Scene Understanding from Aerospace Sensors: What can be Expected ?

S. Herbin, F. Champagnat, J. Israel, F. Janez, B. Le Saux, V. Leung, Michel Aillerie

S. Herbin , F. Champagnat, J. Israel, F. Janez, B. Le Saux, V. Leung, A. Michel
(Onera)

E-mail: stephane.herbin@onera.fr

Scene Understanding from Aerospace Sensors: What can be Expected?

Automated scene understanding or interpretation is a fundamental problem of computer vision. Its goal is to compute a formal description of the content and events that can be observed in images or videos and distribute it to artificial or human agents for further exploitation or storage. Over the last decade, tremendous progress has been made in the design of algorithms able to analyze images taken under standard viewing conditions. Several of them, e.g., face detection, are already used daily on consumer products. In contrast, the aerospace context has been confined to professional or military applications for a long time, due to its strategic stakes and to the high cost of data production. However, images and videos taken from sensors embedded in airborne or spatial platforms are now being made publicly available, thanks to easily deployable UAVs and web based access data repositories. This article examines the state of the art of automated scene interpretation from aerospace sensors. It will examine how the general techniques of object detection and recognition can be applied to this specific context, as well as what their limitations are and what kind of extensions are possible. The interpretation will be focused on the analysis of movable objects such as vehicles, airplanes and persons. Results will be illustrated with past and ongoing projects.

Introduction

What is scene understanding?

Scene understanding or interpretation is a traditional field of computer vision. It consists of designing algorithms able to associate data produced by image sensors with a formal informative description enunciated in a shared language. The target description is defined by the context of use and it is often reduced in practice to detecting, characterizing and locating in space and time the entities and events of interest.

The role of scene understanding in a global system is to generate a formal description that can be communicated, stored or enhanced by various agents, either artificial or human and is therefore not the ultimate output of a processing chain. This also means that the requirements for the interpretation result quality depend on the purpose that it will be exploited for.

Humans have no difficulty in describing what they see in an image or in a video and in reasoning about the cause and consequences of the observed phenomena. It is a platitude to state that this easiness is not shared by artificial devices such as computers. The expression “semantic gap” has been coined to refer to this problem and expresses the fact that the information encoded in computers does not spontaneously match the inner structure of sense-data.

One possible explanation of this difficulty lies in the complexity of the function relating data to description: the input space is a point in a vector space of high dimension – the number of pixels – that maps to an often hybrid space mixing continuous and discrete representations. The data distribution is therefore very sparse in its representational space, with no obvious regularities that can be captured in a simple form. Physical models may help by introducing some constraints, but are not sufficiently accurate or general to account for all phenomena occurring in real situations.

What is special with the aerospace context?

By the aerospace context, we mean in this article two families of data: large still images from remote sensing satellites and images or videos from airborne platforms. Aerospace data acquisition for scene understanding is interesting for several reasons: it can be discreet and non-intrusive; it allows wide area views; it can provide information from isolated regions; it can produce various viewing conditions to remove ambiguities. One of the first applications has been intelligence through image analysis: aerospace sensors for scene understanding have been deployed since the beginning of photography, for instance for tactical information gathering on the battlefield using aerostats. Similar applications include surveillance or targeting. Information acquisition for search and rescue purposes after a natural disaster are currently being studied. Environment monitoring is a traditional

application of remote sensing data. The availability of large quantities of image and video data drives the need for smart archiving and retrieving schemes and the construction of semantic keys describing their content.

Sense-data in the aerospace context is specific in several respects. It is usually of high dimensions – it is not unusual to handle giga-pixel images in modern remote sensing data; it is often acquired under non-intuitive viewing conditions (nadir or oblique point of view); it may be produced by unusual sensors (radar, infrared, laser, hyperspectral, etc.). However it often comes with extra metadata, typically describing the date, the viewing conditions within a given error range, or the weather conditions. Other sources of knowledge, such as maps or other similar data with informative ground truth, can also be exploited to introduce informative priors.

Although image and video data is now made easily available thanks to huge repositories (Google, Flickr, Getty, etc.) aerospace data is still scarce. A first explanation is that the aerospace context has always been strategic for governments and is therefore carefully controlled. A second explanation lies in the cost of producing good quality aerial or remote sensing data, limiting its use to professional applications. This situation is becoming less true nowadays: flight platforms with embedded sensors can be deployed more easily, high resolution remote sensing and "bird's eye view" images are readily available on any computer connected to the Internet. This new trend was recognized in the recent and first workshop on aerial video processing (http://manufacture.nimte.ac.cn/vision/wavp/) addressing the need for automatic tools for aerospace image data analysis.

The goal of this article is to provide a broad view of the state of the art of automatic scene understanding from aerospace image and video data. The next section will give a short description of the current techniques used in modern automatic scene understanding with specific emphasis on movable objects. The further two sections will concentrate on the problems of detecting and characterizing the entities and events of interest. Outputs of environment reconstruction and data registration will be assumed and are presented in a companion article of this issue [99], or can be found in other reviews [59]. The following section will give some insight on more complex settings. The last section will give some clues in regard to the state of the art performances, in terms of accuracy and processing time. The conclusion will state a few forthcoming challenges.

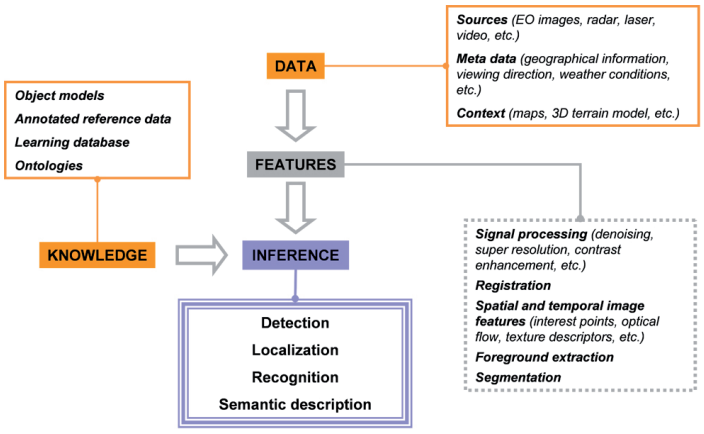

Figure 1 - General structure of processing chains for scene understanding from aerospace image sensor

Techniques for automated scene understanding

The main problem facing automatic scene understanding is a complexity issue: how to map a high dimensional sparse sense data space to a hybrid discrete and continuous interpretation. The general paradigm applied to solving this problem is to project the input data into an intermediate representation, commonly called a feature space, and then making knowledge based inferences (localization, recognition, description, etc.) from this space (figure 1).

Feature extraction + inference

The objective of such a feature space is manifold: it is expected to reduce the dimension, to make computations easier, to normalize heterogeneous data in a common framework, to reveal information, to remove noise and biases and to be invariant to known nuisance while staying discriminative in the inference space. It therefore plays a very critical role in the processing chain.

Driven by industrial vision applications, the first type of features that have been used were inspired by geometric considerations [78]. They mainly consist of simple elements, such as corners, segments or lines and are rather easy to extract. However, their lack of robustness to various illumination conditions and their limited expressive power have restricted their use in real operations for interpretation purposes.

The next generation of feature spaces were built on geometric or spatio-temporal landmarks [75][61], but were augmented with local image descriptors to better characterize the image and to introduce textural patterns in the intermediate representation [76]. This kind of feature space has higher dimensions compared with a pure geometrically based description, but is still far more compact than the original data.

Features are local, i.e., they only characterize a small part of the original data. Scene understanding is global, even if the interpretation contains local information (location of entities). Before going into an inferential step, there is a need to gather and encapsulate the local features in formal objects based on some sort of geometric extension: regions, bounding boxes, spatio-temporal tubes, 3D structure, collection of patches, etc. Many schemes have been proposed [62] [53] and evaluated for several types of data.

The next issue, given this simplified or intermediate feature space, is to produce reliable inferences to generate the interpretation in a given language or set of hypotheses. Ideally, features would be considered as direct indexes to the interpretation space. Unfortunately, it does not happen this way: features are still noisy, not sufficiently discriminative and still too complex. The inference step, i.e., the stage that actually produces the interpretation, also requires rather sophisticated processes. A popular solution is to use a learning algorithm to build the interpretation function from a set of reference data and handle the various levels of uncertainty or noise left in the feature data (see box 1).

Coupling features, models, reference data and inference

In practice, segmenting the processing chain into two uncorrelated steps – feature extraction and inference – is conceptually appealing but not optimal. Good features depend on what kind of information they carry, the quality of the information being measured by the targeted interpretation problem. Indeed, the huge volume of research

studies shows that there is no consensus on universal features, or on a general inference engine. The feature extraction and inference stages are in general designed jointly and purposely.

A first common practice making the two stages cooperate is the classical task of feature selection or construction. In this scheme, the goal of the first step is to provide an over complete set of informative features that will be selected or combined by the inference step according to an error criterion. In machine learning, the well-studied boosting family is a powerful instantiation of an integrated feature extraction + inference design (see box 1).

The inference step is heavily dependent on the nature of the reference data and on the available models. Objects represented by CAD models, collections of images, logical descriptions or deformable templates are not exploited in the same way. They constrain both the type of features that can be profitably extracted and the structure of the inference algorithm.

Prior or contextual data is often available in aerospace data: maps with various levels of semantic information or geo-referencing, viewing and weather conditions, 3D environments, knowledge representations such as ontologies, etc. They contain useful information that can be exploited to reduce the number of hypotheses to handle or map the features onto a cleaner and lower dimensional space. Though using extra sources of knowledge is appealing, there is no systematic way to introduce them into the processing chain. They can also bring their own type of noise and make data interpretation less robust if too much trust is given to their value.

Scene understanding processing chains can therefore be very complex. Though as a first approximation they all follow the same basic feature extraction and inference scheme, the research literature proposes many variations around it. One of the reasons for this high volume of research is the current performance level: with a few notable exceptions, it has difficulties to be really operational. Although improving and addressing new issues each year, it is hard to claim that automated scene understanding is a solved problem.

Box 1 - Machine learning for scene understanding

Object models have been restricted for a long time to physical models such as CAD polyhedrons with optical description of materials, illumination and viewing conditions. Physical modeling is limited – it cannot predict all of the observed phenomena in a simple form – and relies on knowledge of parameters that are hidden most of the time and must be inferred from the data or from reasonable hypotheses.

Machine learning offers a series of empirical techniques able to produce models from sample data with minimal assumptions. Its cornerstone concept is generalization, i.e., the capacity of producing meaningful inferences from unseen data. Theoretical results ensure that the empirically generated models have good properties (convergence, bounded generalization errors).

The last decade has seen statistical machine learning techniques invading the area of computer vision, especially the field of object recognition. A conjunction of events can explain this fact: new powerful learning techniques, increase in computer power, easily available digital data, stagnation in performances in pure geometric approaches and a new generation of researchers. Almost all modern and efficient scene understanding algorithms include in their processing chain a module whose parameters have been estimated by machine learning.

What is machine learning?

A collection of algorithms able to mimic a function from sample empirical data. We talk of classification when the output is discrete and of regression when it is continuous.

What are the main achievements?

Powerful algorithms and software toolboxes relying on solid theoretical grounds can now be exploited quite easily, without too much theoretical knowledge.

What are the main algorithms used in scene understanding?

Large margin (Support Vector Machines), neural networks and ensemble classifiers (boosting, bagging, random forest) for classification [17] , Kernelized Gaussian processes for regression [95] . Other more classical techniques such as Principal Component Analysis and Discriminant Analysis may produce acceptable results in low dimensional problems.

What are the limitations?

Data must be representative. No guarantee that the output will be meaningful on outliers or biased data, although several techniques are currently being developed to handle this problem.
And what are they in the aerospace context? The availability of data mainly limits the use of machine learning techniques to on-line approaches, or generic problems (person and vehicles detection). Another limitation is the processing time required by several approaches.

The following sections will present in more details the achievements to date and the specificities of several scene understanding problems. The first logical step of scene understanding will be described first: detecting the presence of entities of interest and locating them. However, as will be made clear, the detection step may require more specific object descriptions, implying that recognition is often logically antecedent to detection: bottom-up (detection) and top-down (characterization) processes are intimately linked in a global interpretation loop. Detected entities have some qualities and therefore carry pieces of information that must be revealed: how to extract this information and how to communicate it will then be presented. The aerospace context offers unconventional ways to acquire sense-data that will finally be described.

Object detection and localization

Detection is the first logical objective of image data interpretation, since it reveals the presence of entities of interest. Aerospace sensors are used in multiple situations and produce data of various types and qualities. Non-conventional sensors can make use of the specificities of the aerospace sense-data acquisition mode (see box 2). Detection algorithms depend mainly on the apparent object size: entities observed as fewer than ten pixels are not handled in the same way as entities spanning thousands. Moving objects are also a special case. The sections below will present a broad view of these issues.

Small objects

In many automatic surveillance activities, objects in the sensor range appear small or even unresolved. Early detection of such objects is fundamental, in order to perform higher level recognition tasks by pointing a better resolved sensor onto them. The scope of this section is detection from an image or a sequence of images. An example of such an application considered at Onera is the detection of objects on a runway (DROP project): in such a context, a 2cm part fallen on the runway appears as less than the size of a pixel on the image because of the wide field to be covered. When one considers small objects, very few features can be used from the object itself; typically, in many application contexts, only its position and intensity are available. Conversely, image backgrounds may have a huge variety of appearances; being non-stationary temporally and spatially, this variation emphasizes adaptive processing, where background behavior on a current image is deduced from previous images, or from spatial neighbors [73].

We have developed different methods based on building background statistics against which a pixel or a group of pixels is tested. The first one [29] considers a “detection by rejection” method. In this context, we have proposed several statistical approaches to correctly estimate a model of the environment. In particular, we have proposed the use of a mixture of densities, to guarantee a good estimation in case of background transition. An example of such processing is given in figure 2, on a SAR image. A second trend of research consists in building robust means for background pixels, by fetching pixels in a much wider area than the usual local windows and weighting them, using a patch similarity measure like the Buades “Non-Local means” [24]. The approach, denoted detection by NL-means (D-NLM), proved to be very efficient on non-stationary structured backgrounds, such as clouds [39].

When the object is moving w.r.t. the background, one can take advantage of image sequences, instead of single images. In such cases, a very simple cascade of motion compensation, threshold and short-term temporal association yields very good performances, provided that the background motion is described by simple parametric models. We showed that a particular spatio-temporal extension of D-NLM deals very efficiently with the alternative complex motion context, requiring only rough motion compensation [39].

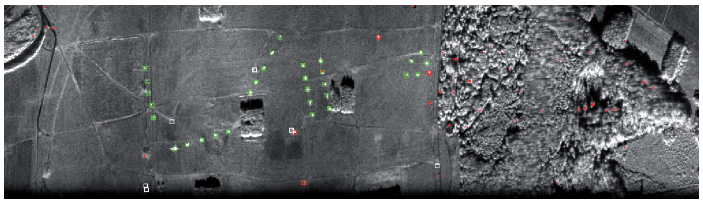

Figure 2 - Detection result on a SAR image.
Green squares mark good detection, red squares indicate false alarms.
(better viewed by magnification on screen)

Extended objects

When the resolution is sufficiently high, object appearance is associated with a region typically of the size of several hundreds to several thousands of pixels in an aerospace context. This means than the object appearance contains textural information, but this intermediate size also dismisses approaches relying on fine details.

The algorithms for object detection conform to a structure in three successive stages sharing the load of calculation:

- Saliency detection. This computes a low-cost list or regions potentially containing an object, with high level of confidence. It may use “generic” object models for persons or vehicles.
- Detection by recognition. This computes a confidence measure or a likelihood for each region, given the object models. It possibly provides a segmentation mask of the target to improve localization.
- Spatial filtering. This ensures that the detected objects have a consistent spatial distribution. This step eliminates obvious artifacts, by using simple and inexpensive rejection mechanisms.

The first step requires fast algorithms: stationary filter banks [26] [70], cascades [60] [111], coarse to fine schemes [11] , etc. It often greatly benefits from a parallel implementation (GP-GPU). The challenge is to identify salient structures from a background that is also structured. The algorithms are based on features able to discriminate textures or patterns characteristic of artifacts from those corresponding to natural elements. When the ground resolution is unknown, it is usually necessary to perform multi-scale analysis (pyramid filters, wavelets, scale space, etc.) to scan the potential size of the observed objects (figure 3). These techniques are in general increasing the number of false alarms to be filtered in the subsequent phase. In the case of urban settings, containing many artifacts (buildings, infrastructure), the exploitation of registered maps and the knowledge of image resolution can be a great help to constrain the decisions.

Though the first step is to be considered purely bottom-up, the second step (detection by recognition) implements a top-down process. It uses more sophisticated features to characterize the hypothetical object, or to reject it. It often contains an extra class for background identification, or a rejection scheme able to state that a given instance is out of the scope of the hypothesized objects. The models used to compute the likelihood are similar to those developed for object classification or categorization: configuration of parts [35], accumulation scheme [38] [63], histograms of local features[62] [90].

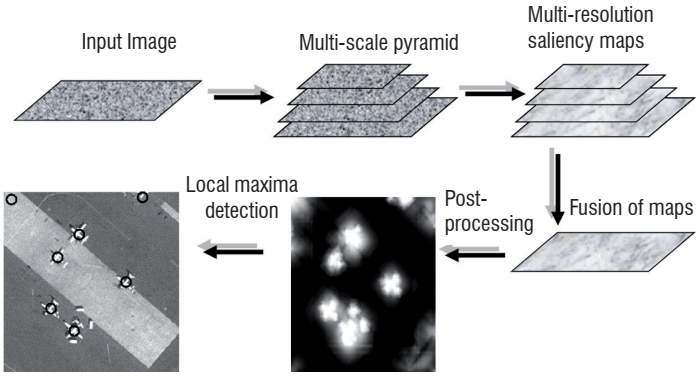

Figure 3 - Typical chain for multi-scale saliency detection. The saliency maps are built using a combination of contrast detectors (Gabor filters) and local scale estimation [26].

Moving objects

Detecting and localizing moving objects in the scene is fundamental to many higher-level interpretation processes, both in the aerospace and the security context. Examples include detecting other airplanes while airborne, maritime surveillance, pursuit of land vehicles from the air, as well as visual surveillance where either individuals or faces must be detected and localized over time.

The detection of moving objects can be divided roughly into two phases: acquisition, which reveals the presence of an object of interest and pursuit, which filters its location over time.

Many existing acquisition approaches rely on change detection and inter-frame correspondence, i.e., comparing the incoming image with a reference, which could range from a single image to a model obtained from data. However, the underlying assumption of this approach is that the sensor is fixed in location, which severely limits its applicability in the aerospace context. Image registration techniques can compensate for the apparent motion between frames (see article [99] in this issue), but cannot get rid of the parallax phenomenon, i.e., the apparent motion due to tri-dimensional structures in the scene. Furthermore, objects may enter and leave the observed area because of the moving platform, breaking the time continuity of its appearance. The literature solves these problems in two ways.

The first idea is to mix the two phases of object acquisition and pursuit. Object features are learned either offline or online to allow redetection and localization during runtime. Approaches involving offline learning are applicable if the appearance of the objects of interest does not change overtime and if sufficient training data is available. However, in order to handle possible variations of the objects of interest, adaptive techniques (i.e., online learning) are required to incrementally update their representation. Multiple variations around this scheme have been proposed, especially in the case of single object tracking ([40]). The detection and tracking of multiple objects from moving sensors, exploiting machine learning techniques, have been studied more recently [22]. These approaches rely on a good appearance characterization and are therefore more suitable to rather large extended objects.

A second idea is to filter out the residual noise after motion compensation, either by introducing object motion models and/or contextual information [9]. These techniques are applicable to smaller size objects. Longer time scale filters allow a more global analysis on blocks of data [113], or post processing on elementary tracks [71]. We have proposed to introduce a light learning step to take into account local context estimation in the inference [41] to specialize the processing chain under various viewing conditions and scene contents easily. Figure 4 shows several detection results on very different styles of image content.

Figure 4 - Results of moving object detection using local context estimation on videos with different viewing conditions [41].

Box 2 - 3D object recognition based on laser data

Ladar (Laser Detection And Ranging) is a powerful technology that provides direct access to three-dimensional information. Originally limited to macroscopic meteorological and atmospheric research, recent technical developments of ladar (miniaturization, lower cost of ownership and maintenance, eye-safe operation, performance) has led to considerable diversification of potential application including 3D object recognition.
Recent competitions on this topic (see, for instance, SHREC, http://www.aimatshape.net/event/SHREC) have encouraged the development of efficient 3D recognition algorithms. The following approach (representation / feature extraction / database inspection / model matching) is quite generic and representative of most algorithms from the state of art [16][23][68][87][86].

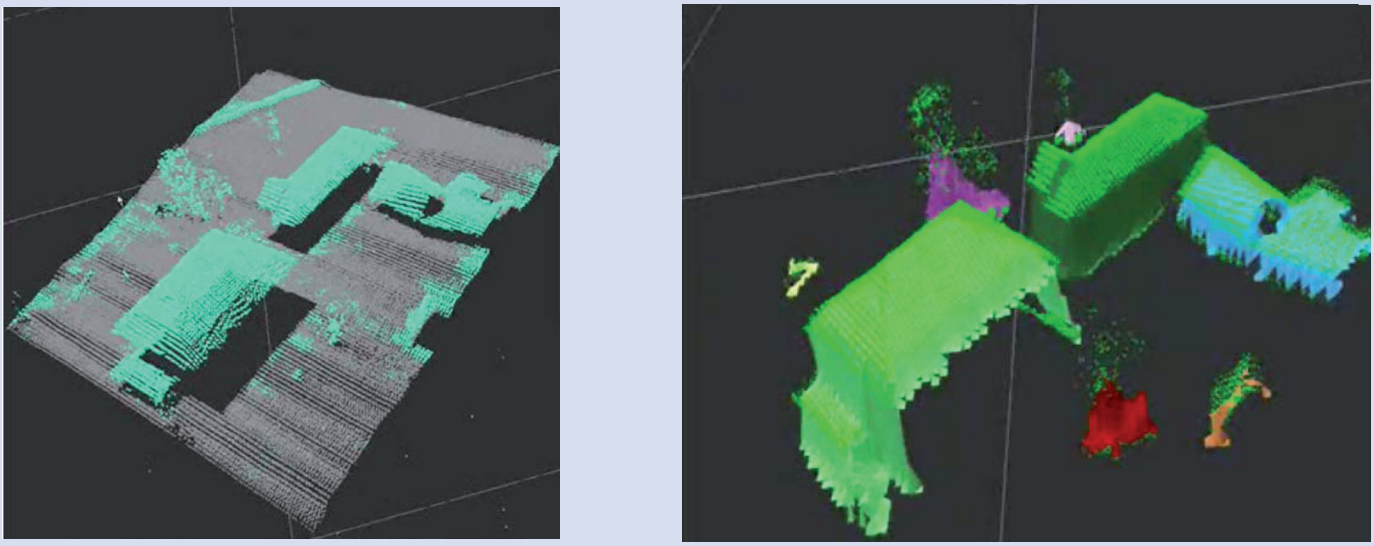

Figure B2-01 - 3D laser acquired by the Onera UAV Ressac. Detected clusters of 3D points are shown in green (left picture) and allow the object extraction phase (right picture).

Ongoing research is currently focused on the following topics:

- The model representation: It should present simultaneously a highly discriminative power and a limited complexity.
- The constitution of the database. Its structure determines the speed and performance of any algorithm for automatic and real time recognition.
- The efficient and robust extraction of characteristic features of any object, even in the presence of noise or clutter.
- Effective recognition (matching) from the set of characteristic elements that has been introduced previously into the database.

Current developments are aimed at improving and embedding 3D recognition algorithms for the exploitation of data acquired by Onera UAVs. (see figure B2-01).

Content description

This part will concentrate on what can be inferred once the entities of interest have been detected in space and time. It will concentrate on three types of interpretation: 3D object classification, action and behavior description and it will present a focus on the prospective research area of high level semantics.

Object recognition

Object recognition is a rather imprecise expression referring to various types of discrete decisions from image data. In this section, we will focus on two types of functions: classification according to a given list of hypotheses and re-identification based on a similarity measure. The typical targets of choice are vehicles, which have generated many studies driven by traffic surveillance, battle field situational awareness or intelligence applications. Person recognition in an aerial context has been less investigated [87], but the increase in video resolution is likely to stimulate new approaches. To face the challenging conditions of the aerospace context, various approaches have been proposed. We will focus on the solutions that can be deployed in practical situations.

Classification, i.e., the choice of a hypothesis from a set of possibilities – category, object model, brand, aspect, etc. –can be used as a final interpretation or as a means to filter out outliers using a rejection mechanism. It is a critical function in image understanding and has therefore received considerable attention. In re-identification, the set of hypotheses is a list of previously observed objects and relies on a similarity measure or on a conditional likelihood: it is mainly used to associate observations at different dates between distant fields of view to increase the temporal continuity of interpretation.

Aerospace sense-data (still images or videos) can be obtained from various sources, the sole common feature of which is the fact they are located above ground, from low-flying micro-uavs to satellites. This implies a wide variety of viewing conditions: points of view are usually oblique, which leads to unusual appearances of the objects. One of the main difficulties is the 3D nature of objects and the variety of appearances that a single object can produce. Moreover, aerospace data is often noisy, due to motion blur, bad color calibration, low contrast or saturations, bad focus, etc. A second practical issue is the need for an object model that can handle all of these types of nuisance.

Three dimensional object modeling is an old problem in computer vision, and was originally studied to locate an object and estimate its position from well-defined CAD geometric models [78]. This kind of approach is limited to a specific object shape. When it comes to a more general category of objects, new types of models must be defined.

A first idea is to exploit learning-based methods and blend them with geometric models. Several approaches have been used to estimate the position of vehicles [69] [88], by learning the visual appearance from various 3D points of view and the 3D geometric relationships to produce a global model [101] or by trying to fit a complete 3D appearance model [64] [56] [45]. All of these new approaches are particularly greedy for training data, a condition not often satisfied in the aerospace context. In practice, only simple models are manageable.

In object re-identification the algorithms depend mainly on the difference of viewing conditions: same sensor or not, same point of view or not. Many of the re-identification approaches in the aerial context aimed at correcting detection gaps [44]. Several approaches also make use of learning techniques to build a similarity measure between images adapted to a given setting [83] [36]. In [42] we proposed several solutions to object re-identification exploiting 3D modeling for aspect extrapolation and self-occlusion handling. Global and sparse appearance descriptions have been evaluated in an experimental set-up aimed at urban surveillance using a camera network with oblique points of view.

Action and behavior analysis

Action recognition refers to the classification of spatio-temporal patterns over a short time interval and seldom involves more than two or three agents. Usually the set of actions that we wish to recognize is defined and action recognition is the process of determining which action class the given data belongs to. Behavior analysis, on the other hand, refers to the recognition of phenomena that are more long-term and usually involves multiple interacting agents. Often, “normal” behavior is learnt from data and any deviation from this norm can be used to detect anomalies.

Many human action recognition approaches share the same global structure, where extracted spatio-temporal features of the different actions are used to train a classifier. The difference between silhouettes several timesteps apart is used as features in [94]. In [79], an action is represented by a sequence of primitive actions and learning is accomplished on a small dataset. A bag-of-words-based approach with a hierarchy of class-specific vocabularies using neighborhoods of spatio-temporal feature points is used in [57]. In [67], view-point and style independent manifolds are learnt to improve robustness. A Hough Transform-based voting where random trees are trained to learn a mapping between feature patches and spatio-temporal action Hough space is used in [112]. The problem of selecting the most discriminant part of the data by proposing an automatic optimal cropping applicable to action recognition techniques in general is addressed in [100].

Behavior analysis also relies on machine learning, except that a global representation of normal behavior is automatically extracted from the data and “anomalous” behavior can be detected by the fact that it cannot be explained by this representation. In [19], any new query that cannot be constructed using the video segment database components is classified as abnormal. Global motion flow fields are used to determine the dominant motion patterns called “supertracks” in [48]. A global representation using spatio-temporal co-occurrence between motion vectors is proposed in [103].

Figure 5 - Example of vehicle re-identification with various aspects [43].
On the left, images in the first column are “queries” that are to be matched with the color framed images. Green means good match, red means false match.
On the right, two kinds of models used to extrapolate aspects: sparse and global.

In [58], spatio-temporal features are extracted to be used in a coupled-HMM. Behavior analysis on crowds is also starting to receive much attention, its interest shown by the inaugural workshop on large crowds at the last ICCV [7]. For a comprehensive review on crowd analysis, see [51]. In the aerospace context, research on behavior analysis focuses on learning normal traffic patterns from tracking information, so that deviations from this norm can be detected as an anomaly [49] [21] [98] [77].

A behavior analysis problem that has been studied for many years is the long term trajectory characterization of humans, vehicles and airborne platforms, in the ultimate objective of supporting long-term reasoning; however, it is only recently that solutions are being proposed. The technology for tracking an object for a short time interval has reached a relatively mature level. However, maintaining continuous tracks over a longer duration is still a difficult problem, since occlusions, both static and dynamic, mean that the labels assigned to individual objects are not unique. Interactions between entities, for example the formation and splitting of groups, add another layer of difficulty to the problem.

A new trend for addressing this problem draws on the analysis of groups. Social force models have been proposed to explain the physical dynamics and groupings of individuals [74] and Pellegrini et al. proposed a joint estimation of tracks and groups [89]. We propose a solution based on Markov Logic Networks (MLN) [66][96]. An MLN is the application of a Markov network to first order logic and combines both logical statements and probabilities [109] into a single framework. This approach is promising, since it allows higher-level reasoning and multiple types of queries on the data structure. Figure 6 shows a typical complex situation with the formation and splitting of groups whose interpretation will benefit from such an approach. In such a situation, traditional trackers will return eight tracklets, but will be unable to infer the tracks of the three people. Applying the MLN-based solution here allows the three tracks to be recovered.

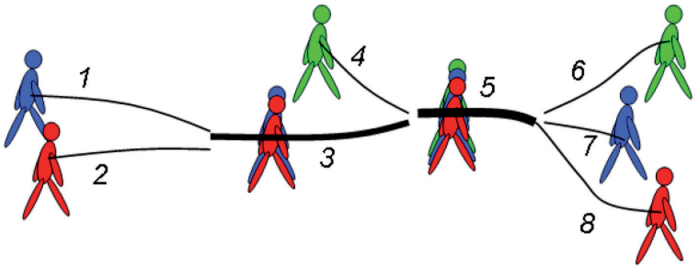

Figure 6 - Illustration of the complex behavioral patterns that can be interpreted using our Markov Logic Network approach ([66]). Tracklets 3 and 5 are tagged as groups. The resulting structured pattern allows different levels of reasoning for further interpretation.

Complex settings

The presentation above has described scene understanding issues from a single image or video stream, concentrating on the specificities of the aerospace origin of data. Acquiring images from aerospace platforms allows increased flexibility in the acquisition settings. This part discusses three of these: the exploitation of multiple points of view, of multiple sources and of multiple sensors.

Multiple points of view

Aerial sensors are embedded in moving platforms. Movement can be considered as a nuisance requiring compensation (see section on moving objects). But it can also be considered as a chance for information gathering by allowing multiple viewing conditions and therefore different ways to look at scene content.

Multiple points of view can be considered either as a redundant or as a complementary source of information. Redundancy can be used to remove noise or enhance the quality of the input signal. Interpretation performances are improved by exploitation of complementarity properties.

One of the main difficulties of vision is the management of occlusions; objects can be hidden by others and by the environment; they also occlude themselves and show to the camera only one aspect, which may be not informative for several reasons (no corresponding reference data, ambiguous appearance, or incomplete model).
The management of multiple points of view for interpretation is generally closed loop and addresses three different problems:

- Inference: what multiple view combination schemes use in order to build the final interpretation;
- Informative state modeling: how to encode and sequentially update the current information level reached;
- Control: what action or sequence of actions may improve the information state, and on what grounds.

This “active vision” approach has received a lot of attention, especially in the robotics community, since it implements a perception/action loop. The main problems addressed have been 3D environment or object reconstruction, object search in complex environment or unknown position and object recognition [97] . The problem of sensor placement and information fusion in multiple camera networks [52] [8] [104] can also be interpreted as an active vision question and will be presented in more detail in another section.

In the aerospace context, the position of an object relative to the observer, its aspect, is hard to anticipate. Each item of image data carries some information, but often not enough to discriminate between all of the hypotheses of a given set. Several studies have addressed this multiclass problem using an active recognition scheme, where the next view [30] [20] [13] [31] [32] or the observation strategy [46] [47] is optimized. Figure 7 shows an example of the influence of the number of views acquired on the recognition accuracy for a problem of vehicle classification with rather similar shapes.

Box 3 - Rich semantics perception

Object recognition achievements have considerably improved with the development of new techniques, especially the coupling of machine learning and multiple feature representations of objects. However, when the number of classes or categories to be discriminated increases, performance seems to plateau: the smallest classification error on an item of the well-studied benchmark data containing 101 categories [5] is about 30%, a performance level that cannot be considered high enough for real operational applications.

One possibility for overcoming these limitations is to increase the size of the data set and hope that learning algorithms will scale accordingly. A second idea is to consider "flat" classification as a simple instance of a more structured description language with richer semantics, where more subtle and potentially more reliable scene interpretations can be generated.

A first instantiation of this principle is simply to build a hierarchy of classes and exploit this structure, both in the algorithm design and in the output: for example, in a problem of vehicle classification, if the precise model ("Citroen C3 5 doors") cannot be issued reliably enough, a simpler higher confidence level "Hatchback" description would be preferable (figure B2-01). [108] presents a recent review of the use of hierarchies for image understanding.

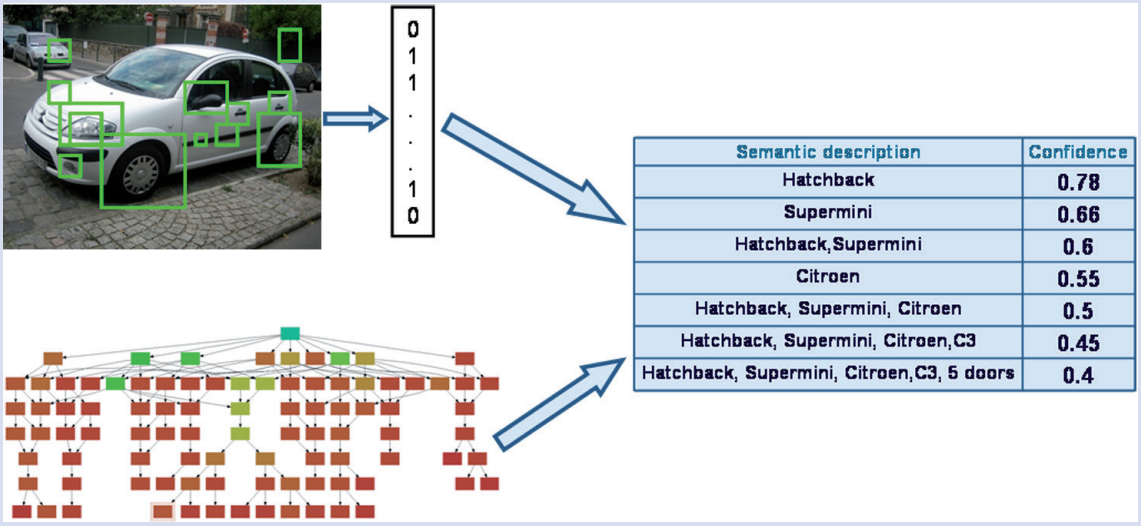

Figure B2-01 - Example of a multiple level semantic description of a vehicle. The output is a distribution of tags constrained by a hierarchy and with associated confidence. The less semantically precise it is, the more confident [107] it is.

The introduction and development of tools for richer semantics management in the description of image and video data is a rather new trend of research. It takes inspiration from various other fields, such as natural or computation linguistics, knowledge engineering, semantic web, structured data processing and multimedia database and has made concepts such as stochastic grammar [116] [93] or ontology [6] meaningful for scene understanding.

One of the key issues is the description and management of uncertainty. Indeed – this is especially true in the aerospace context – scene description is often not the ultimate output of an artificial system exploiting sense data and is likely to be exploited by other agents: since scene understanding outputs cannot be produced with infinite confidence, there is a need to provide the results in a form that contains a usable representation of uncertainty. The trade-off between confidence and complexity of description or semantic precision is one aspect of such a question.

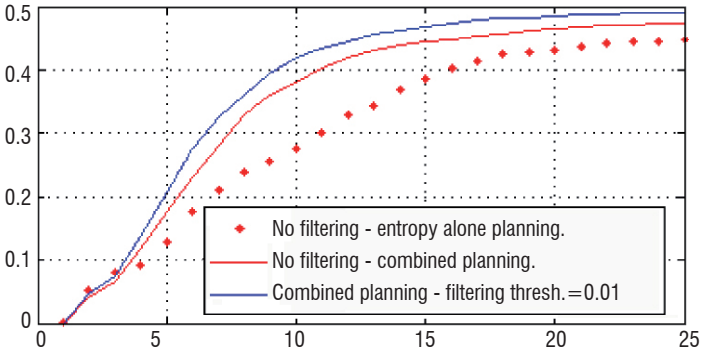

Figure 7 - One example of an active recognition problem [30]. The 8 objects to discriminate are all vehicles (left). The three view planning strategies generate various recognition rates (right).

Multiple sources

The multiplicity and variety of sensors available today that are capable of delivering complex digital information strongly encourage interest in their joint use in current or future intelligence systems. This fusion of information is a particular need in the context of a complete C4ISR chain (Communication, Command and Control, and Computers, Intelligence, Surveillance and Reconnaissance) [25] . The expected benefits are a greater capacity to analyze complex situations and robustness to the environment [18] .

Onera has been conducting research on this subject for several years. In particular, in [15] we proposed the definition of a functional architecture of SAR/optic images fusion for automatic target recognition, in a satellite or aerial context. The approach is based on the use of conventional methods of recognition, on the one hand a bottom-up method that allows us to make different assumptions on the basis of target information extracted from the images and, on the other hand, a top-down method that verifies each of these hypotheses using a matching model/image technique (see figure 8). The originality of the approach lies in the reasoning mechanisms in place. These occur mainly in the ascending phase, controlling the extraction of information by using the concepts of fusion or cooperation of sources, and, secondly, by allowing a gradual exploitation of this information

In the area of UCAV (Unmanned Combat Air Vehicle), we have developed a perception module whose aim is to increase the independence of the system by proposing fully automatic functionalities for image understanding of sensor outputs. The implemented functionalities are “automatic target detection” resulting from “fusion of the detections from 2 SAR images” and “automatic target recognition from optro-electronic (EO/IR) images”.

Thanks to its stand-off acquisition, its wide field, and its all-weather capabilities, SAR imagery is particularly well suited to detecting metallic objects in a natural environment. The method of detection is based on censoring techniques, a target being regarded as an anomaly compared to its close environment (figure 9).

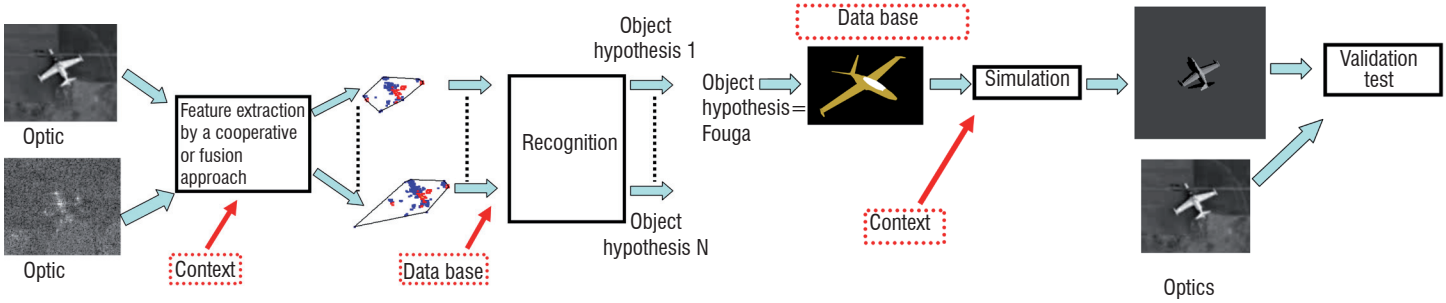

Figure 8 - Left: bottom-up process – hypothesis generation. Right: Top-down process – hypothesis verification

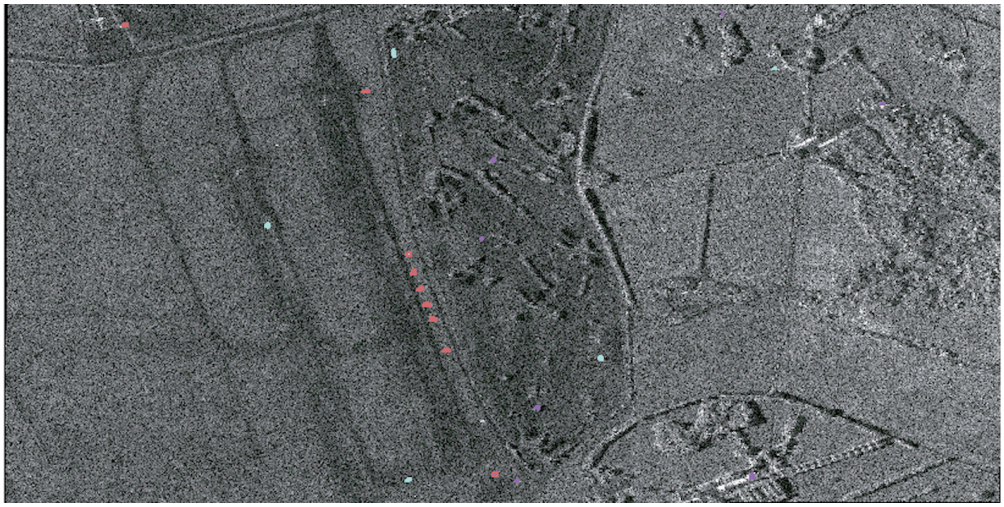

Figure 9 - SAR detection (blue: image#1 detections, purple : image#2 detections) and fusion (red)

In order to limit the amount of information coming through the data links, the result of the detection of each SAR image is transmitted as a list of points located by their geographical positions. The fusion of detections is then carried out in a decentralized scheme, producing a list of objects of interest characterized by a confidence index (plausibility).

Vehicle recognition is carried out on the basis of two triplets of high-resolution images (visible and infra-red), since the current performance of the identification process with SAR images is not sufficient to consider automation. Each triplet consists of an image acquired at nadir and two images acquired using an oblique optical axis (±45°). Recognition is based on a template matching method that uses a local planar geometric model to fit 3D models to the vehicle silhouette in the image (figure 10)[72] .

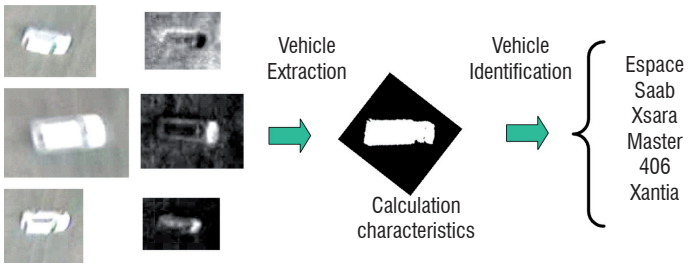

Figure 10 - Block diagram of the baseline ATR system [72].

Sensor networks

Sensor networks are often encountered in distributed systems. Their configurations range from sensors with a shared, overlapping field of view, to sensors that are non-overlapping. The primary purpose of using sensor networks is to increase the amount of information available for subsequent processing. For overlapping cameras, multiple views of the same scene can, for example, improve localization accuracy [28] [92]; for a series of non-overlapping cameras, the coverage area is increased.

One of the main difficulties in using a sensor network is to ensure correspondences amongst the cameras. On a basic level, a geometrical calibration process handles positional correspondences [105] [115] . However, for an extended network, the configuration of the cameras is also required for handling “sensor handoff” [28]. This addresses the following question: when an object leaves the field of view of one camera, which are the possible cameras with which the object can be viewed next? In addition, since the viewing conditions and the response of each camera can be different, the same object viewed by different cameras can be different. In order to associate objects across different cameras, a color calibration process is also required [10] [50] [54] [92].

Perhaps one of the most common deployments of sensor networks is in public transport networks, e.g., the use of surveillance cameras in underground systems. The sheer extent of such networks poses a challenge to scene interpretation; nevertheless, there is much a priori information and physical constraints (e.g. a train can only move according to a predefined route) to facilitate this task [65] .

The deployment of sensor networks in defense or security applications is primarily for the purpose of providing a common operating picture (COP) and for including redundancy in the system. For the networks to be scalable and be able to provide consistent and succinct information to the users, techniques of distributed and decentralized data fusion [27] [80] [81] have been studied extensively. These systems face a different set of difficulties. These include the problem of ensuring that the information provided by all the sources is trustworthy [105] [114] and the potential for reconfiguration in the event of the loss of one or more sensors [10].

Performance evaluation

A shared concern

How far are we from an operational solution? This question did not have any answer until a series of benchmark data and associated competitions were put forward [91] . Several international and national initiatives have distributed annotated data and organized image understanding competitions. The most famous and still active series of competitions are Pascal VOC [1] for object detection and recognition in internet data, TrecVid [2] video interpretation and PETS [3] for video surveillance problems. These competitions have encouraged research teams to compare their results on the same data and have prompted a global emulation. They have revealed a few things: no tested solution clearly outperforms all of the others on all interpretation problems; performances increase each year, but at a slow rate; some problems are much easier than others. Figure 11 illustrates the evolution of a 20 category detection/localization competition for three different years.

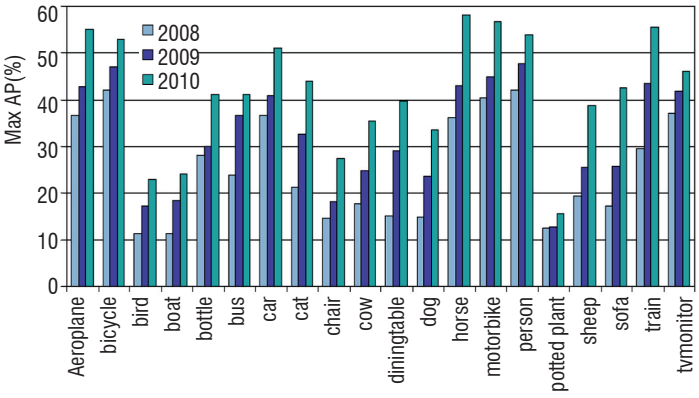

Figure 11 - Progress on the Pascal VOC challenge on three years for the best results on the object localization from 20 categories (http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/workshop/voc_det.pdf)

The aerospace context has no equivalent. The CLIF data set [4] is no longer available outside of the US. The French initiative TechnoVision-ROBIN [33] [34] is no longer maintained. With the growth of interest in aerial and satellite data, it would be beneficial for quantitative evaluation and innovation stimulation to produce and maintain comparable sets of aerospace benchmark data. Possibly the recent dataset for wide area surveillance and containing aerial video [84] will partially address this need.

Processing time

Most of the advanced algorithms for scene understanding are not real-time: the design energy is usually placed on algorithmic innovation, rather than on computing time control. However, with the constant growth of data size, fast algorithms are needed: this is especially true for videos. Several studies are now explicitly addressing the comput-

ing time issue, especially with the availability of GP-GPU devices and programming toolkits.

"Fast" image understanding algorithms claim a computing time between a few tenths of a second to a few seconds per image, in general with a parallel implementation. Several bottlenecks are still limiting: low level features used by every stage of scene understanding are, in general, time consuming and should be carefully optimized. Entity detection remains the most demanding step.

Conclusion: What can be expected?

Scene understanding from aerospace sensors follows the general trend of computer vision progress: more robust processing chains, larger domains of exploitation, higher and more refined interpretation levels and better performances both in accuracy and computing time. There has been a noteworthy evolution over the last decade, with the acclimatization of machine learning techniques and a constant development of new image features.

Are the newly developed principles and algorithms really operational? By operational we mean that the processing output can be reliably exploited by non-experts. The answer is not clear-cut. As mentioned in the introduction, simple image understanding algorithms are already used daily, but in situations where their failure is not critical. When embedded into a complex system, a failure may contaminate the entire chain and ruin the confidence in the interpretation.

One should not project too many anthropomorphic expectations on automated scene understanding: algorithms do not think or reason, and have limited experience. However, they are tireless tools, insofar as we can anticipate what they are good at. The next generation of algorithms should therefore integrate reflexive analysis and develop self-diagnosis tools ■

Acronyms

CAD (Computer Aided Design)
C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance)
EO (Electro Optic)
GP-GPU (General Purpose Graphical Processing Unit)
IR (Infrared)
SAR (Synthetic Aperture Radar)
UCAV (Unmanned Combat Aerial Vehicle)
UAV (Unmanned Aerial Vehicle)

References

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/
[2] http://trecvid.nist.gov/
[3] http://www.cvg.rdg.ac.uk/slides/pets.html, http://pets2010.net/
[4] https://www.sdms.afrl.af.mil/index.php?collection=clif2007
[5] http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[6] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
[7] http://server.cs.ucf.edu/~vision/ICCVWorkshop/home.html
[8] B. R. ABIDI, N. R. ARAGAM, Y. YAO, M. A. ABIDI - *Survey and Analysis of Multimodal Sensor Planning and Integration for wide Area Surveillance*. ACM Comput. Surv. 41, 1, Article 7, 2009
[9] S. ALI, V. REILLY, M. SHAH - *Motion and Appearance Contexts for Tracking and Re-Acquiring Targets in Aerial Videos*. CVPR, 2007
[10] Y. AL-OBAISAT AND R. BRAUN - *On Wireless Sensor Networks: Architectures*. Protocols, Applications, and Management, Auswireless Conference, 2006
[11] Y. AMIT, D. GEMAN, X. FAN - *A Coarse-to-Fine Strategy for Multiclass Shape Detection*. IEEE Trans. Pattern Anal. Mach. Intell. 26(12): 1606-1621, 2004
[12] J. ANNESLEY, J. ORWELL - *On the use of MPEG-7 for Visual Surveillance.* Sixth IEEE International Workshop on Visual Surveillance, May, Graz, Austria, 2006
[13] T. ARBEL, F-P. FERRIE - *Viewpoint Selection by Navigation through Entropy Maps.* CVPR, 1999
[14] B. BABENKO, M.-H. YANG, S. BELONGIE - *Robust Object Tracking with Online Multiple Instance Learning.* IEEE Trans. Pattern Anal. Mach. Intell. 33(8): 1619-1632, 2011
[15] C. BELIARD, R. REYNAUD, F. JANEZ - *An Architecture to Recognize Land Target in SAR and Optical Images.* Cognitive Systems with Interacting Sensors (COGIS), 2005
[16] S. BERRETTI, A. DEL BIMBO, P. PALA - *SHREC'08 Entry: 3D Face Recognition Using Integral Shape Information*. Shape Modeling and Applications (SMI), 2008
[17] C. M. BISHOP - *Pattern Recognition and Machine Learning*, Springer, 2006
[18] I. BLOCH - *Fusion d'informations en traitement du signal et des images.* Traité IC2, Série Traitement du signal et de l'image, 2003
[19] O. BOIMAN, M. IRANI - *Detecting Irregularities in Images and Video.* Int. Journal of Computer Vision, 74(1), pp.17-31, 2007
[20] H. BOROTSCHNIG, L. PALETTA, M. PRANTL, A. PINZ - *Appearance-Based Active Object Recognition.* Image and Vision Computing, 18(9): 715–727, 2000
[21] M. BREITENSTEIN, H. GRABNER, L. VAN GOOL. *Hunting Nessie: Real Time Abnormality Detection from Webcams.* Proceedings ICCV'09 Workshop on Visual Surveillance, 2009

[22] M. D. BREITENSTEIN, F. REICHLIN, B. LEIBE, E. KOLLER-MEIER, L. VAN GOOL - *Online Multiperson Tracking-by-Detection from a Single.* Uncalibrated Camera, IEEE Trans. Pattern Anal. Mach. Intell., 33(9): 1820-1833, 2011
[23] M. BRUSCO, M. ANDREETTO, A. GIORGI, G-M. CORTELAZZO - *3D Registration by Textured Spin-Images*. 3-D Digital Imaging and Modeling (3DIM), 2005
[24] A. BUADES, B. COLL, J.M. MOREL - *A non-Local Algorithm for Image Denoising*. IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, 2005
[25] C4ISR Architecture Framework, C4ISR Architecture Working Group (AWG), 1997
[26] B. CHALMOND, B. FRANCESCONI, S. HERBIN - *Using Hidden Scale for Salient Object Detection*. IEEE Transactions on Image Processing 15(9): 2644-2656, 2006
[27] K.C. CHANG, C.-Y. CHONG, S. MORI - *On Scalable Distributed Sensor Fusion*. 11th International Conference on Information Fusion, 2008
[28] R- T. COLLINS, A- J. LIPTON, T. KANADE, H. FUJIYOSHI, D. DUGGINS, Y. TSIN, D. TOLLIVER, N. ENOMOTO, O. HASEGAWA, P. BURT, L. WIXSON - *A System for Video Surveillance and Monitoring*. Technical report CMU-RI-TR-00-12, 2000
[29] D. J. CRISP - *The State-of-the-Art in Ship Detection in Synthetic Aperture Radar Imagery.* Rapport DSTO-RR-0272, 2004
[30] J. DEFRETIN, S. HERBIN, G. LE BESNERAIS, N. VAYATIS - *Adaptive Planification in Active 3D Object Recognition for Many Classes of Objects.* Workshop "Towards Closing the Loop: Active Learning for Robotics", RSS Robotics: Science and Systems Conference, 2010
[31] F. DEINZER, J. DENZLER, H. NIEMANN - *On Fusion of Multiple Views for Active Object Recognition*. DAGM, pp. 239–253, 2001
[32] J. DENZLER, C-M. BROWN - *Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2): 145–157, 2002
[33] D. DUCLOS, J. LONNOY, Q. GUILLERM, F. JURIE, S. HERBIN, E. D'ANGELO - *ROBIN: a Platform for Evaluating Automatic Target Recognition Algorithms: I. Overview of the Project and Presentation of the SAGEM DS Competition*. Proc. SPIE 6967, 2008.
[34] D. DUCLOS, J. LONNOY, Q. GUILLERM, F. JURIE, S. HERBIN, E. D'ANGELO - *ROBIN: a Platform for Evaluating Automatic Target Recognition Algorithms: II. Protocols used for Evaluating Algorithms and Results Obtained on the SAGEM DS Database*. Proc. SPIE 6967, 2008
[35] P. FELZENSZWALB, R. GIRSHICK, D. MCALLESTER, D. RAMANAN - *Object Detection with Discriminatively Trained Part Based Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, N° 9, 2010
[36] A. FERENCZ, E. LEARNED-MILLER, J. MALIK - *Learning to Locate Informative Features for Visual Identification*. International Journal of Computer Vision, Volume 77, N° 1, pp. 3-24, 2008
[37] F. FLEURET, J. BERCLAZ, R. LENGAGNE, P. FUA - *Multi-Camera People Tracking with a Probabilistic Occupancy Map*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2): 267-282, 2008
[38] J. GALL, V. LEMPITSKY - *Class-Specific Hough Forests for Object Detection.* IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009
[39] L. GENIN, F. CHAMPAGNAT, G. LE BESNERAIS, L. CORET - *Point Object Detection Using a NL-Means Type Filter*. ICIP, 2011
[40] H. GRABNER, C. LEISTNER, AND H. BISCHOF - *Semi-Supervised On-Line Boosting for Robust Tracking*, ECCV, 2008
[41] C. GUILMART, S. HERBIN, P. PÉREZ - *Context-Driven Moving Object Detection in Aerial Scenes with User Input*, ICIP, 2011
[42] J. GUINET, S. HERBIN, G. LE BESNERAIS, S. PHILIPP-FOLIGUET - *Extrapolation d'aspect pour la reconnaissance d'objets sur séquences d'images,* GRETSI, 2007.
[43] J. GUINET. *Reconnaissance multi-vues de véhicules sur séquences d'images*, PhD thesis, University of Cergy-Pontoise, Oct. 2008
[44] Y. GUO; S. HSU; Y. SHAN, H. SAWHNEY, K. RAKESH - V*ehicle Fingerprinting for Reacquisition & Tracking in Videos*. CVPR, 2005
[45] Y. GUO, C. RAO, S. SAMARASEKERA, J. KIM, R. KUMAR, H.SAWHNEY - *Matching Vehicles under Large pose Transformations Using Approximate 3D Models and Piecewise MRF model*. CVPR, 2008
[46] S. HERBIN - *Combining Geometric and Probabilistic Structure for Active Recognition of 3D Objects*. ECCV, pages 748–764, 1998.
[47] S. HERBIN - *Active Sampling Strategies for Multihypothesis Testing.* Energy minimization methods in computer vision and pattern recognition, LNCS vol. 2683, 2003
[48] M. HU, S. ALI, M. SHAH. *Detecting Global Motion Patterns in Complex Videos*, ICPR 2008
[49] W. HU, X. XIAO, Z. FU, D. XIE, T. TAN, S. MAYBANK - *A System for Learning Statistical Motion Patterns*. IEEE. Trans Pattern Analysis and Machine Intelligence, 28(9), pp.1450-1464, 2006
[50] A. ILIE, G. WELCH - *Ensuring Color Consistency Across Multiple Cameras*, Technical Report TR05-011, 2005
[51] J.C.S. JACQUES JUNIOR, S.R. MUSSE, C.R. JUNG - *Crowd Analysis Using Computer Vision Techniques*. IEEE Signal Processing Magazine, 27(5), pp. 66-77, 2010
[52] O. JAVED, M. SHAH - *Automated Multi-Camera Surveillance: Algorithms and Practice*. Springer, 2008
[53] H. JEGOU, M. DOUZE, C. SCHMID, P. PEREZ - *Aggregating Local Descriptors into a Compact Image Representation*, CVPR, 2010
[54] N. JOSHI - *Color Calibrator for Arrays of Inexpensive Image Sensors,* MS Thesis, Department of Computer Science, Stanford University, 2004
[55] S.M. KHAN, M. SHAH - *Tracking Multiple Occluding People by Localizing on Multiple Scene Planes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(3), pp. 505-519, 2009
[56] S-M. KHAN, CHENG H., D. MATTHIES, H. SAWHNEY. *3D Model Based vehicle classification in aerial imagery*. CVPR, 2010.
[57] A. KOVASHKA AND K. GRAUMAN. *Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition*. CVPR 2010
[58] L. KRATZ AND K. NISHINO - *Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models*. CVPR '09, pp. 1446-1453, 2009
[59] R. KUMAR, H. SAWHNEY, S. SAMARASEKERA, S. HSU, T. HAI; G. YANLIN, K. HANNA, A. POPE, R. WILDES, D. HIRVONEN, M. HANSEN, P. BURT - *Aerial Video Surveillance and Exploitation.* Proceedings of the IEEE , vol. 89, no.10, pp.1518-1539, 2001
[60] C. H. LAMPERT, M. B. BLASCHKO, T. HOFMANN - *Efficient Subwindow Search: A Branch and Bound Framework for Object Localization.* IEEE Trans. Pattern Anal. Mach. Intell. 31(12): 2129-2142, 2009
[61] I. LAPTEV - *On Space-Time Interest Points*. International Journal of Computer Vision 64(2-3): 107-123, 2005
[62] S. LAZEBNIK, C. SCHMID, J. PONCE - *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*. CVPR, 2006

[63] B. LEIBE, A. LEONARDIS, B. SCHIELE - *Robust Object Detection with Interleaved Categorization and Segmentation*. International Journal of Computer Vision 77(1-3):259-289, 2008
[64] M. J. LEOTTA, J. L. MUNDY - *Predicting High Resolution Image Edges with a Generic, Adaptive, 3-D Vehicle Model.* CVPR, 2009
[65] V. LEUNG, J. ORWELL AN S.A. VELASTIN. *Performance Evaluation of Tracking for Public Transport Surveillance.* Annals of the British Machine Vision Association, 6, British Machine Vision Association, pp. 1-12, 2010
[66] V. LEUNG, S. HERBIN. Flexible *Tracklet Association for Complex Scenarios Using a Markov Logic Network*. To appear in ICCV-VS 2011
[67] M LEWANDOWSKI, D. MAKRIS AND J.-C. NEBEL. *View and Style-Independent Action Manifolds for Human Activity Recognition*. European Conference on Computer Vision (ECCV), 2010
[68] X. LI, A. GODIL, A. WAGAN - *SHREC'08 entry: Visual based 3D CAD retrieval using Fourier Mellin Transform*. Shape Modeling and Applications (SMI), 2008
[69] Y. LI, L. GU, T. KANADE - *A Robust Shape Model for Multi-View car Alignment*. CVPR, 2009
[70] Z. LI, L. ITTI - *Saliency and Gist Features for Target Detection in Satellite Images*. IEEE Transactions on Image Processing 20(7): 2017-2029, 2011
[71] Y. LIN, Q. YU, G. MEDIONI. *Efficient Detection and Tracking of Moving Objects in Geo-Coordinates*. International Journal of Machine Vision and Applications, Vol. 22, No. 3, pp. 505-520, 2011
[72] B. LE SAUX, M. SANFOURCHE - *Robust Vehicle Categorization from Aerial Images by 3D-Template Matching and Multiple Classifier System*. International Symposium on Image and Signal Processing and Analysis (ISPA), 2011
[73] A. MARGALIT, I.S. REED, R.M. GAGLIARDI - *Adaptive Optical Target Detection Using Correlated Images*. IEEE Transactions on Aerospace and Electronic Systems, vol.21, n°. 3, pp. 394–405, 1985
[74] R. MEHRAN, A. OYAMA AND M. SHAH - *Abnormal Crowd Behavior Detection Using Social Force Model*. CVPR 2009
[75] K. MIKOLAJCZYK, C. SCHMID - *A Performance Evaluation of Local Descriptors*. IEEE Trans. Pattern Anal. Mach. Intell. 27(10): 1615-1630, 2005
[76] K. MIKOLAJCZYK, T. TUYTELAARS, C. SCHMID, A. ZISSERMAN, J. MATAS, F. SCHAFFALITZKY, T. KADIR, L. J. VAN GOOL - *A Comparison of Affine Region Detectors.* International Journal of Computer Vision 65(1-2): 43-72, 2005
[77] B.T. MORRIS AND M.M. TRIVEDI - *Learning, Modeling, and Classification of Vehicle Track Patterns from Live Video*. IEEE Trans. Intelligent Transportation Systems, 9(3), pp.425-437, 2008
[78] J. L. MUNDY - *Object Recognition in the Geometric era: A Retrospective. Toward Category-Level Object Recognition*. LNCS vol. 4170, 2006
[79] P. NATARAJAN, V.K. SINGH AND R. NEVATIA - *Learning 3D Action Models from a few 2D videos for View Invariant Action Recognition*. CVPR 2010
[80] E. NETTLETON, M. RIDLEY, S. SUKKARIEH, A. GÖKTOĞAN AND H. DURRANT-WHYTE - *Implementation of a Decentralised Sensing Network aboard Multiple UAVs*. Telecommunication Systems, 26(2-4), pp. 253-284, 2004
[81] D. NICHOLSON, C.M. LLOYD, SJ. JULIER, AND J.K. UHLMANN - *Scalable Distributed Data Fusion*. Proceedings of the fifth International Conference on Information Fusion, pp. 630-635, 2002
[82] J.C. NIEBLES, C.-W. CHEN, L. FEI-FEI - *Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification*. ECCV 2010
[83] E. NOWAK, F. JURIE - *Learning Visual Similarity Measures for Comparing never seen Objects*. CVPR, 2007
[84] S. OH, A. HOOGS, A. PERERA, N. CUNTOOR, C. CHEN, J. TAEK LEE, S. MUKHERJEE, J. K. AGGARWAL, H. LEE, L. DAVIS, E. SWEARS, X. WANG, Q. JI, K. REDDY, M. SHAH, C. VONDRICK, H. PIRSIAVASH, D. RAMANAN, J. YUEN, A. TORRALBA, B. SONG, A. FONG, A. ROY-CHOWDHURY, AND M. DESAI - *A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video*. CVPR, 2011
[85] R. OHBUCHI, T. TAKEI - *Shape Similarity Comparison of 3D models using alpha Shapes.* Pacific Conference on Computer Graphics and Applications, 2003
[86] R. OHBUCHI, T. SHIMIZU - *Ranking on Semantic Manifold for Semantic 3D Model Retrieval*. International Conference on Multimedia Information Retrieval (MIR), 2008
[87] O. OREIFEJ, R. MEHRAN, M. SHAH - *Human Identity Recognition in Aerial Images*. CVPR, 2010
[88] M. OZUYSAL, V. LEPETIT, P. FUA - *Pose Estimation for Category Specific Multiview Object Localization*. CVPR, 2009
[89] S. PELLEGRINI, A. ESS AND L. VAN GOOL - *Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings*. ECCV 2010
[90] F. PERRONNIN, C. R. DANCE, G. CSURKA, M. BRESSAN - *Adapted Vocabularies for Generic Visual Categorization*. ECCV, 2006
[91] J. PONCE, T. L. BERG, M. EVERINGHAM, D. A. FORSYTH, M. HEBERT, S. LAZEBNIK, M. MARSZALEK, C. SCHMID, B. C. RUSSELL, A. TORRALBA, C. K. I. WILLIAMS, J. ZHANG, A. ZISSERMAN - *Dataset Issues in Object Recognition. Toward Category-Level Object Recognition*. Lecture Notes in Computer Science. Vol. 4170, 2006
[92] F. PORIKLI - *Inter-Camera Color Calibration by Cross-Correlation model function*. IEEE International Conference on Image Processing (ICIP), 2, pp. 133–136, 2003
[93] J. PORWAY, K. WANG, S.C. ZHU - *Hierarchical and Contextual Model for Aerial Image Understanding*. Int'l Journal of Computer Vision, vol. 88, n°.2, pp. 254-283, 2010
[94] H. QU, L. WANG, C. LECKIE - *Action Recognition Using Space-Time Shape Difference Images*. ICPR 2010
[95] C. E. RASMUSSEN, C. WILLIAMS - *Gaussian Processes for Machine Learning*. The MIT Press, 2006
[96] M. RICHARDSON, P. DOMINGOS - *Markov Logic Networks*. Machine Learning, 62: 107-136, 2006
[97] S. D. ROY, S. CHAUDHURY, S. BANERJEE - *Active Recognition through next View Planning: a Survey.* Pattern Recognition, vol. 37, Issue 3, pp. 429-446, 2004
[98] J. SALAS, H. JIMENEZ, J. GONZALEZ AND J. HURTADO - *Detecting Unusual Activities at Vehicular Intersections*. IEEE Int Conf Robotics and Automation, 2007
[99] M. SANFOURCHE M, J. ISRAEL, P. CORNIC, Y. WATANABE, A. TREIL, H. DE PLINVAL, J. DELAUNE, A. PLYER, G. LE BESNERAIS - *Perception for UAV: Vision-Based Navigation and Environment Modeling*. Aerospace Lab Issue 4, May 2012.
[100] S. SATKIN AND M. HEBERT - *Modeling the Temporal Extent of Actions*. ECCV 2010
[101] S. SAVARESESE, L. FEI-FEI - *3D Generic Object Categorization, Localization and Pose Estimation*. ICCV, 2007
[102] S. SAVARESE, F. LI - *View Synthesis for Recognizing Unseen Poses of Object Classes*. ECCV, 2008
[103] Y. SHI, Y. GAO, R. WANG - *Real-Time Abnormal Event Detection in Complicated Scenes*. ICPR 2010
[104] B. SONG, C. DING, A.T. KAMAL, J.A. FARRELL, A.K. ROY-CHOWDHURY - *Distributed Camera Networks*. Signal Processing Magazine, IEEE , vol. 28, n°.3, pp. 20-31, 2011

[105] Y.L. SUN, W. YU, Z. HAN, K.J.R. LIU - *Information Theoretic Framework of Trust Modeling and Evaluation for Ad Hoc Networks*. IEEE Journal on Selected Areas in Communications, 24(2), pp. 305-317, 2006
[106] T.H. THI AND J. ZHANG - *Human Action Recognition and Localization in Video using Structured Learning of Local Space-Time Features*. AVSS 2010
[107] A.-M. TOUSCH, S. HERBIN, J.-Y. AUDIBERT - *Semantic Lattices for Multiple Annotation of Images*. International Conference on Multimedia Information Retrieval (MIR), 2008
[108] A.-M. TOUSCH, S. HERBIN, J.-Y. AUDIBERT - *Semantic Hierarchies for Image Annotation: a Survey*. Pattern Recognition, vol. 45, Issue 1, pp. 333-345, 2012
[109] S.D. TRAN, L.S. DAVIS - *Event Modeling and Recognition using Markov logic Networks.* ECCV 2008
[110] R.Y. TSAI - *A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses.* IEEE J. Robotics and Automation, 3(4), pp. 323-344, 1987
[111] P. A. VIOLA, M. J. JONES - *Robust Real-Time Face Detection*. International Journal of Computer Vision 57, pp. 137-154, 2004
[112] A. YAO, J. GALL, L. VAN GOOL - *A Hough Transform-Based Voting Framework for Action Recognition*. CVPR 2010
[113] Q. YU, G. MEDIONI - *Motion Pattern Interpretation and Detection for Tracking Moving Vehicles in Airborne Video*. CVPR, 2009.
[114] W. ZHANG, S.K. DAS, Y. LIU - *A Trust Based Framework for Secure Data Aggregation in Wireless Sensor Networks*. Sensor and Ad Hoc Communications and Networks (SECON) '06, pp.60-69, 2006
[115] Z. ZHANG - *A Flexible New Technique for Camera Calibration*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), pp.1330-1334, 2000
[116] S.C. ZHU, D. MUMFORD - *A Stochastic Grammar of Images*. Foundations and Trends in Computer Graphics and Vision, vol. 2, N°.4, pp. 259-362, 2006

AUTHORS

Stéphane Herbin received his engineering degree from the Ecole Supérieure d'Electricité (Supélec), his M. Sc. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign and his PhD degree in applied mathematics from the Ecole Normale Supérieure de Cachan. He was employed by Aérospatiale Matra Missiles (now MBDA) from 1998 to 2000. He joined Onera in 2000 and has been working since then in the Information Processing and Modeling Department. His main research interests are stochastic modeling and analysis for object recognition and scene interpretation in images and video.

Frédéric Champagnat graduated from the École Nationale Supérieure de Techniques Avancées in 1989 and received his Ph.D. degree in physics from the Université de Paris-Sud, Orsay, France, in 1993. Since 1998, he has been with the Department of Information Processing and Modeling (DTIM) at Onera. His main interests are in the field of spatio-temporal or aerial image sequences, in particular, registration, motion estimation, super-resolution and detection.

Jonathan Israel graduated from the Ecole Nationale Supérieure des Télécommunications in 2004 and received a Master's degree in mathematics from the Ecole Normale Supérieure, Cachan, France, in 2005. Since 2006, he has held a research scientist position at Onera in the Department of Information Processing and Modeling. His main interests are related to data registration, segmentation and classification for navigation or interpretation purposes.

Fabrice Janez received a PhD degree in science from the University of Angers in 1996. He has been with Onera in the Department of Information Processing and Modeling since 1997. His research topics are data fusion and image processing applied to target detection on SAR or optical images.

Bertrand Le Saux is a research scientist in Onera's Information Processing and Modeling Department. He received his Ph.D. in content-based image retrieval from the Imedia research group at INRIA (France). Before joining Onera, he worked on Bayesian inference for tomographic imaging in the Applied Mathematics Laboratory (CMLA) at the Ecole Normale Supérieure de Cachan and spent two years as an ERCIM fellow at the University of Bern and the National Research Centre at Pisa (C.N.R. di Pisa) working on image content recognition. His current research interests include machine learning, interactive mining and object detection and recognition in satellite imagery.

Valerie Leung is a research scientist in Onera's Information Processing and Modeling Department. She received her Ph.D. in Electrical and Electronic Engineering from the University of Canterbury in Christchurch (New Zealand) in 2002. After graduating, she worked on visual surveillance and data fusion in sensor networks at BAE Systems (Bristol, UK) and was a post-doctoral researcher at Kingston University (London, UK) in the EU FP6 CARETAKER project working on detection and performance evaluation. Her current research interests include image processing and computer vision.

Alain Michel received a PhD in Geography from the EHESS in 1988. He has been working at Onera since 1989 and is heading the "Multi Source Interpretation" research unit. His research field is mainly focused on image interpretation and information fusion of remote sensing data.