of data and system and *control strategy*: In SPAM [8] four interpretation phases group features on different levels of abstraction. Foresti et al. [10] describe a system with six layers of abstraction. In ERNEST, which was designed for several different applications, different levels of abstraction are represented with a special construct in the knowledge representation and do not require any modifications in the structure of the system. This idea was taken up in AIDA.

While nearly all systems try to cope with the unreliable results of low level vision modules, most can not properly handle 3–D information or do not provide the flexibility to cope with new visual cues. The main problem with 3–D data is the handling of occlusions in the different camera views. To meet these requirements the interpretation system AIDA was developed.

4.1 The Knowledge Representation

In AIDA the knowledge about the scene to be modeled is represented explicitly as a semantic net. For the description of objects and their relations a problem independent net language was defined which mainly resembles the net syntax of ERNEST [13].

Each node of the generic prototype net represents an object and is called concept. Concepts are depicted as boxes in figure 4. The nodes in the scene description are called instances and represent real objects. They are depicted as ellipses in the following figures. The interpretation starts with a copy of the prototype net and creates and inserts those instance nodes into the net that match to the signal.

Each node can contain several attributes like photometric or geometric features. These attributes are employed to test the compatibility of a concept with an instance.

The nodes of the net are connected via links. The most important links are: *instance–of, part–of, cdpart–of, is–a* and *concrete–of*. The *instance–of* link connects instances with their prototype concept. The *part–of* and *cdpart–of* links allow a decomposition of complex structures. The *cdpart–of* (context dependant part of) joins concepts that can not exist on there own, like the concept *Junction* that is only defined in the context of a wall.

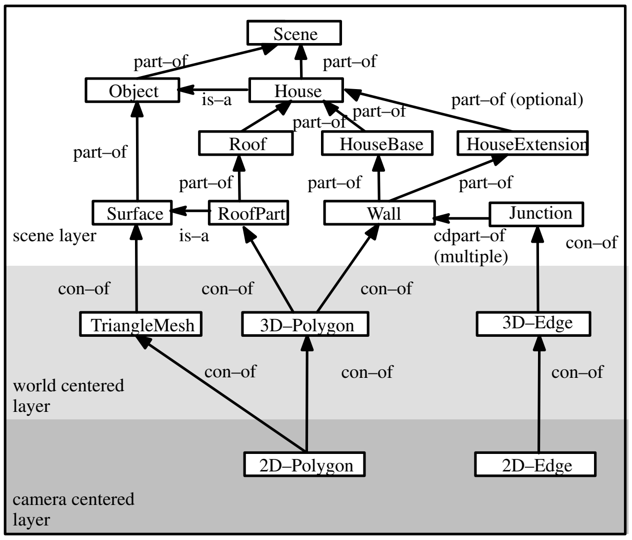


Figure 4: Simplified part of the knowledge base.

The link *concrete–of* has an important role. It links two levels of different abstractions. The knowledge base is structured into three layers: The scene layer, the world centered layer and the camera centered layer. For example the *concrete–of* link assigns, as depicted in figure 4, the concept *Wall* from the most abstract scene layer to its concretization *3D–Polygon* in the world centered layer and to *2D–Polygon* in the camera centered layer. The nodes in the camera centered layer are related to the signal and contain attributes that are directly measured in the image or image primitives.

The *is–a* link implements inheritance and provides the specialization of the object description. In figure 4 the node *Scene* contains the parts *Object* and the specialization *House*. The interpretation module will consider both nodes as competeting hypotheses. In the case an object can be detected as a house more special knowledge can be applied in the latter surface recognition.

Nodes in the prototype net can be conneted to more than one child node, like different parts. Further some nodes in the prototype net have more that one parent node, for example the node *Wall* which has the nodes *HouseBase* and *HouseExtension* as parents. The quantity of obligatory and optional connections (links) is a parameter of the corresponding link in the prototype network. In figure 4 the *HouseExtension* is connected as an optional part.

In the scene description net the number of connections of a node to its parent is usually exactly one. An instance of the concept *Wall* for example has exactly one connection to the appropriate node *House*. For the concept *Junction* in figure 4 a multiple binding is enabled in the *cdpart–of* link from *Junction* to the concept *Wall*. This setting causes the interpretation to search for two parents of an instance of *Junction* during the interpretation and joins these (see figure 9).

4.2 The Interpretation Module

The control of the interpretation is implemented in a rule–based inference engine. The different rules are activated at corresponding stages of the interpretation phase. In [13] a description of the most important rules is given. In AIDA rules for the handling of multiple bindings and topological relations were further added. In the following the principles and the progress of the interpretation is illustrated using the simplified knowledge base depicted in figure 4.

The goal of scene interpretation is the assignment of a meaning to each primitive found in the input image. The used primitives are regions and contours (as depicted in figure 3b+c). For the description of the scene the interpretation successively builds up a network of instance nodes.

At the beginning of the interpretation an initial hypothesis instance is created from a concept of the knowledge base. A hypothesis instance is a state of a node. The possible states of nodes are: Concept, hypothesis instance, partial instance and complete instance. During the interpretation nodes usually change their state from hypothesis instance to complete instance.

The position of the concept of the initial hypothesis instance in the net determines the control flow of the interpretation: If the concept is located in the top of the knowledge base it enforces a top–down analysis. In the case that the concept is located in the bottom or in the middle of the net a bottom–up or mixed (bottom–up and top–down) analysis is enforced. For the modeling of buildings a top–down analysis is chosen. Therefore a hypothesis instance of the concept *Scene* is created at the beginning of the interpretation (step 0 in figure 5).

All instances in a scene description have unique names. The naming convention is to extend the name of the concept with a unique integer number, as depicted in the following figures.

Starting from the initial hypothesis *Scene–1* the interpretation successively creates the parts found in the prototype net. In the steps 1–2 a hypothesis instance *House–1* and its obligatory sub–part *HouseBase–1* are created. In this top down flow the ranges of attributes of the nodes are propagated. For the computation of the initial parameter range of an attribute an optional computation function can be specified for a node. If the function is omitted the range value found in the node concept is copied into the attribute range of the hypothesis instance. The nodes *House–1*, *HouseBase–1* and *Roof–1* contain an attribute ”Height”. The range is copied from the prototype and contains value ranges that are typical for the considered buildings.

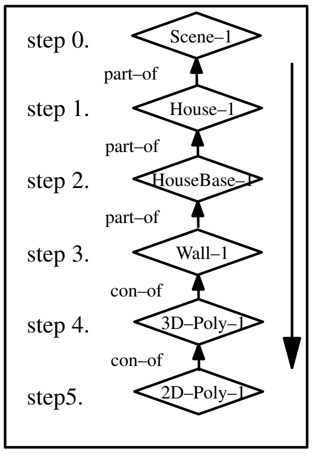


Figure 5: Top–down propagation

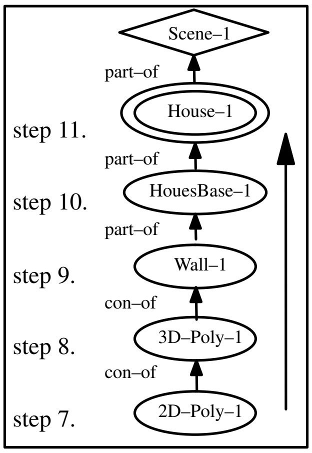


Figure 6: Bottom–up propagation

In the step 3 an instance *Wall–1* is created. As depicted in figure 4 the concept *Wall* is connected to *HouseBase* (containing the mandatory walls or the house) and *HouseExtension*. The concept *Wall* also contains a ”Height” attribute. The parameter range is computed from the parent node. So the range can be adapted to the apriori knowledge.

In steps 4+5 the concretes of the node *Wall–1* are expanded and the parameter values of the expectation are propagated as for the above nodes. It is at this stage that a relation to the signal can be established by searching a region in the label image (figure 3b) that matches the expectations propagated so far (size and shape). Due to the fact that the propagated parameter space restrictions are not very strict the search function of *2D–Poly–1* returns a list of regions that match the parameter range. For each region a complete copy of the current content of the scene description is made. Each copy of the current scene description is handled as a search node. The control of the interpretation computes a judgement value for each search node and selects the search node with the best judgement as the current interpretation state. The judgement is computed using all judgements of the nodes contained in the search node. The node judgements itselves are based on the attributes and the links of the nodes.

With the complex judgement of the complete scene description the selection of the correct region for *2D–Poly–1* can be made more reliable. Further the control is able to recognize a wrong decision of the interpretation: If a binding of a region was incorrect the interpretation is able to step back to an earlier search node and to retry the interpretation from this starting point with a binding to another region. The process of the selection of the search node employs a modified A* algorithm that considers both certainties and costs of an interpretation.

After the binding of data in step 6 (not depicted) the state of the hypothesis instance *2D–Poly–1* is changed into a complete instance (step 7 in figure 6). Complete instances are depicted as ellipses. In the steps 7–10 the state of the parent nodes of 2D–Poly–1 are also changed to complete instances. Furthermore the attribute values can be computed now. The attribute values are computed recursively from the child nodes, which are closer related to the signal, in a bottom–up propagation.

At step 11 the state of *House–1* can not be changed into a complete instance because there are obligatory parts missing. The node becomes the intermediate state partial instance instead– this allows the recalculation of attribute values.

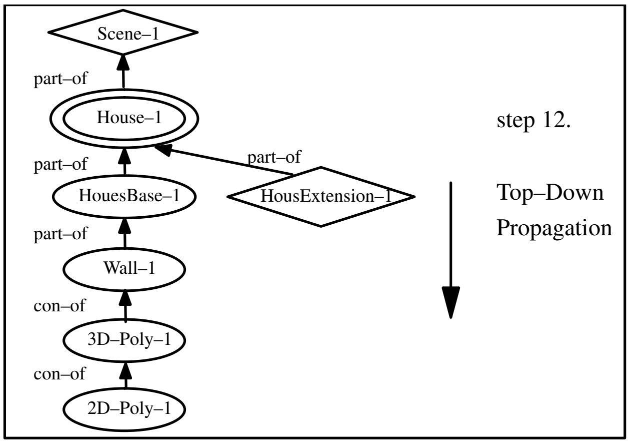


Figure 7: Top–down propagation of updated attribute values

In step 12 (depicted in figure 7) the updated attribute values are used to restrict the attribute ranges of the node *House-Extension–1*. The attribute range for ”Height” can be computed from the informations of the node *House–1* that contains the signal based value form the first wall found. So the search range for the house extension can be estimated more precisely.

After the creation of more than one instances of the concept *Wall*, hypothesis instances of the concept *Junction* are created (not depicted). In a top–down propagation the already found knowledge about the scene is used for a restricted search of the corresponding 2–D line segments: The expectation of the corresponding 3–D edge is calculated by the intersection between the 3–D polygons corresponding to the wall instances. The 2–D edge is computed from the 3–D edge using the camera model.

The interpretation end if there are no further hypothesis instances that could be verified or if a user specified concept was instantiated. The second condition is true if all obligatory parts of the concept are found.

5 Surface Reconstruction

The surface reconstruction [7] integrates the depth measurements into a consistent 3–D model employing the constraints generated during the interpretation. The module restores the 3–D surfaces found in the scene description using all the assigned data that is useful for the reconstruction. In addition to the data found in the depth map and the contours the model selection (3–D polygon or triangular mesh) is directly found in the scene description. The knowledge that a wall is a polygon is in fact a strong and important constraint for the surface reconstruction. For each point in a segmented region the 3–D coordinates are computed from the depth map (figure 3a). All points within the region contribute to an overdetermined equation system. It is solved by a least square minimization to find the 3–D plane which belongs to the region. The boundaries of the polygons are determined by intersection of the found plane with neighboring polygon planes, as described in section 5.2.

5.1 Used Constraints and Data

The considered input data for the surface reconstruction is: (i) the regions of planar polygons in space found in the segmentation (section 3) and (ii) their bounding edges found as grey level contours in the input images.

All points within a labeled region of (i) are reprojected into 3–D space using the depth value and are integrated to a single plane measurement E_m. The plane is determined by a weighted regression with optional certainty factors of the depth estimation serving as the weights.

New camera view points can be integrated easily because the determination equations of the regression can be superposed (see section 4.3).

Both the surface reconstruction and the view point estimation (section 5.3) use a uniform representation of the input data. The measured data from the image processing is handled like the additional constraints derived during the scene interpretation. Therefore the measured contours and depth values are compared to the model.

The measured plane E_m and the bounding edges in the input image refer to the same polygon in 3–D space and can be formulated as cost functions c_{ji} that measure the fit to the model:

c_{1i} : the difference between pose and orientation of the polygonal description from the segmentation and the depth values to the polygonal model,

c_{2i} : the distance between model and corresponding contour edges.

These cost functions that correspond to data outside the model are called external constraint in the following. On the other hand internal constraints are not directly related to the image data. The following are used in this approach:

c_{3i} : The angle between two polygons,
c_{4i} : parallel polygon edges,
c_{5i} : equal length of edges and
c_{6i} : equal angles (between 3 or more polygons).

The use of the internal constraints is optional but very helpful in cases of uncertain or noisy depth measurements.

5.2 Integration into a Consistent 3–D Model

The surface reconstruction integrates the plane and edge measurements together with additional internal constraints into a consistent polygonal 3–D model. The boundings of the polygons are obtained by intersecting neighboring 3–D planes:

$$L_k = \text{Intersect}\ (E_i, E_{ik})$$

The different measurements (plane, edge and constraints) produce a conflicting description due to noise in the image and the depth map. This inconsistency is solved by a numerical optimization that minimizes the objects overall cost function f_{glob}:

$$f_{glob} = \Sigma\Sigma\ w_{ji} \cdot c_{ji}(\mathbf{p}) \ \text{–> Min.} \qquad (1)$$

Each constraint is represented by it's cost function $c_{ji}(\mathbf{p})$ in this equation. The parameter vector **p** is varied during the numerical minimization and contains the current object geometry, i.e. all polygons that describe the object surface. The factor w_i weights the influence of each constraint. It strongly depends on the application.

The minimization of the global cost function leads to a surface description which best meets both the measured information and the constraints derived from the knowledge base. The surface reconstruction progresses incrementally.

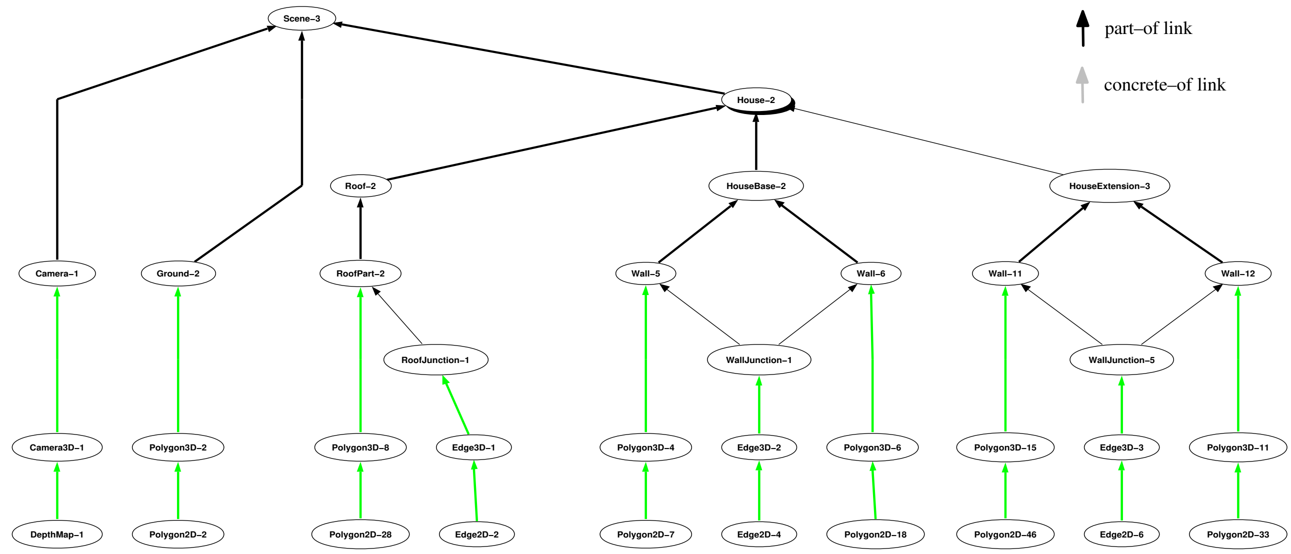


Figure 9: Part of the scene description net of the scene depicted in figure 2

Hence it is possible to include new data like depth values from new camera views and new surface patches from prior invisible object parts into the 3–D model.

5.3 View Point Estimation

For the integration of new camera views the camera position and orientation must be known. For practical tasks like the modeling of outdoor scenes the parameters are usually unknown and must been estimated.

The view point estimation takes into consideration the extracted features of a new camera image and tries to find a set of camera parameters (location and orientation) that optimally fits the already generated 3–D model. In case of planar polygons found in the new camera view the correspondence to the model is assessed by measuring the difference between normal vectors. Additionally the match between the model's projected 3–D edges and the edges found in the image is measured. The view point estimation uses the same minimization of the global cost function f_{glob} as described for the surface reconstruction in equation (1). Here the parameter vector **p** contains the camera pose and orientation rather than the model geometry which is fixed. Precondition for the view point estimation is that a major part of the already modeled object is visible in the new camera view.

The correspondence of object parts in the image and the already modeled 3–D object shall be established by the interpretation. At the current state of the system implementation this step is guided mainly manual.

6 Results

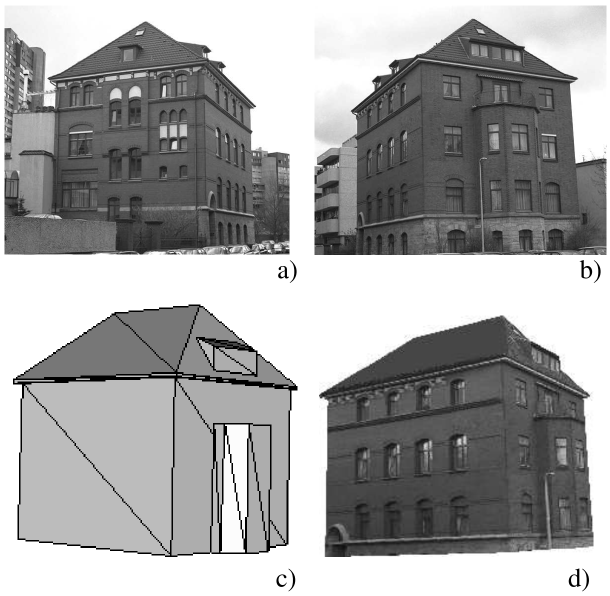


Figure 8: Building modeled from two camera viewpoints a+b. c) and d) show synthesized views of the resulting 3–D model in shaded and textured representation

Figure 9 shows a part of the scene description as the result of the interpretation of the scene depicted in figure 2. The mantatory parts of the house (roof and walls) were found very reliable. In the following the house extensions and the dormer and chimney (both not depicted in figure 9) were found correctly using geometrical and topological features.

Figure 8 shows the result of modeling an outdoor scene. It uses the data extracted in the image processing and additional constraints found in the interpretation. The model is created using two different camera views. For the integration the view point estimation was used. The neccessary correspondance between the views were given manually.

7 Conclusions

The scene interpretation system AIDA was presented, which combines recognition and reconstruction facilities. With the proposed knowledge representation it is possible to formulate explicitly properties of the objects that are to be analyzed. The interpretation module plays a central role. It selects the applicable models and constraints useful for the surface reconstruction. The most important advantage of the approach is the explicit knowledge which describes the objects to be modeled separately from the system implementation. Thus only the description of the objects has to be exchanged to apply the system to a new object class. The performance of the system has been demonstrated for the modeling of buildings. The adaptation to other objects like furniture should be easy.

The system easily allows the introduction of highly specialized high level constraints, like a specific angle between two walls. The applicability of those constraints is proved by asserting and testing hypotheses. In the case of rejection a more general constraint or model is selected.

The surface reconstruction module integrates the different data types (depth maps and conturs) and the additional constraints using a numerical optimization. The optimization is robust and able to cope with a portion of bad assigned data or constraints.

For the considered application of modeling buildings the approach leads to accurate models as depicted in figure 8. Due to the use of polygons the surfaces are planar and edges are straight. Quality of photo realism is reached that satisfies most requirements of computer graphic applications. The critical structures for the modeling are considered in the knowledge base. This is the main difference compared to data driven approaches where the surface approximation is made by a triangular mesh that doesn't consider the semantics.

8 References

[1] Braun, C., Kolbe, Th., Lang, F., Schickler, Steinhage, Cremers, Förstner, W., Plumer, L., "On the Models for Photogrammetric Building Reconstruction". Computer & Graphics, 1995.

[2] L. Falkenhagen, "Depth Estimation from Stereoscopic Image Pairs Assuming Piecewise Continuos Surfaces", Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production. Nov. 1994 Hamburg. ISBN 3–540–19947–0, Springer.

[3] Koch, R., "3–D Surface Reconstruction from Stereoscopic Image Sequences", ICCV'95, Boston, June 1995.

[4] Leonardis, A., Gupta, A., Bajcsy, R., "Segmentation of Range Images as the Search for Geometric Parametric Models", Int. Jour. of Computer Vision, Vol. 14, No.3, April 1995.

[5] O. Grau, R. Tönjes: "Knowledge Based modeling of Natural Scenes". European Workshop on Conbined Real and Synthetic Image Processing for Broadcast and Video Production. Nov. 1994 Hamburg.

[6] R.Tönjes: "Control of Scene Reconstruction Using Explicit Knowledge", IEEE Workshop on Applications of Computer Vision, Sarasota, Florida, Dec. 1996.

[7] S.Weik, O.Grau, "Recovering 3–D Object Geometry using a Generic Constraint Description", ISPRS96 – 18th Congr. of the Int. Soc. for Photogrammetry and Remote Sensing. July 9–19 1996 Vienna

[8] McKeown, D. M. Jr., Harvey, W. A. Jr., McDermott, J., "Rule–Based Interpretation of Aerial Imagery", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI–7, No. 5, pp. 570–585, Sept. 1985.

[9] Matsuyama, T., Hwang, V.S.–S., "SIGMA : A Knowledge–Based Aerial Image Understanding System", *Plenum Press*, New York 1990

[10] Foresti, G.L., Murino, V., Regazzoni, C. S., Vernazza, G., "Distributed spatial reasoning for multisensory image interpretation", Signal Processing 32, May 1993, pp. 217–255.

[11] Clément, V., Giraudon, G., Houzelle, S., Sandakly, F., "Interpretation of Remotely Sensed Images in a Context of Multisensor Fusion Using a Multispecialist Architecture", IEEE Trans. on Geoscience and Remote Sensing, Vol. 31, No. 4, July 1993

[12] Sandakly, F., Giraudon, G., "Multispecialist System for 3D Scene Analysis", ECAI 94. 11th European Conference on Artificial Intelligence

[13] Niemann, H., Sagerer, G., Schröder, S., Kummert, F., "ERNEST: A Semantic Network System for Pattern Understanding", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 9, pp. 883–905, Sept. 1990.