Thematic information extraction and data fusion, a knowledge engineering approach to the problem of integral use of data from diverse sources.

N.J.Mulder. ITC/UT, Enschede,
the Netherlands.

1. ABSTRACT

Multisource data integration is reformulated as a problem of defining a problem space with goals and constraints. Solving the problem of detecting objects, estimating parameters of geometric and radiometric models and classification of objects is described as searching for the goals by navigation through problem space which can be reduced to a tree search. An inference engine provides the mechanism for navigating through problem space. By defining the inference procedure as a backward chaining of rules it is possible to select only data which are relevant to current hypothesis evaluation. Backward chaining also allows the handling of cases of missing data. Information quality and error propagation are treated under the formalism of maximum likelihood / minimum cost decision making. Likelihood vectors are stored or regenerated for future use.

Arguments are given for not using the Dempster,Schaefer method. The approach of defining a search space and use inference engines for navigating from initial state to goal state is contrasted with the usual approach of data merging by colour picture painting. The knowledge based approach is illustrated by hypothesis evaluation using both ordinal (remote sensing) data and nominal (GIS , attribute) data. Examples are provided of the integration of multispectral data with radar data, and on model based image interpretation applied to the recognition of buildings in airphotos.

keywords : multi source, data integration, inference, backward chaining, Bayes, remote sensing, geo information system, knowledge engineering, knowledge based systems.

2. Introduction.

Multi source data integration is not a well posed problem. Integration is not an aim in itself but must serve the purpose of providing information about Earth related objects and processes. The general approach should be to build a model of the world by a process of abstraction which is defined by the world model of the users of information. All querries to a GIS require the identification of objects and an estimation of the state of the object(s) in terms of properties and attributes (for a certain time interval).

Object definitions (3 dimensions + time) can be stored in a database or can be derived when needed from various sources.

The general goal behind the sentence "multisource data integration" is to identify objects in space and time and to determine the value of a set of properties and attributes.

In knowledge engineering (not the same as artificial intelligence, but concerned with handling the product of natural intelligence) several techniques exist for linking querries to answers. Information is a relation defined on the cross product of questions {q }, and answers {a }. For a specific question q is_a member_of{q } and a specific answer a is_a member_of{a }, information is instantiated I = {(a ,q)} , by the relation alq. The space of all (a,q) for a certain domain of modelling can be very large, so finding the right a for a given q may require extensive search. The inverse problem of finding a question matching an answer is also a nontrivial one.

Knowledge engineering can be described as the art of defining the proper sets of questions and answers for a certain domain, and the building and using of engines for the navigation in problem space. The need to

navigate in problem space comes from the need to connect questions and answers for the production of information.

Simple knowledge based systems use a knowledge base of facts and rules and an inference engine for the linking of facts by rules to produce 'new' facts. The linking of facts, to produce new facts is equivalent to linking questions and answers, in (generic)GIS language, which in turn is equivalent to the linking of hypotheses and evidence, in the language of knowledge engineering.

An inference engine working in a forward chaining mode starts with facts like RS and GIS data and finds the most likely matching hypothesis. The forward chaining mode requires pattern matching on the input data and is not very robust in situations of incomplete / missing data. But if the dataset is complete and homogeneous, the standard updating of prior probabilities to posterior probabilities via Bayes' tautology is applicable.

An inference engine working in the backward chaining mode starts with the hypothesis with the largest prior probability and navigates to the data which provide evidence in favour of, or against that hypothesis. When insufficient evidence can be found from (part of) the data, the likelihood for alternative hypotheses is evaluated. The backward rule chaining, hypothesis driven method has as advantages that it can be more efficient if the hypothesis set is smaller than the evidence (data) set, and it does not stop when data items are missing.

Both methods have the advantage that explicit rules must be defined for the transformation of data into evidence. Ideally one would define a meta level knowledge based system containing the relevant rules from the physical sciences. When somebody would map a combination of SPOT XS data and SAR data into a hue and saturation space, the meta system would not allow the rule to be entered before an explanation was given on how coherent reflectance in the microwave domain could be made compatible to photon reflectance. Similarily the system would not allow the combination of emissive and reflective infrared photon radiation in a simple image calculator, also band ratios and principal component transformations would be exclude from the model until a sufficient explanation was given (in terms of axioms and re_writing rules).

Presently, such a moderately 'learned' meta KBS does not exist. Instead a KBS has an explanation sub system, which must provide the reasons and trace the reasoning process.

The treatment of uncertainty can be completely handled by the current statistical tools [Fukanaga,1990]. The (per object) priors come basically from a gGIS, storing likelihoods for the state of objects and the priors are updated to new likelihoods by any new data available such as regularly provided RS data. There are no good arguments for the area of geo informatics to limit the application of statistics to the assumptions made by Dempster&Schaefer. In [Mulder,1990] it is shown in how far the Dempster&Schaefer approach is inferior to the proper use of the minimum cost / maximum (cost weighted) likelihood rule.

The need for a knowledge engineering approach comes mostly from a desire to please the intellect by avoiding nonsense transforms on multiple source data and second to state and solve the problem of information instantiation from hypothesis and evidence in an elegant way.

The schema for extracting information from multi source RS data is: - find class,parameter priors (from a newGIS) - for given data, find the model defining the relation (class, parameters - data). - update the (class,parameter) probabilities for the(sub)set of objects in the model (GIS). - if needed, redefine objects (e.g. new landuse distribution) by split and merge operations.

Research at ITC and UT, is directed at model based image analysis with emphasis on modelling in connection with minimum cost classification and parameter estimation ref. ([Cheng,1989], [Korsten,1989], [Schutte,1992]).

3 Forward and backward reasoning , missing data.

3.1 (Geo)Information and knowledge.

The purpose of a knowledge based system is to establish a link between hypotheses and evidence. In general this link is not binary but probabilistic: P(H|E), the likelyhood of hypothesis H given evidence E.

In a GIS the link between questions and answers has to be made, as discussed in chapter 2. Given a question, the hypotheses are possible answers for which a truth value c.q. likelihood has to be determined.

In RS and GIS applications we have to solve classification problems and problems of estimating the best parameters of radiometric and geometric models. Concentrating for this moment on the estimation of likelihood for class membership : the hypotheses are about the class an object belongs to and the evidence is derived from the data, so P(H|E) -> P(Cl|x). (In a similar way parameters of say a metric model can be estimated : P(parameter|x) under a minimum cost / max likelihood criterion).

3.2 Models for reasoning.

Forward reasoning : usually one starts with a data vector x followed by the evaluation of the posterior probability for each class given the value of the data vector. With equal cost functions for all classes (cost of misclassification) the minimum cost classification rule is the same as the maximum likelihood rule, which is also called the max. a_posteriory rule, the MAP rule. When data are missing, or when there are too many data, the data driven approach tends to fail. In such cases it is better to start with the most probable (a_priory) hypothesis and search for data supporting or negating each of the hypotheses.

Backward reasoning : starting with a hypothesis ,in the case of one source of data with occasionally missing data, the expression P(Class|x1,x2,x3,..) can be evaluated even when only one of the vector elements xi is missing.

3.3 Nominal, GIS data.

The role of data already stored in a model (say a GIS) is to provide the best possible prior probability for the class of the object under consideration, P(Class(time)). The link between P(Class(time+1)|xi) and P(Class(time)) is defined through the Markov and Bayes relations.

Markov : P(Class(time+1)) = function(Class(time),context).

Bayes : P(Class|x) x P(x)= P(x|Class) x P(Class) .

3.4 Missing data.

If the components of x =[x1,x2,...] were independant, then P(x|Class) = P(x1|Class) x P(x2|Class) x ....

Independancy of data components is one of the aims of feature extraction from data, but does not solve the problem of missing data.

The estimation of the interdependancy of x1,x2,.... given class can be done in a non_parametric way which is optimal in terms of minimum error, or in a parametric way, which is minimal in terms of efforts for the human brain.

Within a parametric approach, assuming a Gaussian distribution of frequency(x1,x2,..,Class=constant), the parameters of the distribution are mean(x) and covariance_matrix(x). The MAP decision rule is equivalent to a minimum Mahanalobis distance rule in an anisotropic measurement (feature) space.

The effect of e.g. x2 missing from [x1,x2,x3] is a projection of a 3 dimensional cluster onto a 2 dimensional subspace. In order to let the missing data not influence the likelihood for a class, it is sufficient to substitute for x2, the mean(x2,class) for every class.

In the above schema the data interdependancy is taken care of by the covariance matrix while the missing data gets a default value per class. The dependancy of the data repair operation on the class under consideration indicates a backward chaining mode.

The inference procedure is then :

for all possible classes for the object under consideration do: - look up the prior probability for that object and that class -> P(Class) - evaluate the data vector x , if components are missing then replace them with the most likely xi ,= max P(xi | Class). - update P(Class | x), store it in the gGIS.

3.5 Multiple data sources, multiple models.

With multiple data sources the critical part is in the feature extraction by model inversion. One of the aims in feature extraction is to get non redundant, statistically independant clusters for the classes. As in reality most processes are coupled, it is very likely to find

situations where it is not allowed to write : P(x|Class) = P(x1|Class) x P(x2|Class) x ... P(xi|Class).

This assumption of statistical independance is often assumed in fuzzy logic schemas. Therefore these methods are as weak as their assumptions.

Model based image analysis concentrates on the estimation of model parameters from measurement data under a minimum cost optimisation rule. Good models have defined orthogonal parameters, so a first step in feature extraction should be model inversion.

Each data source provides estimaters for its own model with its specific parameters. For example, multispectral data provides estimators for surface angle -> leaf angle distribution, shading of slopes and spectral reflectance -> the mixture of visible surfaces within a resolution element -> leaf area index. A temperature / heat balance model requires the complement of reflectance = absorption of radiation. From thermal data the emission can be measured and the temperature or emittance can be estimated. The microwave reflection model has parameters for surface normals, surface roughness and dielectric properties.

Dependancies between models are initially estimated on the bases of a dimension analysis and on the form of the models involved. This is followed by an estimation of the cooccurence of parameters. In the end the total dependancy is represented by P(Class | data1, data2, data2,..) or its inverse P(data-i,.. | Class). The dependancy between class membership, context and data is first modelled and then adjusted to observed values for the model components. (context is modelled through local priors).

Given the dependancies of all occuring (class, data), the availability of only one, or few data sources at a time can now be treated as a case of missing data !

4 Practice of multi source, multi class data.

4.1 GIS + RS data.

In the daily practice of RS data analysis and use of a GIS the types of data available for hypothesis evaluation and classification are : (old) maps or GIS(t=0) -> status at time=0, + data ; emitted photon data, reflected photon data, reflected (synthetic aperture/ real aperture) electro magnetic waves -> GIS(t+1).

We assume that error correction and geo referencing procedures have been applied and that the source images have been segmented into surface objects -> scene objects.

In model based scene analysis, a forward model is defined as a relation between scene object parameters and remotely sensed data. The analysis problem is mostly a problem of model inversion: RS data -> object parameters.

Objects in a scene are labelled by attributed class names. The class names serve as a label indication groups of objects which have something in common (which need not be something observable by (all sensors) remote sensing). This leads to the definition of classes and subclasses connected by reasoning about observability. The relation between an observable from a certain source and class membership of the scene object is the feature(vector). Features are often defined through the definition of the interaction model between object(class) and radiation -> data source. Feature determination is in that case equivalent to parameter estimation.

The general classification schema is :

maps, status of GIS -> class priors -> observable source1 -> feature 1 -> class (likelihood) observable source2 -> feature 2 -> etc.

Old maps, GIS data: are very usefull in the definition of prior probabitilies for classes per object, specially when combined with a Markov state transition probability algorithm [Middelkoop,1990,2]. It also helps in object detection in combination with one of many area segmentation procedures.

Digital elevation models (from the GIS) are used for the prediction of shadow and shading effects via ray tracing procedures.

Different sources lead to different features. Features are often parameters of models linking observables to object descriptors.

4.2 Examples.

For example : model(leaf area index, leaf angle) -> multispectral reflection. Inversion of the model : multispectral data -> green vegetation index -> leaf_area index and multispectral data -> sum of photons all bands -> intensity -> leaf angle. leaf area index & leaf angle -> vegetation subclass.

Reflected and emitted photons: Landsat TM data with 6 reflective pass bands and one infrared band. As there are two completely different processes at work it would be senseless to just combine them into one picture. (Temperature is a state variable, the observed emitted radiance is a function of the heat balance over a long time interval. The use and interpretation of the data requires therefore a model for heat flow with many parameters which are not observable by RS techniques). So the reflected radiance is used to measure the amount of radiative energy absorbed at the moment of observation. This is extrapolated over the previous period. Other components of the heat transfer model can be derived through surface class membership. If the emittance of the objects is known then the temperature can be estimated from the IR data. In model based image analysis, the predicted temperature is compared with the estimated temperature, and an optimisation subroutine varies the remaining variable parameters until a minimum cost of estimation / classification is found.

Temperature can only be determined if at least the (directional) emittance of the object under observation is known.

Feature extraction depends also on class hypothesis: for waterbodies and the problem of thermal polution the temperature distribution (flow model) is the feature, for crop monitoring relative temperatures indicate the degree to which plants cope with heat stress and for soils the temperature is related to soil moisture because water has a high heat capacity.

In applications of meteo sat and NOAA VHRS data, the temperature estimation is relevant for the determination of height of clouds, which in turn feeds into a rainfall likelihood model. External data are in this case provide by rainfall gauges and predicted pressure and windfield distribution plus vertical profiles of humidity, pressure and temperature.

Spatial features are defined through geometric (solid) modelling, parameters such as object position, size and orientation are estimated from the RS images. Progress is made in the geometric modelling of buildings [Schutte,1992], trees and homogeneous vegetation canopies .

The parameters of the microwave reflection model are mostly geometric, regular like reflecting plane surfaces or irregular like area roughness and scattering vegetation canopies. Woodhouse,1990 has build a simulation model and has demonstrated how a sequence of parameter adjustments reduces the difference between predicted and actual radar image to a noise picture. The next important parameter is the dielectic constant in combination with (surface) conductivity. Last, anisotropic scatterers / reflectors change the polarisation and phase of the incoming e.m. wave. This leads to an estimater of e.g. tree branch directional distribution or directional distribution of fissures and cracks in rocks and ice. For water applications the parameters describing sea state are important. Further research into the model based analysis of SAR images takes into account the spectral classification at an earlier data plus a Markov estimator for the change with time.

4.3 The method.

As feature extraction depends so much on class definition and the physical model describing the interaction between radiation and matter, the hypothesis driven reasoning of knowledge based systems is selected.

The following strategy is used :

- problem analysis leads to the definition of querries in terms of classes and subclasses of objects and state parameters of the object. - the present state of the model representing the previous state of the world is used to predict a_priory probabilities for class membership and parameter values. - for each {class , parameter, source} combination, the appropriate features are extracted. - for each object in the scene the class likelihoods and the process / state parameters are updated. - the GIS used for modelling (specific views of) the world stores the class likelihoods and relevant state variables / parameters together with a time tag.

The knowledge base of the system consists of two main parts :

I - facts about the status of the world model -> GIS. - procedures for feature extraction / parameter estimation. - procedures for predicting future states.

II - rules for the selection of procedures for feature extraction. - rules for updating the model, given data(source).

Efficiency of hypothesis evaluation is needed because with higher resolution (including digitized photos) and an increasing number of sources the volume of data increases more rapidly that the volume of information. Efficiency can be achieved through the top down, backward chaining of rule by gathering the statistics of the degree of change of probability from prior to posterior as a function of {class , feature -> source}. If for a certain class a certain feature does not significantly change the likelihood for that class then there is no need to evaluate P(Class|feature) for that combination in future. This is also the case if for an object the P(Class) is so low that it is very unlikely to lead to a significant P(Class|f).

P.M. : Bayes , $P(Cl|f)xP(f)=P(f|Cl)xP(Cl)$, $P(f|Cl)$ would have to be very high to compensate for a low P(Cl).

5 Concluding remarks .

Most of the present publications on the subject of using multisource data is at a pre_scientific level of picture processing. One of the more favoured painting receipes is to play with the IHS transform. Another favourite is to throw data of different sources and hence incompatible physical dimension together into a principal components analysis. This disregards not only the incompatibility of the physical units but also the restriction of linear transforms to additive vector models.

Discussion with experts in visual image interpretation who have looked at e.g. SPOT+SAR -> HSI pictures does not provide more representable knowledge than can be derived from physical modelling. The useful knowledge of interpretation experts is in the field of context dependent prior probabilities related to complex spatial relations or to complex processes involving human activies such as destroying the environment. Their expertise is best used in defining sensible hypotheses about object's states and about processes and the relation between priors and context.

Progress in computer assisted image analysis is most rapid in those areas where models can be defined for the relationship between object class, model parameters and data(source). Examples are model based analysis of buildings [Schutte,1992] and plants in digitized aerial photos, the use of vegetation indices and (DEM) model based analysis of SAR radar [Woodhouse,1990].

Backward chaining of classification and parameter estimation rules allows efficient handling of missing data, and the omission of data which is not relevant to the evaluation of a current hypothesis.

The GIS which is used to contain the world model must have the possibility to store the relevant P(class) vectors as these are required for a multi source updation of the model of the world.

The combination of Bayes and Markov relations can be used to estimate states of the system as a function of time.

The above formulated meta rules have resulted in a research agenda at ITC aimed at model based image analysis, in cooperation with the University of Twente. Research into a GIS with likelihoods is executed in cooperation with the Rijks Universiteit of Utrecht (the Camotius project).

The definition of the relationship {class , parameter, data(source)} is central in the treatment of data from various sources. The knowledge base with rules for the relation Class -> image processing procedure, is under construction in a PhD project [Fang,1992]. Rules for Class , parameters, data(source) -> feature extraction ,will have to be added (in the problem analysis part).

5 References.

Cheng, 1990, Design and Implementation of an Image Understanding System: DADS ,Cheng X.S., PhD thesis TU Delft ,1990

Fang, 1990, Computer assisted problem solving in image analysis, Fang Luo & N.J.Mulder, to be published, Proc. ISPRS, Washington, 1992.

Fukanaga K.,1990, Introduction to statistical pattern recognition. Academic press, 1990, ISBN 0-12-269851-7.

Korsten, 1989, Three-dimensional body parameter estimation from digital images, Korsten M.J., PhD-thesis, Febrodruk, Enschede,nl,1998.

Makarovic A.,1991, Parsimony in Model-Based Reasoning . PhD thesis TU Twente , 1991 ,ISBN 90-9004255-5.

Mulder,1990,3, An evaluation of uncertainty handling in expert systems " - Mulder, N.J. and Middelkoop H., in Proceedings ISPRS Symposium, May 1990, Wuhan, P.R. China, pp. 578-597

Middelkoop,1990,1, Parametric versus non-parametric MLH classification - Mulder, N.J. and Middelkoop, H. in Proceedings ISPRS VI Symposium, May 1990, Wuhan, P.R. China , pp. 616-628; accepted for ITC Journal 1991

Middelkoop,1990,2, Uncertainty in a GIS, proposal for a test for quantification and interpretation - Mulder, N.J. and H. Middelkoop, in Proceedings ISPRS VI Symposium, May 1990, Wuhan, P.R. China, pp. 598-615; accepted for ITC Journal 1991

Middelkoop,1990,3, Progress in knowledge engineering for image interpretation and classification - Mulder, N.J., H. Middelkoop & J.W. Miltenburg in ISPRS & Journal of Photogrammetry and Remote Sensing, sept.1990

Pan ,1990,3, Behaviour of Spatial Structure in three Domains: Conceptual, Physical and Appearance - Pan, H.P., C. Austrom and N.J. Mulder in Proceedings of 4th International Symposium on Spatial Data Handling, July 1990, Zuerich, Switzerland

Salamanka,1990, Quadtrees data structure for handlin large data sets of multi-resolution nature,Salamank S.C. MSc. thesis, ITC Enschede,February, 1990.

Schutte ,1992, Knowledge engineering in RS an Knowledge based systems in GIS,- N.J.Mulder an K.Schutte,to be publ. Proc. ISPRS, Washington, 1992.

Sijmons K,1986, Computer assisted detection of linea structures from digital remote sensing data for cartoç raphic purposes , dissertation, Freie Universitaet Berlir Berlin 1986.

Woodhouse I.,1990, The Simulation of Radar Images fc Remote Sensing Applications, Techn.Rep. ITC-IPL Sep tember 1990 under ERASMUS exchange with Dundee University, Scotland, UK.