Interpretation of remotely sensed images in a context of multisensor fusion using a multi-specialist architecture

Veronique Clement, Gerard Giraudon, Stéphane Houzelle, Fadi Sandakly

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

2004 route des Lucioles
B.P. 93
06902 Sophia-Antipolis
France

Rapports de Recherche

N° 1768

Programme 4

Robotique, Image et Vision

INTERPRETATION OF REMOTELY SENSED IMAGES IN A CONTEXT OF MULTISENSOR FUSION USING A MULTI-SPECIALIST ARCHITECTURE.

Véronique CLÉMENT
Gérard GIRAUDON
Stéphane HOUZELLE
Fadi SANDAKLY

Octobre 1992

Interpretation of Remotely Sensed Images in a Context of Multisensor Fusion using a Multi-specialist Architecture.[1]

V. CLÉMENT, G. GIRAUDON, S. HOUZELLE and F. SANDAKLY
INRIA, BP 93, F-06902 SOPHIA ANTIPOLIS Cedex, FRANCE
Tel.: 93 65 78 62 Fax: 93 65 76 43 Telex: 970 050 F
Email: giraudon@sophia.inria.fr

Abstract: This report presents a scene interpretation system in a context of multisensor fusion; it has been applied to the interpretation of remotely sensed images. First we present a typology of the multisensor fusion concepts involved, and we derive the consequences of modeling problems for objects, scene and strategy. The proposed multi-specialist architecture generalizes the ideas of our previous work [GG90a] by taking into account the knowledge about sensors, the multiple viewing notion (*shot*), and the uncertainty and impresision of models and data modeled with the Possibility Theory. In particular, generic models of objects are represented by concepts independent of sensors (geometry, materials, and spatial context). Three kinds of specialists are present in the architecture: generic specialists (scene and conflict), semantic object specialists, and low level specialists. A blackboard structure with a centralized control is used. The interpreted scene is implemented as a matrix of pointers enabling conflicts to be detected very easily. Under the control of the scene specialist, the conflict specialist resolves conflicts using the spatial context knowledge of objects. Finally, an interpretation system with SAR/SPOT sensors is described, and an example of a session concerned with bridge, urban area and road detection is shown.

Interprétation d'images de télédétection dans un contexte de fusion multi-capteurs avec utilisation d'une architecture multi-spécialistes.

Résumé : Ce rapport présente un système d'interprétation de scènes dans un contexte de fusion multi-capteurs ; il a été appliqué à l'interprétation d'images de télédétection. Tout d'abord, nous présentons une typologie des concepts mis en jeu pour la fusion multi-capteurs, puis les problèmes de modélisation pour les objets, la scène et les stratégies. L'architecture multi-spécialistes proposée généralise les idées de nos travaux précédents [GG90a] en prenant en compte la connaissance sur les capteurs, la notion de points de vue multiples (*shot*), ainsi que l'incertitude et l'imprécision des modèles et des données modélisées par la théorie des possibilités. En particulier, des modèles génériques d'objets sont représentés par des concepts indépendants des capteurs (géométrie, matériaux, et contexte spatial). Trois types de spécialistes sont présents dans l'architecture : les spécialistes génériques (scène et conflits), les spécialistes d'objets sémantiques, et les spécialistes bas niveau. Une structure de tableau noir avec un contrôle centralisé est utilisé. La scène interprétée est implémentée par une matrice de pointeurs, ce qui permet de détecter les conflits très facilement. Sous le contrôle du spécialiste scène, le spécialiste conflits résoud les conflits en utilisant la connaissance sur le contexte spatial des objets. Finalement, un système d'interprétation utilisant des données provenant de capteurs SAR et SPOT est décrit ; un exemple de session montrant la détection de ponts, zones urbaines et routes est commenté.

[1]This work is in part supported by AEROSPATIALE, Department E/ETRI, 1 Rue Pablo Picasso F-78114 MAGNY-les-HAMEAUX, FRANCE, and in part supported by ORASIS contract, PRC/Communication Homme-Machine.

1 Introduction

An extensive literature has accumulated during the last decade on the problem of scene interpretation, especially for aerial and satellite images. Among these publications, one can cite the survey article by Binford [Bin82] on knowledge-based image analysis systems, the work of Nagao and Matsuyama [NM80, Mat87, MH90] on interpretation of suburban scenes in multispectral imagery, those of Riseman's team on the VISIONS scene analysis system [RH84, RH89], and in particular the application of contour–region hierarchical cooperation [RIHR84] for the interpretation of airports. McKeown's work [MHM85, MWHW89] is significant because it implements a knowledge base with hierarchic control, building objects incrementally by using explicit knowledge at multiple levels of detail. McKeown uses many sources of information: multispectral images, stereo images, and a given model of the airport (position and distance constraints) onto which object hypotheses are projected. We can also cite the papers by Huertas [HN88] and Fua [Fua88]. As for us, we proposed in [GGM89, GG90a, GG90b] a blackboard multi-specialist architecture called MESSIE for extraction of buildings and roads in aerial images.

One of the main difficulties of these applications is the knowledge representation of objects, of scene, and of interpretation strategy. Previously mentioned systems use various knowledge such as object geometry, mapping, sensor specifications, spatial relations, etc... The knowledge representation is also very various including production rules, scripts, frames, semantic networks... The system MESSIE modelizes objects of the world using 4 points of view [GG90a]: geometric, radiometric, spatial context and functionality. The spatial context is an heuristic knowledge which connects objects among each other through a spatial relation. In [GG90b], we presented an implementation of generic spatial operators like *on, along, between etc...* In MESSIE, the types of knowledge representation used are frames (implemented as object-oriented language) and production rules.

On the other hand, there is a growing interest in using multiple sensors to increase both the reliability and capabilities of intelligent systems [LK89, HSH84, HWH88]. Typical applications involved in this growth are robotic systems, manufacturing processes, autonomous vehicles, and more generally any system which has to take decisions based upon externally sensed information. Because of the development of various image sensors (visible, infrared, SAR...), computer vision domain and especially the scene analysis field are concerned by fusion of data provided by these various sensors (multisensor fusion). However, if the multisensor fusion is a way to increase the number of measures on the world by complementary or redundant sensors, problems of control of the data flow, strategies of object detection, and modeling of objects and sensors are also increased.

Another increased problem is the management of imprecise and uncertain data; in fact, imprecision and uncertainty are more complex in multisensor fusion, but they can be helpful to model and process the information redundancy. So interpretation systems have also to be able to handle and to deal with this kind of information. Two problems arise: i) representing uncertainty and imprecision, ii)processing and combining the uncertain and imprecise information.

Several models are used in uncertainty and imprecision representation, like the Probability Theory (Bayesian Model), Evidence Theory [Sha76] and the Possibility Theory [Zad78] for uncertain information, fuzzy sets and linguistic variables for imprecision. Each of these models have its own operations to combine and process information. Much work in classification or interpretation has investigated this problem. [DVV92] presents a blackboard system which has been developed for biomedical, industrial and remote sensing applications in which models and data structures are based on the use of a fuzzy approach. The Dempster-Shafer Theory [Sha76] is known in the field of expert systems to combine the judgments of different experts; it is used in [COF+91] to validate segment matching in multi-temporal images. VISIONS team has investigated the Dempster-Shafer Theory to be the bases of evidential reasoning, as well as an unified computational framework for generating and combining evidence based on the concept of a plausibility distribution [HR87]. Another approach is used in [DZZ92] where uncertainty is represented by an importance degree attached to each rule in the knowledge base and a likehood degree attached to the facts during reasoning; a MYCIN like rule [BS84] is used to combine uncertainty. In SIGMA a very simple reliability computation method (addition of fact reliability values) is used; first all possible interpretations are constructed in parallel, and then the best one is selected.

This report presents a blackboard multi-specialist architecture for scene interpretation in the context of multisensor fusion. This architecture is currently used for remotely sensed images, and generalizes the ideas of MESSIE. We show the highly modular structure of the system, and the ease of application developments.

First of all, we present an overview on the sensor fusion domain. We structure involved concepts in three levels: sensor juxtaposition, sensor selection and intermediate fusion, our system being based on the latter one. Then, section 3 introduces the modeling problem (real world and interpretation modeling). We derive the fact to stand in need of object representation independently of sensors, of knowledge about sensors, and of multiple viewing notion. In section 4, the architecture implementation is detailed. First, knowledge is structured in three representation levels: architecture, knowledge base and fact base. Secondly, characteristics of the architecture are described: generic, semantic and low level specialists, and communication mechanisms. Thirdly, the different implemented strategies are presented. Possibility Theory is used to model uncertainty and imprecision, and to validate object hypotheses. To conclude this section, interpretation steps of the system are presented. Finally, in section 5, an application of SAR/SPOT fusion developed with this architecture is presented including object and sensor modeling, and implemented specialists. Results of an interpretation session are shown.

2 Concepts Involved in Multisensor Fusion

A lot of different approaches are used in multisensor fusion: statistical theory [HDH82, Wu85, WE88], Dempster-Shafer theory [GLF81, LRS87, LRG87] or neural network architecture [Jak88].

Moreover, fusion is performed at different information levels such as pixels [Wu85] or image features (tokens) [HRW88, MWHW89, MH90]. But only some authors try to formalize the problems such as interactions between information streams from sensors, and control of the information flow. Luo *et al* [LK89] tries to classify multisensor fusion approaches by making a distinction between data fusion and data integration. Henderson *et al* [HSH84, HWH88] introduce logical sensor notion. A logical sensor is either a physical sensor or a process which is only constrained by its I/O. This approach is interesting in a functional analysis because a logical sensor is only modeled by a black box with its I/O. Clark and Yuille [CY90] classify fusional methods into two concepts: weakly and strongly coupled fusion. In weak coupling, the outputs of two or more sensory modules that produce information independently are combined. These outputs are in general of the same type (homogeneous data). So such a fusional method can give a result even if any of sensory modules fails. On the other hand, in strong coupling, the operation of one sensory module is affected by the output of another sensory module. The inputs of sensory modules are in general heterogeneous. In this case, all modules are dependent, and if any of sensory modules fails, the system is down. Thus weak coupling notion can be related to those of redundant sensor and data fusion; strong coupling can be related to complementary sensor and data integration notions. So, taking into account works described above, we propose a fusion classification for modeling and control processes, which is composed of three fusion levels:

- the first level is called sensor juxtaposition. The detection algorithm acts immediately on the whole available sensor data. For example, if we have superimposed images, we can create a multispectral vector data per pixel and apply a classifier directly (example [Wu85]). This type of fusion is generaly weakly coupled [CY90]. It is an upstream fusion approach. For scene analysis, such a method is not trivial because we need a global modeling, in terms of pixels, for each object and for all sensors.

- the second level is called sensor selection. We can consider that two (or more) independent detection chains (one per sensor) are available, and fusion can consist of choosing the chain which is more accurate or more reliable. So, this approach needs a quality measure on the results to be able to compare them. The modeling problem is not fundamentally more difficult than for only one sensor. The object model is described independently for each sensor. It is a downstream fusion approach; we can link it to Henderson's logical sensor [HSH84] and to the weakly coupled fusion [CY90].

- the third level is intermediate fusion. In computer vision, objects are described in terms of features (or tokens) like regions, segments etc... and in terms of relations between these features. So, fusion processes can be made at this level [RH89, MWHW89]. In fact, two modelings can be pointed out: i) *composite object*: Case of complementary sensors. Each feature is extracted from data provided by only one sensor (example: polygonal shape depends on visible sensor, and temperature depends on infra-red sensor). So, composite object can be related to sensor selection and to strongly coupled fusion as defined above;

ii) *composite primitive*: Case of redundant sensors. Each feature can be detected by any sensor (example: a composite polygonal shape is built from data provided by visible *and* infra-red sensors). In this case, composite primitive is close to sensor juxtaposition, and is a weakly coupled fusion.

Intermediate fusion level is the most adaptive and the most general for the different applications of scene analysis. This level enables upstream as well as downstream fusion; moreover, it is the only one that can handle strongly and weakly coupled fusion. So we choose this type of fusion for our application. From this choice, we derive in the next section the modeling problem.

3 Modeling

We are concerned with an interpretation system of real scenes, performing intermediate fusion, and using an architecture based on the blackboard and specialists concepts [HR83]. In this section, we present the main modeling problems arising when developing such a system.

Various models must be used to express the *a priori* knowledge which can be descriptive or operating and may contain imprecision and uncertainty. This knowledge can be divided into knowledge about the real world and knowledge about the interpretation:

- knowledge on the real world: knowledge on objects themselves, on the relations between objects, on sensors, and on the objects in the sensor representations.
- modeling of the interpretation process: modeling of the interpreted scene, and of strategy.

3.1 Real World Modeling

For an interpretation system, *a priori* knowledge of the scene to be observed is necessary: for example, a raw description of objects which might be present in the scene. Moreover, in order to perform multisensor fusion at different levels of representation, and to use the various data in an optimal way, characteristics of available sensors are also to be known: this enables the selection of the best ones for a given task. In the following, we first develop object modeling, then sensor modeling.

3.1.1 Objects to Detect and their Relations

Detecting specific objects (for example, rivers or fields) involves *a priori* knowledge on the specific type of object. Usually, in single-sensor systems, two main descriptions are used: the

geometric description, and the radiometric one. Criteria on the geometric aspect of the object can be, for instance, a raw description of the geometric shape such as rectangular or round, elongated or compact shape, or a range of various dimensions of the objects. The radiometric description is generally a description of how objects appear in the image (a grey level interval, or a characteristic such as pale or dark). These two criteria can be used to detect an object (by allowing the choice of the best-adapted algorithm, for instance), or to validate the presence of an object (by matching computed sizes with model sizes, for example).

In a multisensor system, the distinction must be made between knowledge which is intrinsic to an object, and knowledge which depends on the observation. Geometric properties can be modeled on the real world, however the geometric *aspects* have to be computed depending on the sensor. Concerning radiometric properties, there is no intrinsic description; radiometric descriptions are sensor-dependent. In fact, only the observation of an object can be pale or dark, textured or not. Thus, the notion of material has to be introduced to describe an object intrinsically. Materials describe the composition of an object: for example, a bridge is built of metal, cement and/or asphalt. So, radiometric properties of an object can be deduced from its composition: an object mainly made of cement, and another one mainly made of water would not have the same radiometry in an image taken by an infra-red sensor. These descriptive criteria (geometry, material components) can be used in a deterministic way.

Another intrinsic sensor-independent knowledge very important in human interpretation of images is spatial knowledge, which corresponds to the spatial relationships between objects. Spatial knowledge can link objects of the same type, as well as objects of different types. For instance, we know that a bridge passes over a river, a road or a railway, and a building is often beside another building. The heuristic spatial knowledge can be used to facilitate detection, validation, and conflict solving among various hypotheses. For example, as we know that a bridge will be over a road, a river or a railway, it is not necessary to look for a bridge in the whole image; the search area can be limited to the roads, rivers and railways previously detected in the scene. In multisensor interpretation, we can even detect the river on one image, and look for the bridges on another one. Moreover, we are more confident in the detected objects having spatial relations verified with other objects; for instance, if roads go through a region which is hypothesized as an urban area, it will reinforce our confidence in the region label as an urban area. In the same way, we use the neighboring objects to discriminate between a building and a field.

These three kinds of knowledge (geometric, radiometric, and spatial relations) have not the same importance for every object type. To validate a river hypothesis for example, the geometric criteria is more important than the radiometric one and the spatial relation one. On the other hand, the spatial relations of a bridge are more important than its radiometric and geometric aspects.

Furthermore, object descriptions (geometric and radiometric deduced from materials) are often imprecise, because, in most cases, objects to detect have not an exact model describing their features. For example, the width of a river can be very variable: different rivers have

different widths, and the width of each river can vary all along its course. Moreover, there is not an exact minimum and exact maximum values for a river width. The same can be applied on the radiometric aspects of a material in a given sensor. On the other hand, spatial knowledge can be uncertain. For example, if we say that airports are generally near urban areas, it does not mean that an airport must be found near every urban area. As seen in the introduction, several models can be used to model and process uncertainty and imprecision. The Bayesian Model constrains the use of probability to express the uncertainty; but in scene interpretation we generaly have not this knowledge. Evidence Theory [Sha76] is useful to represent uncertainty and to express the ignorance case; but this theory does not provide handling and processing of imprecise events. Possibility Theory [Zad78] presents a flexible model to express and combine uncertainty; it has many common representations with fuzzy sets imprecision modeling [DP80]. As, in scene interpretation many events are uncertain and imprecise at the same time, we choose these two approches to model uncertainty and imprecision. This allows the processing of uncertainty and imprecision under the same formalism.

3.1.2 Sensors

Some sensors are sensitive to object reflectance, others to their position, or to their shape... Radiometric features mainly come from materials whose objects are composed of, and more precisely from aspects of these materials such as *warm, homogeneous, rough, textured, smooth....* The response to each aspect is quite different depending on the sensor. Among the main sensor characteristics, we can cite the sensitivity to the aspects of various materials, to the geometry and the orientation of objects, the band width , the imprecision of measurements, and the type (*active* or *passive*).

Due to their properties, some objects will be well detected by one sensor, and not by another one; other objects will be well detected by various sensors. To be able to easily and correctly detect an object, we have to choose the image(s), i.e. the sensor(s), in which it is best represented. For that, the sensitivities of the sensors, and the material composition of objects are used. In fact, a sensor is not sensitive to a given material but only to specific aspects of the material; for instance, the response of an infra-red sensor is depending on the heat of the material; such a sensor is non-sensitive to the aspect *cold* (the amplitude data value is weak). The same applies to the orientation of the objects. A sensor can be particularly sensitive to objects which have a perpendicular position in relation to the sensor; this is the case of a radar sensor. Knowing the position of the sensor, and its resolution is also important to be able to determine whether an object would be well detected.

3.2 Interpretation

Concerning scene interpretation process, two main problems arise: how to represent the scene being interpreted, and what is the strategy to get the best interpretation.

3.2.1 Interpreted Scene

First of all, we are going to precise what we call an interpreted scene, and which information must be present in an interpretation. Our goal is not to classify each pixel of the image; it is to build a semantic description of the real observed scene. This description must include: the precise location of each detected object, its characteristics (such as shape, color...), its relations with other objects present in the scene, and the certainty of its existence in the real observed scene. It means that the model has to represent information such as: there is certainly a building at a given location, it is pale, its shape is regular, it is surrounded by a car park, only one road accesses to the building, it is not very far from what seems like an agglomeration, the road leads in a straight line to the urban area.

To capture such information, it is necessary to have a spatial representation of the scene; in the 2D-case, this can be done using a location matrix. This representation is useful for focusing attention on precise areas using location operators such as *surrounded by, near...*, and to detect location conflicts. Location conflict occurs when areas of different objects overlap.

3.2.2 Strategies

There exist in such a system different levels of strategies: global and local strategies. A scene interpretation can be made for various goals; depending on these goals, the global strategies to apply are different. For example, to help making a map of a geographical area, all the objects present in the scene must be detected, and described precisely. But for aerial navigation, only some specific landmarks have to be detected. For the landing of an aircraft, as a first step, only the airport must be detected roughly, then only details on the area of the airport (landed aircrafts, ways...) must be determined precisely. So various global strategies are necessary: detect the maximum of real scene objects, or detect only one type of object (and even one precise object), then describe its area in detail. We also see in the last example the concept of focus of attention.

For every global strategy, three main tasks have to be managed: detection of object hypotheses, validation of these hypotheses, and solving of conflicts arising between hypotheses. Each one of these tasks can have different local strategies. The detection strategy can be bottom-up or top-down; for instance, first look for salient objects in the image, and then use this information to detect the other objects. The validation strategy can be made by a comparison between object hypotheses and models, or by using information redundancy of the different sensors. Conflict solving can be based on a top-down strategy where new low-level processes are required to verify or to precise the previously extracted features. Generally, these strategies to manage hypotheses use spatial relations between objects.

4 Architecture Implementation

In this section, we present the system we have implemented. Our goal is to develop a general framework to interpret various kinds of images such as aerial images, or satellite images. Such a generic system has been designed as a shell to develop interpretation systems: it offers knowledge representation capabilities to express knowledge about an application in an explicit way. The system has been developed using the SMECI expert system generator [II91], and the NMS multi-specialist shell [Cor89, CAN90]; it is based on the blackboard and specialist concepts [HR83]. SMECI and NMS are both flexible, and have a user-friendly interface. Different strategies of interpretation can easily be tested on various images provided by several sensors, aiming at detecting different objects.

The architecture is presented in figure 1. The blackboard approach has been widely used in computer vision [HR87, MH90], and in multisensor fusion [SST86]. We have simplified the blackboard structure presented by Hayes-Roth, and we have built a centralized architecture with a specialist managing the strategy at scene level, and with one specialist per kind of object to detect. The various specialists work at different levels of representation. They are independent, and work only on a strategy request. So the system is generic and incremental: the specialists do not need to know each other to communicate.

In the next sections, we present how knowledge is explicitly described in the system, the components of the architecture, and the interpretation steps.

4.1 Knowledge Representation

Interpretation needs *a priori* knowledge; we showed in section 3 that it can be divided in knowledge concerning the real world (knowledge on objects themselves, on the relations between objects, on sensors, and on the objects in the sensor representation), and knowledge involved in the interpretation process (modeling of the interpreted scene, and of the strategies). An important distinction have to be made between knowledge which is application-independent (architecture), knowledge which is application-dependent (knowledge base), and the representation of the current problem to solve (fact base). So, three levels of representation exist:

1) architecture level: the structure to describe a kind of object in a generic way (*class*). For example: *O.semantic_object, O.scene,* or *O.sensor.*

2) knowledge base level: the effective description of a kind of object (*sub-class*) which is represented by a prototype. For example: the sub-class *Urban_area* of the class *O.semantic_object,* or the sub-class *O.infra_red* of the class *O.sensor* (see examples in section 5).