Knowledge based system for the interpretation of complex scenes

C.-E. Liedtke, J. Bückner, M. Pahl, O. Stahlhut
Institut für Theoretische Nachrichtentechnik (TNT), University of Hannover, Germany

ABSTRACT: Future tasks in remote sensing require automated and robust analysis methods for image data from airborne and satellite sensor platforms. The scenes to be observed exhibit a high degree of complexity. Complexity refers to the large variety of pictorial representations of objects with the same semantic meaning and also to the extensive amount of details of the scenes, i.e. the components which make up a settlement, a rural area, an industrial plant, a street net etc. In order to handle the problem a new approach is presented, which uses structural and holistic methods for object recognition and scene interpretation. The system works on multisensor and multitemporal data. Expectations about the objects and scene content to be recognized can be formulated using different paradigms for knowledge representation. Structural dependencies can be represented explicitely as a semantic net, simple primitives like lines, edges, patches and complex primitives like streets and buildings, etc. can be described and extracted by holistic methods. The system and its components are explained using examples for the recognition of complex patterns like a purification plant, a fairground or the task of the verification of land usage.

1 INTRODUCTION

With the rising number of sensors carried on board of planes and satellites and the increasing performance of these sensors new tasks originate, which are handled more efficiently by remote sensing techniques than by standard procedures. These tasks include data acquisition for geographic information systems (GIS), refinement of GIS data using higher level of detail, update and verification of existing GIS databases. Presently most of the analysis is done manually by human experts. There are only few exceptions like data pre- and post-processing routines. However, in order to reach economic solutions, automated and robust image analysis techniques are required.

The described tasks belong to the area of pictorial pattern recognition in complex scenes. Complexity refers to two aspects. First the patterns themselves are complex and it is difficult to establish a general model for the automated analysis: What is a purification plant, how do streets in an urban scene look like, how do I recognize flooded land or how can I differentiate between industrial and suburban areas? The second aspect is that the scenes themselves are very complex in their details, like fields, streets, residential houses, buildings, bridges, vehicles, rivers, different types of vegetation appearing in rural areas, etc. The problems in automated image analysis are manyfold. Some are related to the large variability of pictorial appearances represented by a single semantic expression, other problems are due to the uncertainty and imprecision of the data caused by the nature of the sensors and natural variations in the scene, like weather conditions, seasons, scene illumination, flight parameters, etc.

A definite optimal solution to the mentioned problems doesn't exist, but we observe, that the human expert can perform the requested tasks despite all the obstacles. Therefore our approach is to observe the human expert and make all the information and methods available, which he uses as well. This means,

- we apply a multisensor and multitemporal analysis,
- we use prior knowledge, like interpretations from existing GIS databases, knowledge about scene features from learning sets and expert knowledge
- we apply flexible image analysis strategies and model-based approaches instead of fixed procedures. That means the knowledge is represented explicitly and system control is independent of the particular knowledge base and the data under investigation.

In the following chapters a knowledge based approach will be described in more detail, which we have successfully applied to complex scenes using structural pattern recognition techniques, holistic methods and expert knowledge, represented by semantic nets.

Various other approaches addressing these problems have been presented in the literature. The systems SPAM (McKeown et al. 1985), SIGMA (Matsuyama & Hwang 1990) and MESSIE (Clement et al. 1993) represent the first generation of knowledge based systems for the interpretation of aerial images. Application of rules for knowledge representation is a common method. In the BPI-system (Stilla & Michaelsen 1997) the rule base is structured in a network describing a part-of-hierarchy of the scene components. (Mees & Perneel 1998) suggests the distinction of strategy, global and sensor-dependent knowledge. The knowledge is represented in AND/OR-trees, fuzzy production rules and attributed prototypes with local image processing operators. In ERNEST (Sagerer & Niemann 1997) the knowledge base is formulated by a semantic net which describes the scene objects and their relations. A well defined network syntax facilitates the automatic reasoning. The MOSES system (Quint 1997) extends the ERNEST approach to extract man-made objects from aerial images using hints from a map.

2 STRUCTURAL APPROACH

2.1 *System structure*

Structural pattern recognition is advantageous if objects can be described as a composition of particular parts interrelated by some typical geometric, radiometric, functional or other observable feature. This is often true for man-made objects and structures. Figure 1 shows the overall scheme of the image analysis system AIDA which has been developed for this purpose.

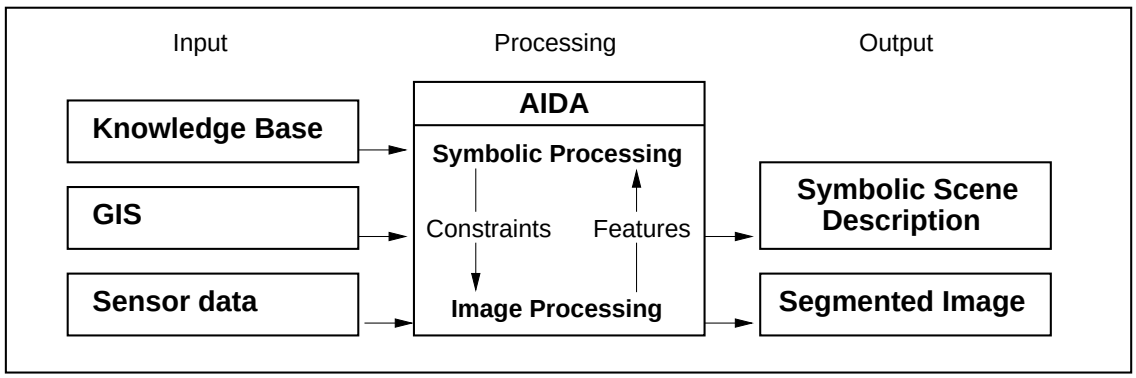

Figure 1. System components of AIDA.

For input the system uses multisensor and/or multitemporal image data of the scene to be analysed, a knowledge base which is relevant for understanding the investigated region and supplementary knowledge from a GIS. AIDA uses the knowledge base to set up hypotheses about the scene objects expected to be found and to derive constraints about the object properties appearing in the images. The output of the system is a symbolic description of the scene content and a segmented image describing the recognized objects and their geometric relations.

2.2 *Knowledge representation and system control*

For the explicit representation of expert knowledge a semantic net has been chosen. An example net is shown in Figure 2. Two classes of nodes can be distinguished: the concepts are generic models of the object and the instances are realizations of their corresponding concepts in the observed scene. Thus, the prior knowledge is formulated by concepts. During interpretation a symbolic scene description is generated consisting of instances. The object properties are described by attributes attached to the nodes. The relations between the objects are modelled by edges in the semantic net. The decomposition of objects is represented by the part-of edge. Thus the detection of an object can be reduced to the detection of its parts. The transformation of an abstract description into its more concrete representation in the data is modelled by the con-of relation. This relation allows to structure the knowledge in different conceptual layers like for example the scene layer, a material and geometry layer and a sensor layer, shown in Figure 2. Topological relations, like close-to provide information about the kind and the properties of neighbouring objects.

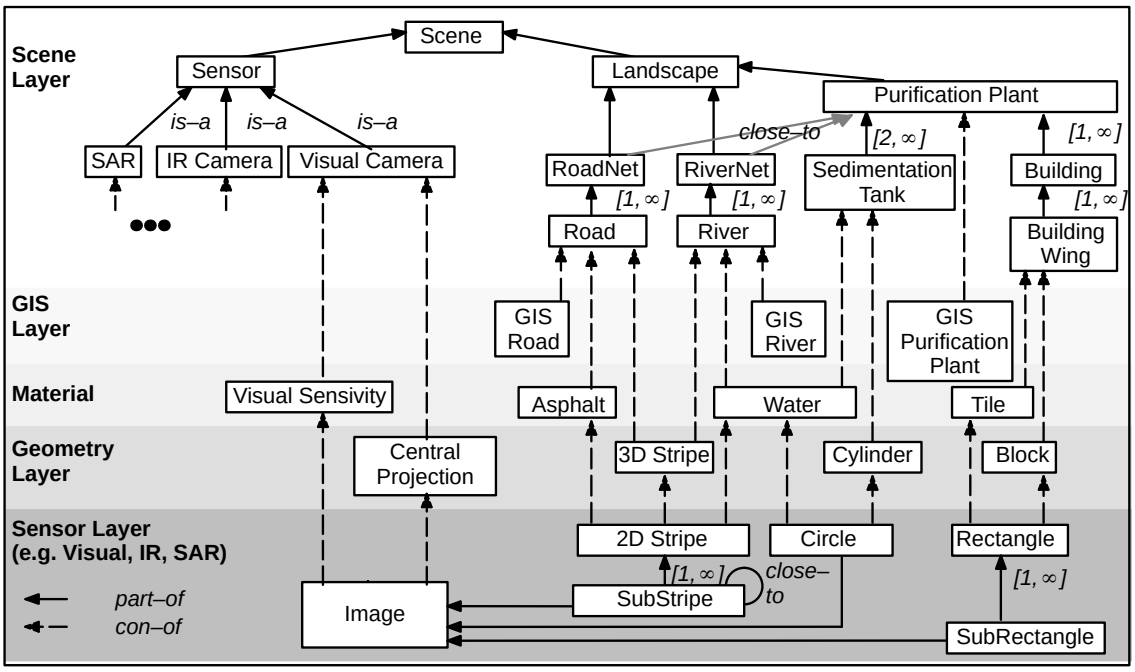

Figure 2. Semantic net representing the the generic scene model of a purification plant and its relation to the sensor image.

To use the knowledge represented in the semantic net, control knowledge is required that states how and in which order scene analysis has to proceed. The control knowledge is represented explicitly by a set of rules. An inference engine determines the sequence of rule executions. Whenever ambiguous interpretations occur they are treated as competing alternatives and stored in the leaf nodes of a search tree. Each alternative is judged by comparing the measured object properties with the expected ones. The judgement calculus models imprecision by fuzzy sets and considers uncertainties by distinguishing the degrees of necessity and possibility. The judgements of attributes and nodes are fused to one

numerical figure of merit for the whole interpretation state. The best judged alternative is selected for further investigation. Using a mixed top-down and bottom-up strategy the system generates model-driven hypotheses for scene objects and constraints about the object properties to appear in the sensor images. During the bottom up phase features are extracted from the image data, which are compared against the expected values and are used to reject, accept or refine hypotheses and constraints.

For an object recognition only those features are relevant which on one hand can be observed by the sensor and on the other hand give a cue for the presence of the object of interest. Hence the knowledge base contains only the necessary and visible object classes and properties. The network language described above is used to represent the prior knowledge by a semantic net. In Figure 2 a part of a generic model for the interpretation of remote sensing data for a scene containing a purification plant is shown. It is divided into the 3D scene layer and the 2D sensor layer. The 3D scene domain is split into a (semantic) scene layer and a physical layer, describing the observable geometric and material properties. If a GIS is available and applicable an additional GIS layer can be defined, representing the scene specific knowledge extracted from the GIS. The 2D image domain contains the sensor layers adapted to the current sensors and the data layer.

For the objects of the 2D image domain general knowledge about the sensors and methods for the extraction and grouping of image primitives like lines and regions is needed. The primitives are extracted by image processing algorithms and stored in the semantic net as instances of concepts like SubStripe or SubRectangle in Figure 2. The sensor layer can be adapted to the current sensor type like optical camera, SAR, range sensor, etc. All information of the 2D image domain is specified in the image coordinate system. As each transformation between image and scene domain depends on the sensor type and its projection parameters, the transformations are modelled explicitly in the semantic net by the concept Sensor and its specializations for the different sensor types.

2.3 *Multisensor and multitemporal analysis*

The automatic analysis of multisensor data requires fusion of the data sets. The presented concept eases the integration and simultaneous interpretation of images from multiple sensors by strict separation of the sensor independent knowledge of the 3D scene domain from the sensor dependent knowledge in the 2D image domain. New sensor types can be introduced by simply defining a new specialization of the Sensor node with the corresponding geometrical and radiometrical transformations. According to the images to be interpreted the different sensor layers (SAR, IR, Optical, Range) are activated.

Temporal changes can be formulated in a state transition graph where the nodes represent the temporal states and the edges model the state transitions. To integrate the graph into a semantic net the states are represented by concept nodes which are connected by a temporal relation. For each temporal relation a transition probability can be defined. As states can either be stable or transient the corresponding state transitions differ in their transition time which can also be specified for the temporal relation. For the exploitation of the temporal knowledge a timestamp is attached to each node of the semantic net. In contrast to hierarchical relations like part-of or con-of the start and end node of temporal relations may be identical - forming a loop - which means that the state stays unchanged over time.

An industrial fairground is an example for a complex structure which is only detectable by a multitemporal image interpretation. Using a single image the fairground would be classified as an industrial area consisting of a number of halls. The different states of a fairground are represented by the concepts FairIdle, FairConstruction, FairActive, and FairDismantling. The states representing the actual fair like the construction-state, the active-state and the dismantling-state are transient compared to the FairIdle-state which is valid most of the year. Therefore transition times of four to eight days are defined for the corresponding temporal relations and the node FairIdle is looped back to itself. The analysis starts with the first image, looking for an Industrial Area. In the given example

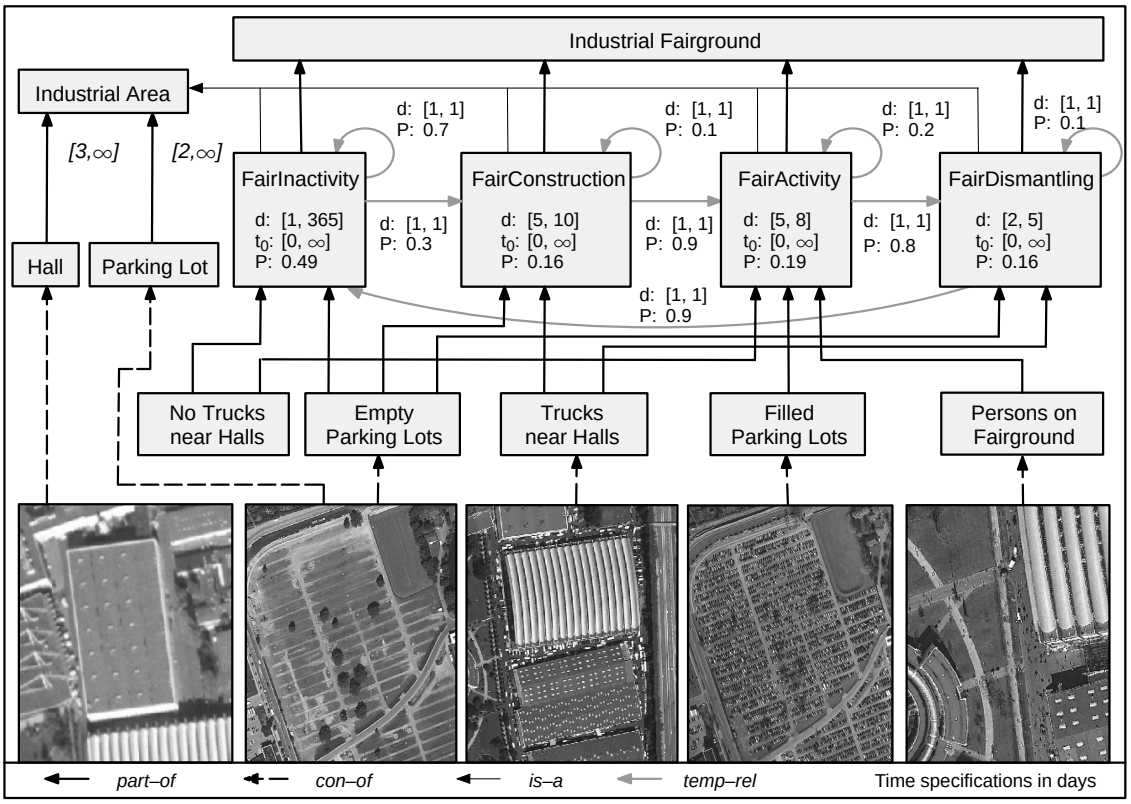

Figure 3. Semantic net for the detection of an industrial fairground using a state transition graph.

the system searches for at least three halls and one parking lot. If the Industrial Area can be instantiated completely the system tries to refine the interpretation by exchanging the Industrial Area by a more special concept. There are four possible specializations (FairIdle to FairDismantling) and the search tree splits into four leaf nodes. Each hypothesis is tested in the image data. A construction or dismantling phase is characterized by trucks parking close to the halls which deliver the equipment for the booths. Hence the system searches for bright rectangles placed around the halls. An active fair can be recognized by parking lots filled with cars and - if the image accuracy is sufficient - by persons walking on the fairground. If one of the four states can be verified the temporal inference is activated. The system switches to the next image in the sequence and generates hypotheses for the successor state. According to the elapsed time and considering the transition times all possible successors are determined. Having found hints for all obligatory states a complete instance of Industrial Fairground can be generated and the interpretation goal is reached.

3 HOLISTIC OPERATORS

Any type of structural pattern recognition method is based on features, so called primitives, which, in a holistic approach, are directly extracted from the image data. Often these features consist of edges, line-elements, patches, etc. Since the knowledge interpretation process is expensive from the computing point of view, the efficiency of the overall approach can be significantly improved by using preprocessed primitive sets which match the particular image analysis task. In the following three methods are described in more detail. One method extracts closed regions by detecting planar areas in the gradient image of distance maps (laserscans). This method is useful for detecting buildings. A second method extracts patches with homogeneous radiometric properties. It can be used to find regions, which are homogeneous in their surface properties, like homogeneous vegetation, water, settlement areas, homogeneous geological areas, etc. A third method extracts line-

like structures of predefined profiles as they are typical for transportation nets like roads, rails or canals.

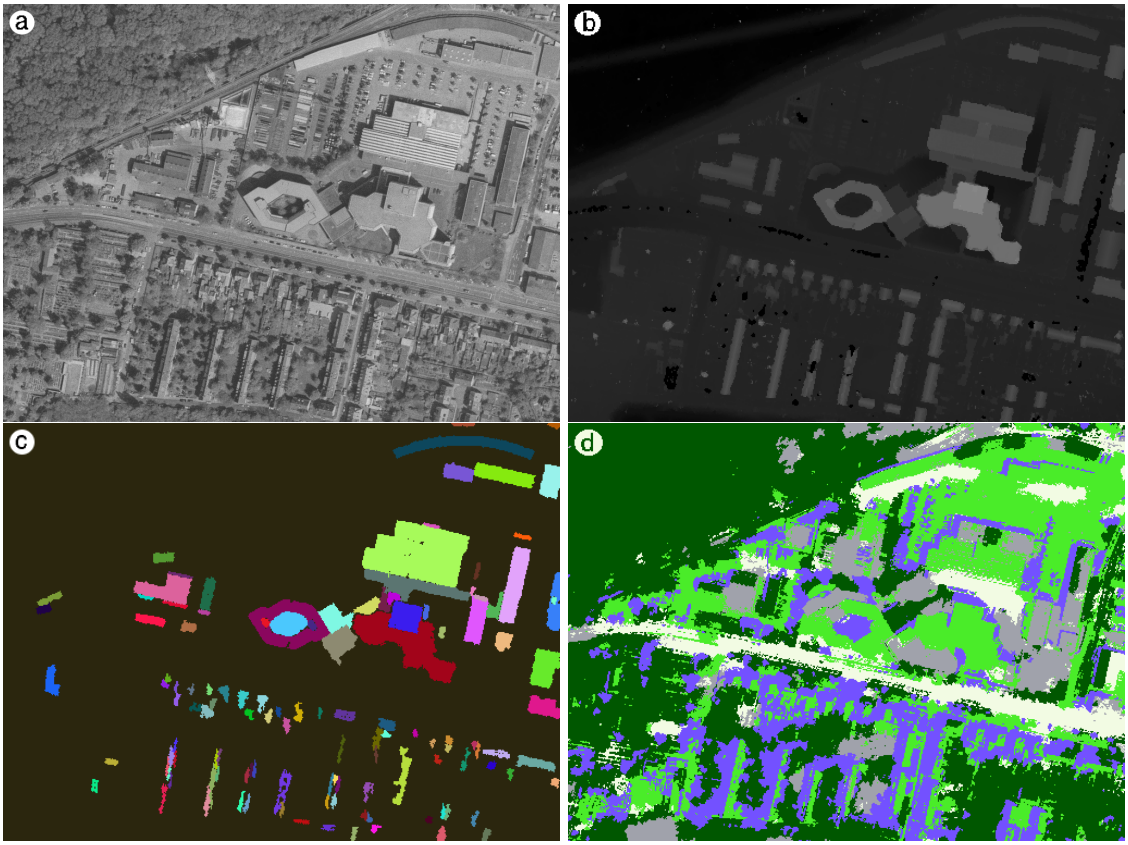

Figure 4. Investigated area - (a) aerial image, res. 0.5m/pix, (b) laserscan, res. 1m/pix and Holistic operator results - (c) Building detection, (d) Texture classification.

3.1 *Detection of planar objects*

Man made objects can be detected by processing the laserscan height information, shown in Figure 4b). Neighbouring regions with local height differences below a given threshold are connected and a unique label is assigned. Labelled regions smaller than a preset minimum surface size, as they appear for example in forest areas, are reassigned to the surrounding region, if the regions are too large they are set to the background. This segmentation process delivers a label image containing unique labels for each object, as it is shown for buildings in Figure 4c).

3.2 *Texture classification*

Regions with homogeneous surface properties can be described efficiently by texture models based on 2D stochastic processes. One possible implementation of a stochastic model for supervised texture segmentation is described by (Gimel'farb & Zalesny 1993). Supervised means that a teaching sample containing the texture classes expected to be found in the image data is analysed regarding first and second order stochastical features. The first order model describing the probability density of the data is determined from the luminance histogram of all elements of each training class. In general the occurrence of a certain luminance value is statistically dependent on the values of its neighbours. The second order texture model considers pairwise pixel interactions of neighbouring pixels, called cliques, and describes their local interactions by Gibbs random fields. Only the

statistical significant clique types are selected for definition of a neighbourhood system κ. The texture model considers four probabilities: The luminance probability P_1 for each class (first order), the probability P_2 of each clique belonging to the same class, the probability P_3 of luminance difference for each clique within a region, and the probability P_4 of luminance difference for each clique at the regions' border. If more than one image band is available the interdependence of the bands is also evaluated. The probabilities P_i specify the potential V_κ of a clique κ.

Gibbs random fields describe the joint probability of a segmentation of a neighbourhood system κ The segmentation has to maximize the joint probability for the whole image, which is an iterative process due to the neighbourhood relations. A result image of the texture segmentation is shown in Figure 4d). The texture model of this example contains the four classes forest, meadow, streets and settlement.

3.3 *Road detection*

Traffic nets like roads are characterized by typical luminance profiles, smoothness and continuity along the main direction. An algorithm for the extraction of traffic nets has been developed using three processing steps. In the first step a gradient and a direction image are calculated. In areas which match user-defined parameters all road candidates are extended using an A*-search-algorithm. The parameters contain the road width specified as interval (min, max value), the brightness of the road area, the gradient profile, shown in Figure 5a) and 5b) and the minimum road length. The last step of the algorithm is a vectorization of the resulting line segments.

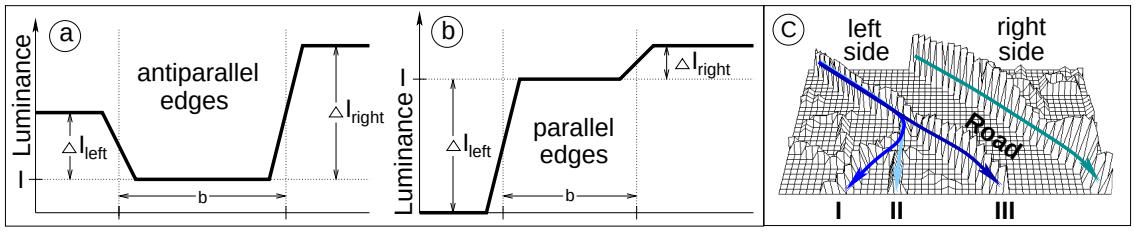

Figure 5. (a),(b) Vertical profiles of the road gradient and (c) alternative results of the A*-algorithm.

4 HYBRID APPROACH

The system AIDA described in Section 2 follows a purely structural approach, which requires that hypotheses of the top semantic level always have to be verified by holistic operators on the bottom image level. Especially in image segmentation tasks it is useful and more efficient to directly employ specialized holistic operators for region extraction, like those mentioned in Section 3, rather than to use a structural pattern recognition scheme leading down to the very detailed components of the region. For example it is not necessary to detect each individual house in order to recognize a settlement area. As a consequence of such requirements a more general redesign has been made with the name geo`AIDA` permitting the use of holistic operators on all levels of the semantic net.

4.1 *Structure of geo`AIDA`*

Nodes and Links Similar to AIDA the domain specific knowledge in geo`AIDA` is coded as semantic net with concept nodes and edges describing the relations between the nodes. According to the requirements of remote image analysis the semantic net is structured hierarchically, based on the idea that a region can be separated into more detailed regions which form true subsets. For example the concept *house* in Figure 7 is a component

of the more abstract region *settlement*. Therefore the most important edge type is the *part_of* edge. In contrast to AIDA the strictly hierarchical structure of the concept net has been extended to all other relations as well. Experience has shown that this request imposes no limitations and eases system control considerably. Another essential difference to other systems using semantic nets is that geo`AIDA` transfers the complete edge functionality into two classes of operators attached to the net nodes (`TopDown` and `BottomUp`).

Analysis The concept net represents the knowledge about the objects expected to be found in the scene. The result net (instance net) depends on the investigated scene. The analysis process is able to exploit scene knowledge and to generate hypotheses (`TopDown`-analysis) as well as to suitably rate extracted objects and to perform a grouping of them (`BottomUp`-analysis). The system control is independent of the problem, generates and deletes instances and controls the execution and parameterization of all necessary operators. Besides a structural scene interpretation approach as used by most of the semantic net based systems (Niemann et al. 1990) (Liedtke et al. 1997) an additional holistic approach was integrated. Holistic operations allow to connect the input data sources on all interpretational levels. Holistic instantiation enables instantiation of a concept without looking at it's structural components. The user can utilize these features to define an analysis depth and to limit the analysis process to certain parts of the concept net.

Operators Each node of the concept has two slots for connection of operators for `TopDown`- and `BottomUp`-analysis. As it is possible to use arbitrary operators which were designed for very different purposes, a dynamic interface description for linkage of operators to the concept net was realized. This mechanism allows to use any existing segmentation or evaluation operator to be connected to the system.

The general task of the `TopDown` operator is to split a region into subareas, to classify and to assign a meaning to these areas and further to provide figures of certainty for this segmentation process. A holistic operator can be attached to any concept node to achieve a direct and fast classification or object extraction as described in Sections 3.1 and 3.2. Depending on the interpretation depth determined by the user the subregions created by a segmentation are further investigated. An alternative procedure would be to recognize an object by its structure if there is no holistic operator available. Also a combination of both strategies is possible: the holistic operator might be used to initially segment a region geometrically, afterwards a structural analysis of the subregions is performed.

The `BottomUp` operator evaluates the attributes of the child instance nodes. Based on this evaluation the instance nodes are grouped and rated. Alternative group arrangements are compared and only the best combination is propagated. During the grouping process a region can be split into subregions if this increases the overall rating and probability of a certain segmentation. In the last step the newly formed instances are valued.

4.2 *Application land use*

The interpretation process is illustrated on the verification of land use. Input data is an excerpt, see Figure 6a), of the gray-value aerial image with a resolution of 0,5m/pixel shown in Figure 4a) and the corresponding part, Figure 6b), of the laserscan with a resolution of 1m/pixel shown in Figure 4b). The laserscan data can be used for a building detection, see Section 3.1, Figure 6d) and Figure 4c). The knowledge base for this example is shown in Figure 7.

The initial analysis step generates a hypothetical instance *scene*. This instance links the input images to the system and produces an instance *region* which splits the whole scene into subregions using information of street positions, see Figure 6c). The holistic operator of the concept *region* queries the database of a GIS for the street coordinates and generates a matching label image for the investigated scene.

The semantic net is designed to determine the type of land usage for regions. All instances generated by the concept *region* have to be processed further to acquire this type.

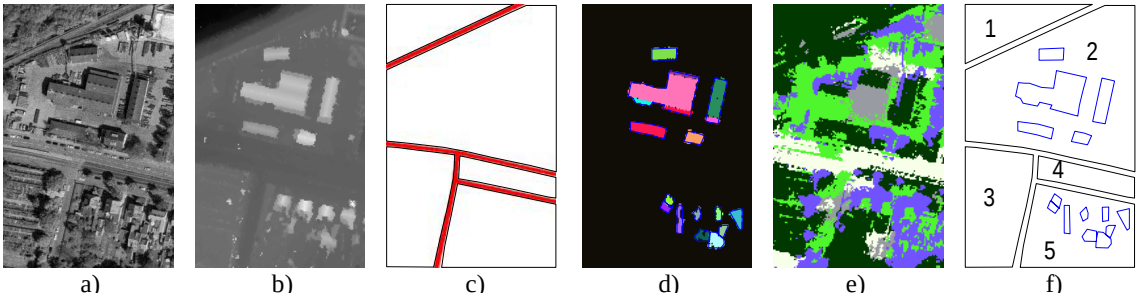

Figure 6. a) ortho photo, b) laserscan, c) initial GIS segmentation, d) building extraction from laser scan data e) texture segmentation of ortho photo, f) final scene interpretation by geoAIDA (1: forest, 2: industry, 3: forest, 4: unknown, 5: settlement).

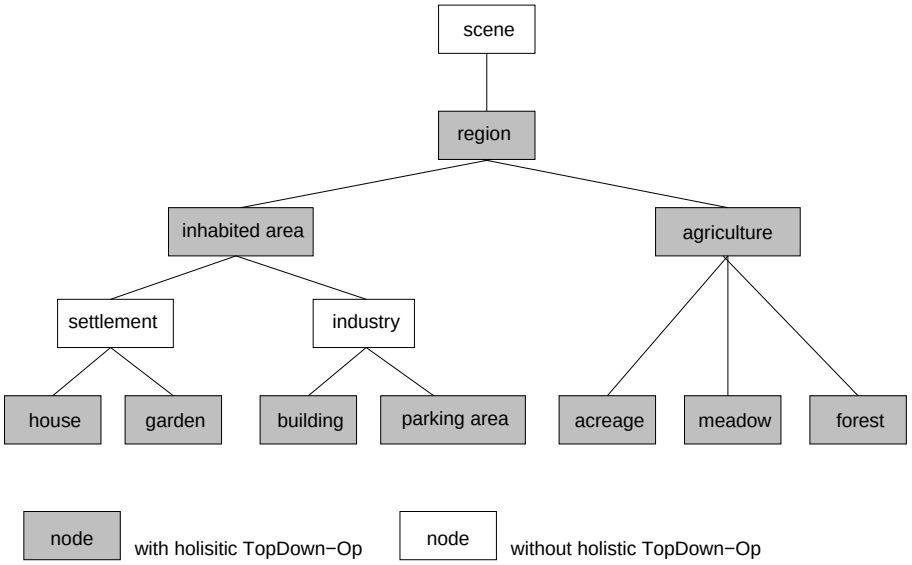

Figure 7. Semantic concept net.

According to the concept net each region can be hypothetically instantiated as *inhabited area*, *agriculture* or ambiguous, if the region contains areas with both types of land usage. This step utilizes the holistic building detection described in Section 3.1. In the following an instance *settlement* as well as *industry* is created for each instance *inhabited area* by instantiation of all the necessary structural components (*house*, *garden*, *building*, *parking areas*). In parallel all hypothetical instances *agriculture* are further processed using the results of the holistic texture classification shown in Figure 6e) and Figure 4b).

Now that the `TopDown`-analysis has reached the leaves of the semantic net, an extensive set of instances for each observed region is available which is submitted to the `BottomUp`-analysis for grouping, evaluation, deletion, etc. Competing interpretations of the same region are differentiated according to the probabilities gained during the `TopDown`-progression. The `BottomUp`-analysis propagates back to the top node *scene* and delivers the scene interpretation result shown in Figure 6f). geoAIDA classifies region one and three as forest, two as industry and five as settlement.

5 CONCLUSION

This contribution presents an approach for knowledge-based image interpretation of complex scenes. Like ERNEST (Niemann et al. 1990) AIDA uses concept nets to model expert knowledge about the expected scene content. These concepts are turned into a symbolic scene description when detected and verified during the analysis process. AIDA features operation on multisensor data, usage of GIS information as an underlying interpretation model for data verification and processing of multitemporal input data for very sophisticated monitoring tasks (moorland, industrial fairs). The ideas of AIDA were transferred into the new system geo`AIDA` which adds region-based and holistic analysis to it's purely structural working predecessor. geo`AIDA` is a promising production tool for image analysis applications in remote sensing. One major application which is presently pursued is the verification of geographic information systems from orthophotos.

References

Clement, V., Giraudon, G., Houzelle, S. & Sadakly, F. 1993. Interpretation of Remotely Sensed Images in a Context of Multisensor Fusion Using a Multispecialist Architecture. *IEEE Trans. on Geoscience and Remote Sensing* 31(4): 779–791.

Gimel'farb, G.L. & Zalesny, A.V. 1993. Probalistic models of digital region maps based on Markov random fields with short and long-range interaction. *Pattern Recognition Letters* 14:10 789–797.

Liedtke, C.-E., Bückner, J., Grau, O., Growe, S. & Tönjes, R. 1997. AIDA: A System for the Knowledge Based Interpretation of Remote Sensing Data. *3rd Int. Airborne Remote Sensing Conference and Exhibition, 313–320, Copenhagen, Denmark.*

Matsuyama, T. & Hwang, V.S.-S. 1990. *SIGMA: A Knowledge-Based Aerial Image Understanding System.* New York: Plenum Press, p. 277.

McKeown, D., Wilson, A. & McDermott, J. 1985. Rule-Based Interpretation of Aerial Imagery. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 7(5): 570–585.

Mees, W. & Perneel, C. 1998. Advances in computer assisted image interpretation. *Informatica - International Journal of Computing and Informatics* 22(2): 231–243.

Niemann, H., Sagerer, G., Schröder, S. & Kummert, F. 1990. ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 12(9): 883–905.

Quint, F. 1997. MOSES: A structural approach to aerial image understanding. In A. Gruen, E. Baltsavias & O. Henricsson (eds), *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II)*: 323–332, Basel: Birkhäuser.

Sagerer, G. & Niemann, H. 1997. *Semantic Networks for Understanding Scences.* Advances in Computer Vision and Machine Intelligence. New York: Plenum Press.

Stilla, U. & Michaelsen, E. 1997. Semantic modelling of man-made objects by production nets. In A. Gruen, E. Baltsavias & O. Henricsson (eds), *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II)*: 43–52. Basel: Birkhäuser.