RESEARCH ACTIVITY

Francois BREMOND, PULSAR team

1. Summary of past research activities (1994 - 2010)

Web-site for publications: http://www-sop.inria.fr/members/Francois.Bremond/topicsText/myPublications.html

My research activities aim at designing a holistic approach for **Scene Understanding systems** which combines different knowledge types such as, extracted visual features, a priori knowledge of the 3D scene, feedback information from higher level components and learned activity models. The issue consists in proposing a general framework for taking advantage of all this information to optimize the performance of a given system depending on its requirements.
This approach is one of the few which can enable a system to recognize in real time complex human activities filling the gap between sensor information (pixel level) and behaviour understanding (semantic level).
Here, scene understanding corresponds to the real time process 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 a large variety of models which humans are using to understand the scene. This 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. Despite few success stories, such as traffic monitoring (e.g. Citilog) and intrusion detection (e.g. ObjectVideo, Keeneo), scene understanding systems remain brittle and can function only under restrictive conditions (e.g. during day rather than night, diffuse lighting conditions, no shadows). To answer these issues, most researchers have tried to develop original vision algorithms but with focused functionalities, robust only to handle a limited number of real world conditions.
Thus my research activities have consisted in designing a framework for the **easy generation of autonomous** and effective scene understanding systems. In order to achieve this ambitious objective, I have proposed a holistic approach where the main scene understanding process relies on the maintenance of the coherency of the representation of the global 3D scene throughout time [web-site, journal, 6]. This approach which can be called 4D semantic reasoning, 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. video, audio), 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. 3D position 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 are general rules characterising the scene dynamics. For instance, the intensity of a pixel can change significantly only in two cases: change of lighting conditions (e.g. shadow) or change due to a physical object (e.g. occlusion). There is still an open issue which consists in determining whether these models and invariants are given a priori or are learned. The whole challenge consists in managing this huge amount of information and in structuring all these knowledge in order to capitalise experiences, to share them with others and to update them along experimentations. To face this challenge knowledge engineering tools such as ontology are needed.

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 and the main contributions I have brought to scene understanding.
A first axis has been to 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. My main contributions have been first to design detection algorithms to separate physical objects from different categories of noise (e.g. due to light change, ghost, moving contextual object). I have also conceived a second set of algorithms to extract meaningful features (e.g. 3D HOG, Haar based descriptors, colour histograms) characterising the objects of interest. These algorithms compute features relatively to the trajectory and the shape of the physical objects. For instance, during M. Zuniga PhD, for characterizing a moving object we have proposed to use a 3D parallelepiped bounding the object detection. A third contribution has been to establish

during B. Georis PhD for which hypotheses the algorithms were valid, and to understand their limits. In the same way, my concern was to establish the precision and likelihood of these processes.
A second axis has consisted in combining all the features coming from the detection of the physical objects observed by different sensors and in tracking these objects throughout time. Therefore, I have proposed a set of algorithms for **tracking multiple objects** in 2D or 3D with one camera or a network of cameras, [web-site, conference, 13, 14]. For instance, these algorithms take advantage of contextual information and of a graph of tracked moving regions where an object trajectory can be seen as the most probable path in the graph. This property enables to process long video sequences and to ensure the trajectory coherence. Moreover, these tracking algorithms compute the uncertainty of the tracking process by estimating the matching probability of two objects at successive instants. A second contribution has been to fuse the information coming from several sensors at different levels depending on the environment configuration. Information fusion at the signal level can provide more precise information, but information fusion at higher levels is more reliable and easier to realize. In particular, I have designed three types of fusion algorithms: (1) multiple cameras with overlapping field of view, (2) a video camera with pressure sensors and sensors to measure the consumption of water and electrical appliances, and (3) coupled video cameras with other sensors (contact sensors and optical cells).
At the **event level**, the computation of relationships between physical objects has constituted a third axis. The real challenge has been to explore efficiently all the possible spatio-temporal relationships of these objects that may correspond to events (called also actions, situations, activities, behaviours, scenarios, scripts and chronicles). First, I have proposed solutions for event recognition, using different types of formalism: finite state automata, HMM and Bayesian networks [web-site, conference, 11, 22]. Second, during T. Van Vu PhD, we have designed an algorithm based on temporal scenarios recognizing in real-time activities predefined by experts and taking as input the a priori knowledge of the observed environment and the mobile objects tracked by a vision module. Concerning the issue of temporal scenario representation, we have proposed a video event ontology (in collaboration with an ARDA workshop series on video events) that can facilitate the representation of temporal scenarios. Based on this ontology, we have proposed a description language (called Scenario Description Language) helping experts of different domains to describe intuitively their scenarios of interest. This language is/has been used by experts of nine European/French projects for video surveillance and homecare. We have also extended this approach to handle (1) audio-video and other sensor data, (2) scenario uncertainty and (3) learning of scenario models.
To be able to improve scene understanding systems, we need at one point to evaluate their performance. I have proposed a complete framework for **performance evaluation** which consists of a video data set associated with ground-truth, a set of metrics for all the tasks of the understanding process, an automatic evaluation software and a graphical tool to visualise the algorithm performance results (i.e. to highlight algorithm limitations and to perform comparative studies). This framework has been used by more than 16 international teams during the ETISEO Techno-Vision project that I have led. Using this evaluation, I have proposed during A. Nghiem PhD an approach for optimising scene understanding systems using **machine learning** techniques in order to find the best set of program parameters and to obtain an efficient and effective real-time process [web-site, conference, 44, 47]. I have also proposed to improve system performance by adding a higher reasoning stage and a feedback process towards lower processing layers (i.e. approach guided by data). For instance, I have proposed an algorithm (called Global Tracker) to compute the global coherency of the tracking process on a long term basis to correct detection errors at the segmentation level [web-site, journal, 10]. The another 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, configure and install the system on a site. To reach this goal, I have proposed during B. Georis PhD a formalism to express knowledge to control program. In complement of this formalism, I have proposed with A. Toshev, L. Patino and G. Pusiol clustering techniques to be used to mine the frequent activities (i.e. event patterns or time series) occurring in the scene [web-site, conference, 50, 57, 66]. For instance, we were able to compute the most common activities of people travelling in a subway network, on an airport apron and in the apartment of an elderly [web-site, journal, 11].
All along these years, for each axis, I have tried to establish the scientific and technological foundation for Scene Understanding through the design of systems for more than 29 research projects (industrial, national, European and International) dedicated to different applications (e.g. Visual Surveillance, Activities Monitoring, Ambient Intelligence, Perceptual User Interface, Health Care, and Animal Behaviour Analysis), in direct contact with users ranging from end-users (e.g. human operators, managers, domain experts), to integrators, hardware and software providers. These systems have been conceived through a common vision platform which has been transferred first to several industrials and has fostered the creation of a spin-off Keeneo. 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 the scene and finally they are the main way to evaluate the performance of the final system.

To summarize, my objective is not only to propose original computer vision algorithms, but also to build a general paradigm to optimize vision systems (i.e. to automatically configure these systems) and to learn in an unsupervised way the activities to recognize.

RESEARCH PROGRAM

Title of research program : **Scene understanding and activity recognition**

This section aims at summarising my research directions to build an effective paradigm (as explained in my past activity) for the **easy generation of autonomous** scene understanding systems. I have structured these research directions in following the three abstraction levels of the scene understanding process: perceptual world, physical world and semantic world.

The perceptual world includes all information relative to the features (e.g. colour, edge, 2D shape, 3D trajectory, sound, contact information) describing a scene and in particular the physical objects evolving in the scene. This world is characterised by its uncertainty and redundancy. To explore the perceptual world, I have designed the SUP (Scene Understanding Platform, previously named VSIP) platform to build scene understanding systems, based on two types of programs: (1) generic programs for the main video understanding tasks and for common video characteristics, (2) advanced programs for specific tasks and for handling particular situations. Nevertheless, the extraction of perceptual features and object detection will stay an open issue for still a long period of time, in particular in real world applications containing challenging situations, such as moving cameras, crowd, and limited processing capacities.
To improve the perception of dynamic 3D scenes, I am currently working towards three directions. First, I am planning to design algorithms for computing **more reliable perceptual features** for characterising the objects of interest. For instance, I am exploring different types of visual features (e.g. feature points such as KLT (Kanade Lucas Tomasi), local 2D descriptors such as HOG (Histogram of Oriented Gradient) or Haar features (as the ones described by Viola and Jones) and other sensor features (e.g. audio, radar features and features from environmental and physiological sensors). In this objective, I supervise a PhD S. Bak [web-site, conference, 59] on **people detection in complex videos** by defining a generic visual signature of individuals. A second direction consists in designing robust algorithms for characterising **human shapes** in order to infer postures and gestures. Thanks to the proposed new perceptual features and previous work [web-site, journal, 8], these algorithms could become independent from the camera view point and still effective in complex situations (e.g. static and dynamic occlusions, moving background, crowd, interactions between moving objects and interactions with contextual objects). I am exploring these new types of algorithms with PhD P. Bilinski who is studying **gesture recognition** in complex scenes. The third direction is to establish for which **hypotheses the algorithms** are effective, and to understand their limits. This topic is currently under process with PhD A. T. Nghiem who is studying the relationships between algorithm parameters, scene conditions and algorithm performance. The objective is then to optimise the use of perception algorithms and their combinations [web-site, conference, 44, 47, 56]. Based on this algorithm characterisation, the advanced research axes proposed in the following and corresponding to the next abstraction levels can be fruitfully explored. I believe that pursuing these other axes is important for two reasons: first, in some specific conditions (e.g. structured environment), the challenges of the perceptual world can be solved and second, some applications do not require perfect vision results (i.e. perfect object detection and tracking). These other research axes constitute today strong active trends in scene understanding.

The physical world contains the information on the physical objects of the real world, especially the ones in motion. Reasoning in this world is similar to the logical inferences performed by human beings (i.e. common sense). It includes geometric, physical laws, spatio-temporal and logic. Here, the scene understanding process aims at maintaining a coherent and multi-modal representation of the real world throughout time. To study the physical world, I have proposed research work organized following two directions: coherency throughout time and coherency of multi-modal information in the 3D space. This work has brought forth some solutions to bridge the gap between signal and semantic levels. The key issues are:

- Building a common knowledge representation for combining all information describing the scene.
- Modelling and managing the uncertainty and the incompleteness of data and models characterizing the scene and its dynamics.

A natural trend in the tracking and information fusion domains consists first in extending the common knowledge representation to **combine in an easy way all information** coming from different sources throughout the time. This trend also consists in modelling all types of **uncertainty** and incompleteness for both data and scene models. Therefore, a current work is to integrate the largest diversity of sensors to get a complete multi-modal perception of the scene. This objective is addressed in two different aspects. Firstly, with PhD D.P. Chau and M. Souded [web-site, conference, 60], we are designing a framework for

incorporating common sense knowledge and logic. With this PhD we are studying an optimal configuration of the tracking algorithms in function of the application specification and of the scene conditions. Secondly, with PhD N. Zouba and R. Romdhame [Web-site, conference, 45, 54], we are building a multi-sensor scene understanding approach to monitor elderly people so that they can live longer at their home. This scene understanding system requires a **generic definition of sensors** to be able to seamlessly add a new sensor and to link the information provided by this sensor with activities to be detected in the scene. We have first monitored in our experimental laboratory (called Gerhome) 14 elderly people (half day for each person) and now we are planning to install the system in two apartments of Alzheimer patients in Nice Hospital.

The semantic world gathers all types of events, relations and concepts related to the activities occurring in the scene. This world is mostly symbolic and linked to the application domain. Reasoning in this world consists in specific causal and logical inferences and verifying spatio-temporal constraints which make sense only in a particular domain with a specific objective. For the semantic world, I have proposed two types of approach for recognising numerical and symbolic events.
The numerical approaches are mostly graphical methods based on a network which nodes correspond to combinations of visual features. They are well adapted to simple events closely related to vision features involving few physical objects (mostly one individual with or without interactions with his/her environment). However, as the parameters are learned with training video sequences containing positive and negative event samples, the effectiveness of these approaches often depends on this training phase. Therefore, two **learning mechanisms** are needed to ease the construction of event recognition algorithms: (1) to select the training video sets and (2) to learn the structure and the parameters of the network. In this objective, a first goal is to propose a set of **generic primitive event concepts** and a mechanism to link them to the output (e.g. posture and trajectory distribution) of specific algorithms and to a priori knowledge including the scene context (e.g. to be close to a seat or in a narrow corridor). For instance, these event concepts can express changes in object shape (e.g. sitting down) and/or trajectory (e.g. zigzagging). The challenge is to describe this knowledge and the application objectives in a declarative way and to link them to a description of the observed real scene. This topic will be partially studied by PhD G. Pusiol who is working on learning primitive event concepts related to trajectories [web-site, conference, 57, 66]. A second goal is to learn the **semantics of the observed dynamic 3D scene** based on, for instance the statistic analysis of features such as object trajectories. For that, in the ICT COFRIEND and VANAHEIM projects we are planning to learn people activities on airport aprons and subways; we will infer the scene topology (e.g. the main aircraft access points) and we will cluster people trajectories into meaningful categories (e.g. the main routes) through the massive and long-term recording of perceptual data coming from a network of video cameras and GPS sensors.
In a complementary way, **the symbolic approaches** are well adapted to model complex temporal events involving multiple physical objects. The main problem of these approaches is the mechanism to handle the vision algorithm limitations. Most of the time, recognition algorithms suppose that vision algorithms generate perfect tracked objects. Therefore, I am working on two main improvements. First, **managing the uncertainty** of vision features (in particular the lost of tracked objects), is a crucial point. To reach this objective, we are extending the Scenario Description Language for modelling explicitly the uncertainty of vision features. Together with this language extension, I would like to propose mechanisms to **propagate** this uncertainty through all the layers of the event recognition process. These improvements will be explored through the PhD R. Romdhame [web-site, conference, 72]. The second improvement consists in **learning the scenario models** of interest. This issue becomes essential while dealing with video monitoring applications and with a large amount of scenario models. This research direction is related to knowledge acquisition and is detailed below.

After studying the issues and perspectives of the scene understanding process in these perceptual world, physical world and semantic world, I am addressing problems more related to **scene understanding systems**. Once we have shown that a computer program can understand a scene in few situations, it is still remaining to study how this processing can be **real time** and be generalized in various real cases. I believe that new trends in scene understanding rely on this generalising process, on the mechanisms to acquire and **capitalise knowledge** and on making systems **adaptable**, user-centred and in the same time fully autonomous. I am addressing these topics through two directions: (1) performance evaluation and learning system knowledge and (2) knowledge acquisition through end-user interactions.

Concerning **Evaluation and Learning***,* I have proposed several contributions in performance evaluation on video understanding algorithms and in learning techniques for parameter tuning. On evaluation, I have designed a methodology and a tool for evaluating the performance of video understanding systems. On learning system knowledge, I have designed with B. Georis an algorithm to learn automatically the parameters of a segmentation program, by clustering the illumination distributions of a given scene. This is the first stage towards the dynamic configuration of video understanding systems. These evaluation and

learning mechanisms are at a preliminary stage, but are necessary to obtain an effective video understanding system, operational 24/7 at a large scale. Therefore, more efforts still need to be done. In particular, an appropriate formalism needs to be defined to **integrate seamlessly** these mechanisms into new scene understanding systems. Moreover, for **parameter tuning,** I aim at characterising precisely and exhaustively all algorithm parameters, their dependencies between themselves and between a characterisation of the input data (i.e. videos) and their impact on system performance. Given this characterisation and a set of reference videos associated with ground-truth, we should be able to automatically and dynamically configure any scene understanding system. This is the subject of PhD A. T. Nghiem [web-site, conference, 44, 56, 71].

Concerning **knowledge acquisition** through end-user interactions, I have designed two tools for acquiring a priori knowledge and the scenarios to be recognised. The first tool, by simulating **3D animations** helps end-users to define and to visualize their scenarios of interest. It has been successfully used to model several scenarios in two applications. However, the tool is not user-friendly and mature enough to be fully operational and to simulate the whole diversity of the real world. Research in this domain is still an appealing topic. For instance, I am planning to work on a generic and complete formalism to describe realistic dynamic 3D virtual scenes to be visualised by end-users. These 3D scenes could be used as videos associated with ground-truth for assessing the performance of the resulting system. The second tool aims at **learning the frequent combinations of primitive events** called event patterns. These event patterns correspond to the frequent activities occurring in the observed scene and are for end-users potential scenarios of interest. This tool is useful for pre-defining everyday activities, especially in monitoring applications [web-site, conference, 36]. However, the learnt frequent scenarios are not always interesting and scenarios of interest are not necessary frequent. Thus, I am planning to refine this approach to include, for instance, contextual information and **end-users feedback** through an interactive interface to guide the extraction of scenarios of interest. In the same way, I would like to conceive tools to visualise and **explore the event space** structured through the computed event patterns. For me another trend consists in **optimizing the processing time** of the clustering techniques by taking benefit of new data mining algorithms. This topic has started with PhD G. Pusiol [web-site, conference, 57, 66, 67] and will expand through the ICT VANAHEIM project.

All these topics on evaluation, learning and on knowledge acquisition for video understanding systems are sensitive and will become critical soon in order to develop 24/7 resilient video understanding systems working on a large scale and fully adaptable to complex dynamic environment. However, these studies are still at a preliminary stage of development and waiting for the main scene understanding process to be sufficiently mastered. Therefore, more work is expected to flourish in these learning and knowledge acquisition domains. This work is ambitious but constitutes the necessary step to design a new paradigm for the easy building of robust scene understanding systems.