Scene understanding

Introduction :

Scene understanding is the process, often real time, of perceiving, analysing and elaborating an interpretation of a 3D dynamic scene observed through a network of sensors. This process consists mainly in matching signal information coming from sensors observing the scene with models which humans are using to understand the scene. Based on that, scene understanding is both adding and extracting semantic from the sensor data characterizing a scene. This scene can contain a number of physical objects of various types (e.g. people, vehicle) interacting with each others or with their environment (e.g. equipment) more or less structured. The scene can last few instants (e.g. the fall of a person) or few months (e.g. the depression of a person), can be limited to a laboratory slide observed through a microscope or go beyond the size of a city. Sensors include usually cameras (e.g. omni directional, infrared), but also may include microphones and other sensors (e.g. optical cells, contact sensors, physiological sensors, radars, smoke detectors). Scene understanding is influenced by cognitive vision and it requires at least the melding of three areas: computer vision, cognition and software engineering. Scene understanding can achieve four levels of generic computer vision functionality of detection, localisation, recognition and understanding. But scene understanding systems go beyond the detection of visual features such as corners, edges and moving regions to extract information related to the physical world which is meaningful for human operators. Its requirement is also to achieve more robust, resilient, adaptable computer vision functionalities by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, and develop new strategies for analysis and interpretation. The key characteristic of a scene understanding system is its capacity to exhibit robust performance even in circumstances that were not foreseen when it was designed (G. H. Granlund). Furthermore, a scene understanding system should be able to anticipate events and adapt its operation accordingly. Ideally, a scene understanding system should be able to adapt to novel variations of the current environment to generalize to new context and application domains and interpret the intent of underlying behaviours to predict future configurations of the environment, and to communicate an understanding of the scene to other systems, including humans. Related but different domains are robotic, where systems can interfere and modify their environment, and multi-media document analysis (e.g. video retrieval), where limited contextual information is available.

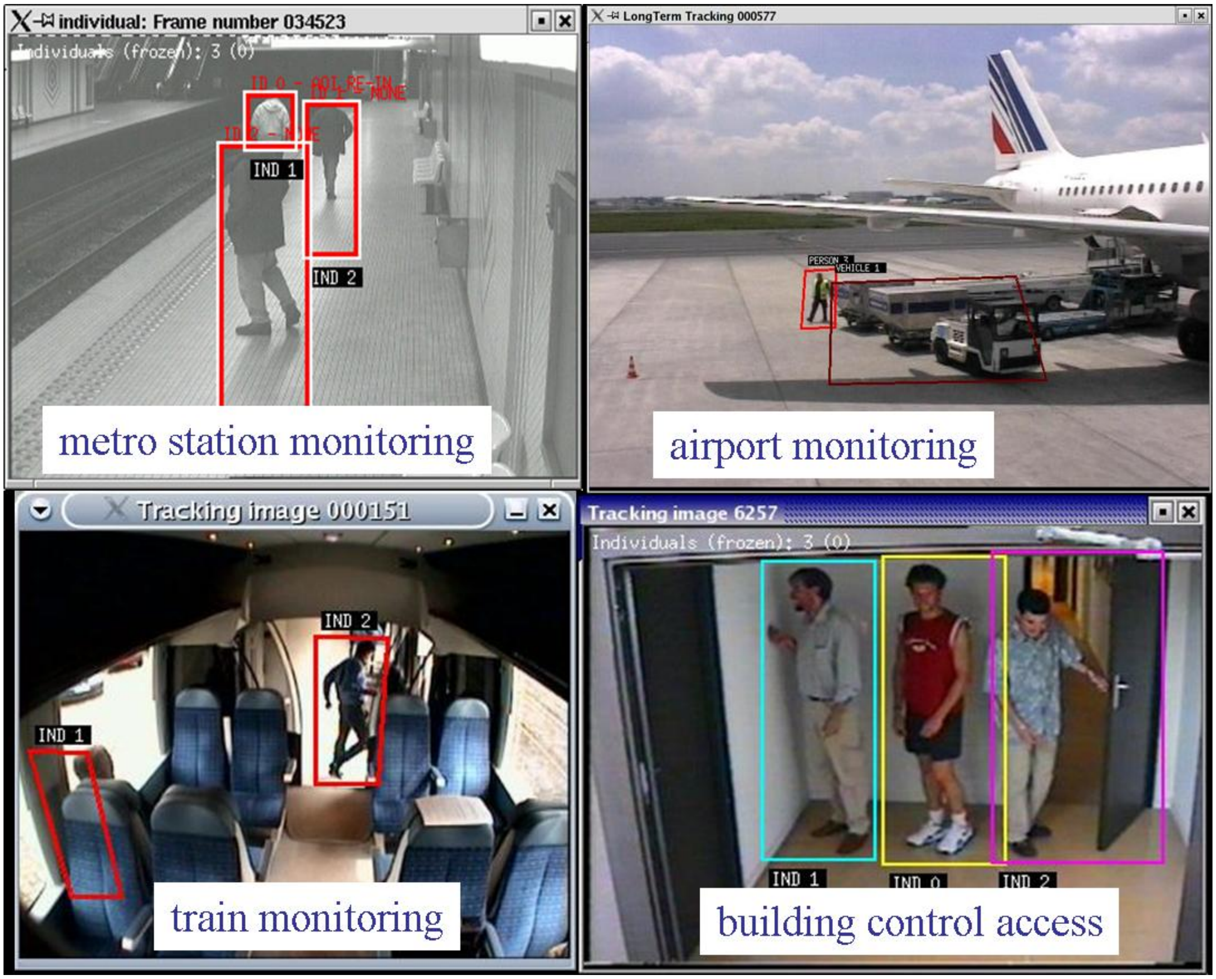

Figure 1. Video understanding applications.

Objectives :

Despite few success stories, such as traffic monitoring (e.g. Citylog), swimming pool monitoring (e.g. Poseidon) and intrusion detection (e.g. ObjectVideo), scene understanding systems remain brittle and can function only under restrictive conditions (e.g. during day rather than night, diffuse lighting conditions, no shadows), having poor performance over time, they are hardly modifiable, containing little a priori knowledge on their environment. Moreover, these systems are very specific and needs to be redeveloped from scratch for other applications. To answer these issues, most researchers have tried to develop new vision algorithms with focused functionalities, robust enough to handle real life conditions. Up to now no vision algorithms were able to address the large varieties of conditions characterising real world scenes, in terms of sensors conditions, hardware requirements, lighting conditions, physical object varieties, application objectives... My goal is to design a framework for the easy generation of autonomous and effective scene understanding systems. This objective is very ambitious; however the current state-of-the-art techniques in cognitive vision have lead to partial solutions [Cohn et al., 2006], [Dee and Hogg, 2005], [Needham et al., 2005], [Nevatia et al., 2004], [Remagnino et al., 2006], [Crowley, 2006b], [Jodogne and Piater, 2005] and [Xiang and Gong, 2006b]. I believe that to reach this goal, a holistic approach is needed where the main scene understanding process relies on the maintenance of the coherency of the representation of the global 3D scene throughout time. This approach which can be called 4D semantic interpretation, is driven by models and invariants characterising the scene and its dynamics. Scene understanding is a complex process where information is abstracted through four levels; signal (e.g. pixel, sound), perceptual features, physical objects, and events. The signal level is characterized by strong noise, ambiguous, corrupted and missing data. The whole process of scene understanding consists in filtering this information to bring forth pertinent insight of the scene and its dynamics. To fulfil this objective, models and invariants are the crucial points to characterise knowledge and insure its consistency at the four abstraction levels. For instance, I have defined formalisms to model the empty scene of the surrounding (e.g. its geometric), the sensors (e.g. calibration matrices of the cameras), the physical objects expected in the scene (e.g. 3D model of human being), and the scenarios of interest for users (e.g. abnormal events). The invariants (called also regularities) are general rules characterising the scene dynamics. For instance, the intensity of a pixel can change significantly only in two cases: change of lighting condition (e.g. shadow) or change due to a physical object (e.g. occlusion). A second rule for example, verifies that physical objects cannot disappear in the middle of the scene. There is still a open issue consists in determining whether these models and invariants are given a priori or are learned. The whole challenge consists in organising all these knowledge in order to capitalise experience, share it with others and updating it along experimentation. To face this challenge, tools in knowledge engineering such as ontology, are needed.

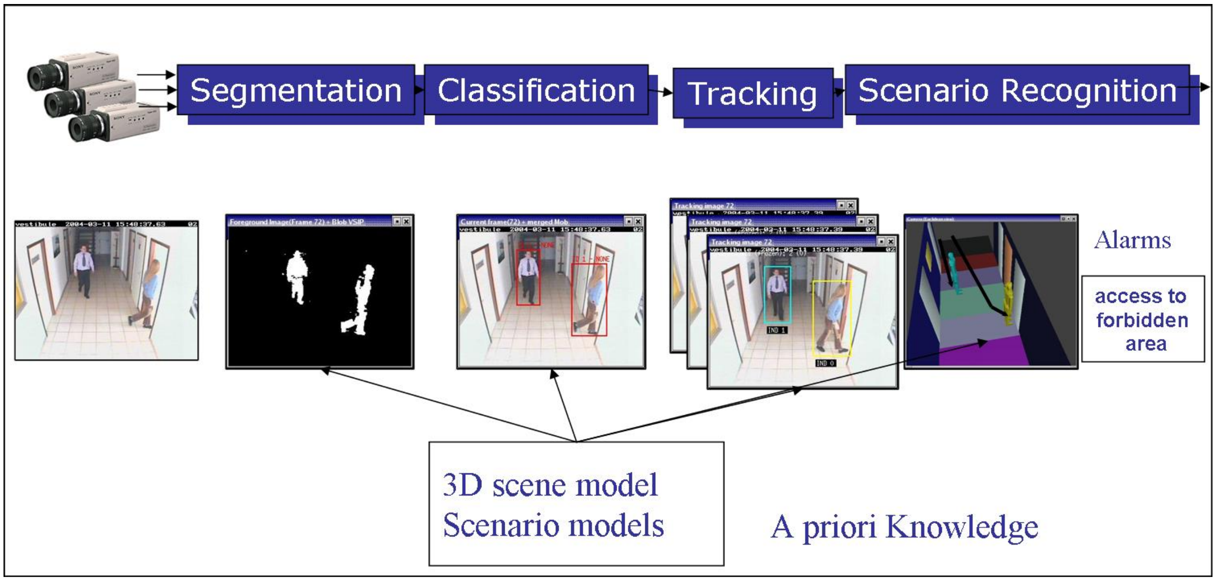

Figure 2. A scene understanding platform: real-time interpretation of videos from pixels to events.

Summary of research work :

To concretize this approach my research activities have been organised within the following five axes. For each axis, I summarize the main scientific challenges I have addressed.

Perception for scene understanding (perceptual world). A first axis is to collect and develop vision algorithms to handle all the varieties of real world conditions. The goal of all these algorithms is to detect and classify the physical objects which are defined as interesting by the users. A first difficulty consists in developing robust segmentation algorithms for detecting the physical objects of interest. The most common algorithms estimate the motion within the videos. These algorithms are based on the hypothesis that the objects of interest are related to what is moving in the video, which can be inferred by detecting signal changes. Unfortunately, these algorithms have the tendency to detect a lot of noise (e.g. due to light changes) together with the objects of interest. A second difficulty consists in extracting meaningful features characterising the objects of interest. Most of algorithms compute features relatively to the trajectory of the physical objects. Robust descriptors characterising the shape of physical objects still need to be developed. A third difficulty is to establish under which hypotheses the algorithms are valid, and to understand their limits. In the same way, algorithms processing other media and modalities (e.g. audio, contact, radar) need more development to complement the information extracted from video streams. A still open issue is to establish the precision and likelihood of these processes.

Maintenance of the 3D coherency throughout time (physical world). A second axis consists in combining all the information coming from the different sensors observing the detected physical objects and in tracking these objects throughout time. Despite all the works done in this domain within the last 20 years, fusion and tracking algorithms remain brittle. To guarantee the coherency of these tracked objects, spatio-temporal reasoning is required. Modelling the uncertainty of these processes is also an open issue. Another question that we need to answer is at which level this information should be combined. Information fusion at the signal level can provide more precise information, but information fusion at higher levels is more reliable and easier to realise. In any case, a precise formalism is needed to combine uncertain information coming from heterogeneous sensors.

Event recognition (semantic world). At the event level, the computation of relationships between physical objects constitutes a third axis. The real challenge is to explore efficiently all the possible spatio-temporal relationships of these objects that may correspond to events (called also actions, situations, activities, behaviors, scenarios, scripts and chronicles). The varieties of these events, called generally video events, are huge and depend on their spatial and temporal granularities, on the number of the physical objects involved in the events, and on the event complexity (number of components constituting the event and the type of temporal relationship). So the challenge is to explore this large event space without getting lost in combinatorial searches.

Evaluation, control and learning (autonomous systems). To be able to improve scene understanding systems, we need at one point to evaluate their performance. The classical methodology for performance evaluation consists in using reference data (called ground truth). However, generating ground truth is tiresome and error prone. Therefore, an issue is to perform the performance evaluation stage using unsupervised techniques. Once evaluation is possible, a real challenge consists in optimising the scene understanding system using machine learning techniques in order to find the best combination of programs, the best set of program parameters with the best control strategies to obtain an efficient and effective real-time process. The difficulty is three fold. First, programs depend on environmental conditions and the program optimisation process has to be dynamic to take into account of environmental changes and available resources. Second, all these programs are interlinked with each others, so the modification of one program parameter can mess the functioning of all other programs. Finally, the knowledge on these programs is not formalised and usually, even the developers cannot tell what will be the program output under even specific conditions. Another way to improve system performance is to add higher reasoning. Scene understanding is essentially a bottom-up approach consisting in abstracting information coming from signal (i.e. approach guided by data). However, in some cases, a top-down approach (i.e. approach guided by models) can improve lower process performance by providing a more global knowledge of the observed scene or by optimising available resources. For instance, the global coherency of the 4D world can help to decide whether some moving regions correspond to noise or to physical objects of interest. So, the fourth axis in my research consists in exploring program supervision (including evaluation) and machine learning techniques for the easy generation of effective real-time scene understanding systems.

Communication, Visualisation and Knowledge Acquisition (interactive systems). Even when the correct interpretation of the scene has been performed, the scene understanding system still has to communicate its understanding to the users. So user interactions constitutes a fifth axis. There are at least three types of users: program developers, experts of the application domain and end-users. The first challenge is to enable program developers to understand all specific components and in the same time, the global architecture of the scene understanding system, so that they can adapt efficiently their programs and configure and install the system on a site. To reach this goal, formalism is required to express program knowledge. Second, if we want an effective system, the a priori knowledge needs to be formalised to enable the domain experts to describe their methodology for analysing the scene. For instance, a tool and a dedicated ontology have to be provided to assist the experts in defining the scenarios that the system has to recognize. To help this process, a graphical tool can be designed to generate and visualise 3D virtual animations illustrating these scenarios. In complement of these tools, clustering techniques can be used to mine the frequent activities (i.e. event patterns or time series) occurring in the scene. Moreover, if we want the system to be used, special care needs to be brought to display what has been understood to the final users. An ergonomic interface on an adapted media (e.g. immersive reality or PDA personal digital assistant), a convenient representation of the scene (e.g. virtual 3D scene, augmented reality), and an intuitive vocabulary are the necessary devices to be provided to the end-users. Besides these devices, the system needs to take into account feedback information from the end-users to be able to adapt its performance to the specific goals of each user.

All along these years, for each axis, I have tried to establish the scientific and technological foundation for a Scene Understanding approach through the design of systems dedicated to more than 20 applications (e.g. Visual Surveillance, Activities Monitoring, Ambient Intelligence, Perceptual User Interface, Health Care, and Animal Behavior Analysis), in direct contact with users ranging from end-users (e.g. human operators, managers, domain experts), to integrators, hardware and software providers. I believe that applications are a key point in conceiving effective scene understanding systems for three reasons: first they enable to answer real challenges, second they are the necessary conditions to enable experts of the application domain to provide the precise knowledge on a scene and finally they are the main way with the help of end-users to evaluate the performance of the final system. These real world systems could not have been conceived, only with the help of 7 PhD students and 9 research engineers that I have supervised.