



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

*Project-Team PRIMA*

*Perception, recognition and integration for  
interactive environments*

*Rhône-Alpes*

THEME 3A

*Activity*  
*R* *eport*

2004



## Table of contents

|  |           |
|--|-----------|
| <b>1. Team</b>   | <b>1</b>  |
| <b>2. Overall Objectives</b>   | <b>1</b>  |
| 2.1. Perception, Recognition and Integration for Interactive Environments.                 | 1         |
| <b>3. Scientific Foundations</b>   | <b>2</b>  |
| 3.1. Context Aware Observation of Activity   | 2         |
| 3.2. A Process Architecture for Observation of Human Activity                              | 3         |
| 3.2.1. Modules and Processes   | 3         |
| 3.2.2. Reflexive Process Supervision   | 4         |
| 3.2.3. Tracking Processes  | 5         |
| 3.2.4. Process Federations   | 6         |
| 3.3. Describing and Matching Local Appearance  | 6         |
| 3.4. Generic Features for Robust Tracking and Recognition                                  | 8         |
| <b>4. Application Domains</b>  | <b>8</b>  |
| 4.1. The Augmented Meeting Environment   | 8         |
| 4.2. The Steerable Camera Projector  | 10        |
| 4.3. Context Aware Video Acquisition   | 12        |
| <b>5. Software</b>   | <b>13</b> |
| 5.1. IMALAB  | 13        |
| 5.2. BrandDetect   | 17        |
| 5.3. CAR: Robust Real-Time Detection and Tracking  | 17        |
| <b>6. New Results</b>  | <b>18</b> |
| 6.1. A Programmable Robust Tracker   | 18        |
| 6.2. Audio Processes for Detection and Tracking  | 22        |
| 6.3. The Process Federation Tool   | 22        |
| 6.3.1. An Example: Distributed Camera Net  | 23        |
| 6.4. Specifying a context model  | 24        |
| 6.5. Context model compiler  | 24        |
| 6.5.1. Situation graphs and temporal relations   | 25        |
| 6.5.2. Synchronized Petri Nets   | 25        |
| 6.5.3. Jess rule generation  | 25        |
| 6.6. Generic Features for detection and classification                                     | 26        |
| 6.6.1. Computation of class-specific feature detectors                                     | 26        |
| 6.6.2. Modeling spatial relations  | 26        |
| <b>7. Contracts and Grants with Industry</b>   | <b>28</b> |
| 7.1. European and National Projects  | 28        |
| 7.1.1. IST-2000-28323 FAME: Facilitating Agent for Multi-Cultural Exchange                 | 28        |
| 7.1.2. IST-2001-32157 DETECT: Real Time Detection of Motion Picture Content in Live Broad- | 29        |
| casts  | 29        |
| 7.1.2.1. Detection of semantic blocks.   | 30        |
| 7.1.2.2. Detection of ROI in order to identify static and dynamic objects.                 | 30        |
| 7.1.2.3. Motion picture analysis.  | 30        |
| 7.1.3. IST 2001 37540 CAVIAR: Context Aware Vision using Image-based Active Recognition    | 30        |
| 7.1.4. IST 506909 CHIL: Computers in the Human Interaction Loop                            | 31        |
| 7.1.5. RNTL/Proact: ContAct Context management for pro-Active computing                    | 32        |
| <b>8. Other Grants and Activities</b>  | <b>33</b> |
| 8.1. European Research Networks  | 33        |

|            |   |           |
|------------|---|-----------|
| 8.1.1.     | IST-2001-35454 ECVision: European Research Network for Cognitive AI-enabled Computer Vision Systems | 33        |
| 8.1.2.     | IST-2000-26434 FGnet: Face and Gesture Recognition Working Group                                    | 34        |
| <b>9.</b>  | <b>Dissemination</b>  | <b>34</b> |
| 9.1.       | Contribution to the Scientific Community  | 34        |
| 9.1.1.     | EUSAI '04: European Symposium on Ambient Intelligence   | 34        |
| 9.1.2.     | PETS '04: Performance Evaluation for Tracking and Surveillance                                      | 34        |
| 9.1.3.     | Pointing '04: International Workshop on Diectic Gestures  | 35        |
| 9.1.4.     | RFIA '04: Reconnaissance de Forme et Intelligence Artificielle                                      | 36        |
| 9.1.5.     | Participation on Conference Program Committees  | 36        |
| 9.1.6.     | Participation on Advisory Panels  | 36        |
| 9.1.7.     | Invited Plenary Presentations at Conferences  | 36        |
| <b>10.</b> | <b>Bibliography</b>   | <b>37</b> |

# 1. Team

## Head of the team

James L. Crowley [Professor INPG]

## Professors

Augustin Lux [Professor INPG]

Patrick Reignier [Assistant professor UJF]

## Team assistant

Natacha Laugier

## Expert Engineers

Daniela Hall

Alban Caporossi

Alba Ferrer-Biosca

Sebastien Pesnel

Jean-Marie Vallet

## Post-doctoral Researchers

Dominique Vaufreydaz

## Doctoral Researchers

Stanislas Borkowski [Bourse EGIDE]

Stephane Guy [Bourse INRIA]

Suphot Chunwiphat [Bourse gouvernement thailandais]

Matthieu Anne [Bourse CIFRE - France Telecom]

Thi-Thanh-Hai Tran [Bourse EGIDE]

Olivier Bertrand [Solde de Normalien]

Nicolas Gourier [Bourse INRIA]

Julien Letessier [Bourse INRIA]

Jerome Maisonnasse [INPG SA - Contrat France Telecom]

Oliver Brdiczka [Bourse INRIA]

# 2. Overall Objectives

## 2.1. Perception, Recognition and Integration for Interactive Environments.

**Keywords:** *Computer Vision, Interactive Environments, Machine Perception, Man-Machine Interaction, Perceptual User Interfaces.*

The objective of Project PRIMA is to develop a scientific and technological foundation for interactive environments. An environment is said to be "interactive" when it is capable of perceiving, acting, and communicating with its occupants. The construction of such environments offers a rich set of problems related to interpretation of sensor information, learning, machine understanding and man-machine interaction. Our goal is make progress on a theoretical foundation for cognitive or "aware" systems by using interactive environments as a source of example problems, as well as to develop new forms of man machine interaction.

An environment is a connected volume of space. An environment is said to be "perceptive" when it is capable of recognizing and describing things, people and activities within its volume. Simple forms of applications-specific perception may be constructed using a single sensor. However, to be general purpose and robust, perception must integrate information from multiple sensors and multiple modalities. Project PRIMA develops and employs machine perception techniques using acoustics, speech, computer vision and mechanical sensors.

An environment is said to be "active" when it is capable of changing its internal state. Trivial forms of state change include regulating ambient temperature and illumination. Automatic presentation of information and

communication constitutes a challenging new form of "action" with many applications. The use of multiple display surfaces coupled with location awareness of occupants offers the possibility of automatically adapting presentation to fit the current activity of groups. The use of activity recognition and acoustic topic spotting offers the possibility to provide relevant information without disruption. The use of steerable video projectors (with integrated visual sensing) offers the possibilities of using any surface as for presentation and interaction with information.

An environment may be considered as "interactive" when it is capable responding to humans using tightly coupled perception and action.

Simple forms of interaction may be based on sensing grasping and manipulation of sensor-enabled devices, or on visual sensing of fingers or objects placed into projected interaction widgets. Richer forms of interaction require perceiving and modeling of the current task of users. PRIMA explores multiple forms of interaction, including projected interaction widgets, observation of manipulation of objects, fusion of acoustic and visual information, and federations of systems that model interaction context in order to predict appropriate action by the environment.

For the design and integration of systems for perception of humans and their actions, PRIMA has developed:

- A new approach to computer vision based on local appearance,
- A software architecture model for reactive control of multi-modal vision systems.
- A conceptual framework and theoretical foundation for context aware perception.

The experiments in project PRIMA are oriented towards perception of human activity. The project is particularly concerned with modeling the interaction between communicating individuals in order to provide video-conferencing and information services. Application domains include context aware video communications, new forms of man-machine interaction, visual surveillance, and new forms of information services and entertainment.

## 3. Scientific Foundations

### 3.1. Context Aware Observation of Activity

**Keywords:** *Context Aware Systems, Context Modeling, Observation of Human Activity.*

Human activity is extremely complex. Current technology allows us to handcraft real-time perception systems for a specific perceptual task. However, such an approach is inadequate for building systems that accommodate the variety of activities that is typical of human environments. To respond to this need, we have defined a conceptual framework for context aware observation of human activity. This framework and methods are used to construct systems for observation of human activity in the PRIMA "Augmented Meeting Environment". Within this framework, contexts are modeled as a network of situations. Situation networks are interpreted as a specification of a federation of processes for observing the entities and relations that define a situation. In this section we review conceptual foundations for such systems. In the following section we describe a process-based software architecture for building systems for observing activity based on this framework.

In the models of activity developed in project PRIMA, human activity is represented as a network of situations. A situation is defined as a configuration of relations computed over observed entities. Relations are predicate functions evaluated over the properties of one or more entities. Changes in relations trigger events that signal a change in situation. The entities and relations that define situations are detected and observed by perceptual processes. A federation of processes is composed and coordinated by a federation controller (a "Federator") in order to predict and observe the situations that describe an activity, and to perform the appropriate actions.

The concept of role is an important (but subtle) tool for simplifying the network of situations. It is common to discover a collection of situations that have the same configuration of relations, but where the identity of one or more entities is varied. A role is an abstract class of person or object that combines a discriminative recognition test with symbolic description of functionality. Within a situation model, a role serves as a variable for the entities to which the relations are applied, thus allowing an equivalent set of situations to have the same representation. A role is played by an entity that can pass an acceptance test for the role.

In that case, it is said that the entity can play or adopt the role for that situation. In our framework, the relations that define a situation are defined with respect to roles, and applied to entities that pass the test for the relevant roles.

Entities are assigned to roles by role assignment processes. A change in the assignment of an entity to a role does not change the situation, unless the result is a change in a relations that define the situation. The context model specifies which roles are to be assigned and launches the necessary role assignment processes. A process federation is formed to detect and observe entities that can play roles as well as the relation between entities. We have found that this architecture provides a foundation for the design of systems that act as a silent partner to assist humans in their activities in order to provide appropriate services without explicit commands and configuration.

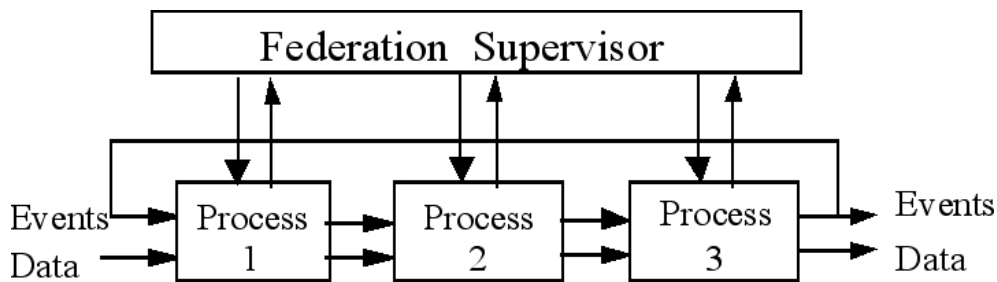


Figure 1. A process supervisor launches and configures a network of perceptual processes

## 3.2. A Process Architecture for Observation of Human Activity

**Keywords:** *Autonomic Systems, Computer Vision Systems, Process Architectures, Reflexive Systems.*

The PRIMA project has developed a data-flow architecture based on dynamically assembled federations [28][42]. This model builds on previous work on process-based architectures for machine perception and computer vision [40][57], as well as on data flow models for software architecture [59]. Processes are launched and configured to observe the entities and relations that define situations. This approach provides an architecture in which reflexive elements are dynamically composed to form federations of processes for observing and predicting the situations that make up a context.

As context changes, the federation is restructured. Restructuring the federation enables the system to adapt to a range of environmental conditions and to provide services that are appropriate over a range of activities.

### 3.2.1. Modules and Processes

Perceptual processes are composed from a collection of modules controlled by a process supervisor as shown in figure 2. Processes operate in a synchronous manner within a shared address space. Such models are related to "work-flow" models increasingly used in modeling human organizations. Process models have been adapted for real time computer vision systems in the ESPRIT BRA project "Vision as Process" [27][40]. Such models permit the dynamic composition of software "federations" in response to events in the scene [42][39].

In our experimental system, the process supervisor is implemented as a multi-language interpreter [49] equipped with a dynamic loader for precompiled libraries. This interpreter allows a processes to receive and interpret messages containing scripts, to add new functions to a process during execution. Inter-process communication is provided by a software bus based on the JORAM middleware from ObjectWeb [30]. This allows us to compose federations of processes distributed on multiple computers.

The modules that compose a process are formally defined as transformations applied to a certain class of data or event. Modules are executed in cyclic manner by the supervisor according to a process schedule. We impose that transformations return an auto-critical report that describes the results of their execution. Examples of information contained in an auto-critical report include elapsed execution time, confidence in the result, and any exceptions that were encountered. The auto-critical report enables a supervisory controller to adapt parameters for the next call in order to maintain a execution cycle time, or other quality of service.

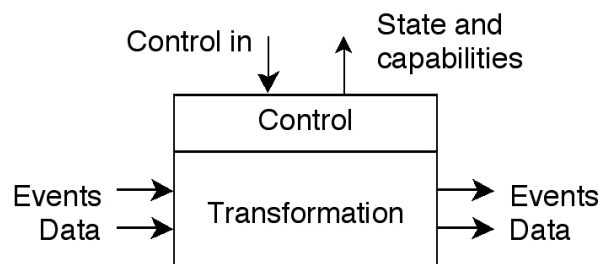


Figure 2. A perceptual process is composed of a set of modules controlled by a supervisor. Processes transform data streams as well as generate and respond to events.

### 3.2.2. Reflexive Process Supervision

The supervisory component of a process provides four fundamental functions: command interpretation, execution scheduling, parameter regulation, and reflexive description. The supervisor acts as a programmable interpreter, receiving snippets of code script that determine the composition and nature of the process execution cycle and the manner in which the process reacts to events. The supervisor acts as a scheduler, invoking execution of modules in a synchronous manner. The supervisor regulates module parameters based on the execution results. Auto-critical reports from modules permit the supervisor to dynamically adapt processing. Finally, the supervisor responds to external queries with a description of the current state and capabilities. We formalize these abilities as the autonomic properties of auto-regulation, auto-description and auto-criticism.

A system requires information about the capabilities and the current state of component processes in order to dynamically assemble and control observational processes. Such information can be provided by assuring that supervisory controllers have the reflexive capabilities of auto-regulation, auto-description and auto-criticism.

A process is auto-regulated when processing is monitored and controlled so as to maintain a certain quality of service. For example, processing time and precision are two important state variables for a tracking process. These two may be traded off against each other.

The process controllers may be instructed to give priority to either the processing rate or precision. The choice of priority is dictated by a more abstract supervisory controller.

An auto-descriptive controller can provide a symbolic description of its capabilities and state. The description of the capabilities includes both the basic command set of the controller and a set of services that the controller may provide to a more abstract controller. Thus when applied to the system's context, our model provides a means for the dynamic composition of federations of controllers. In this view, the observational processes may be seen as entities in the system context. The current state of a process provides its observational



variable. Supervisory controllers are formed into hierarchical federations according to the system context. A controller may be informed of the possible roles that it may play using a meta-language, such as XML.

An auto-critical process maintains an estimate of the confidence for its outputs. For example, the skin-blob detection process maintains a confidence factor based on the ratio of the sum of probabilities to the number of pixels in the ROI. Such a confidence factor is an important feature for the control of processing. Associating a confidence factor to all observations allows a higher-level controller to detect and adapt to changing observational circumstances. When supervisor controllers are programmed to offer "services" to higher-level controllers, it can be very useful to include an estimate of the confidence for the role. A higher-level controller can compare these responses from several processes and determine the assignment of roles to processes.

### 3.2.3. Tracking Processes

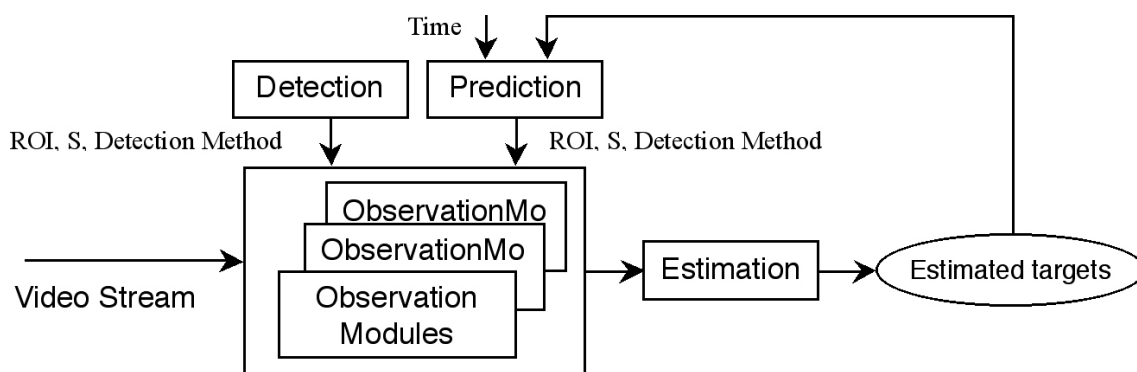


Figure 3. Tracking is a cyclic process of four phases: Predict, Observe, Detect and Estimate. Observation is provided by the observation and grouping modules described above.

Tracking provides a number of fundamentally important functions for a perception system. Tracking aids interpretation by integrating information over time. Tracking makes it possible to conserve information, assuring that a label applied to an entity remains associated with the entity at future times. Tracking provides a means to focus attention, by predicting the region or interest and the observation module that should be applied to a specific region of an image. Tracking processes can be designed to provide information about position speed and acceleration that can be useful in describing situations.

Tracking is a cyclic process of recursive estimation applied to a data stream. In perception systems, a tracking process is generally composed of three phases: predict, observe and estimate, as illustrated in figure 3. Tracking maintains a list of entities, known as "targets". Each target is described by a unique ID, a target type, a confidence (or probability of existence), a vector of properties and a matrix of uncertainties (or precisions) for the properties.

The prediction phase uses a temporal model (called a "process model" in the tracking literature) to predict the properties that should be observed at a specified time for each target. For many applications of tracking, a simply linear model is adequate for such prediction. A linear model maintains estimates of the temporal derivatives for each target property and uses these to predict the observed property values.

The prediction phase also updates the uncertainty (or precision model) of properties. Uncertainty is generally represented as a covariance matrix for errors between estimated and observed properties. These uncertainties are assumed to arise from imperfections in the process model as well as errors in the observation process. Restricting processing to a region of interest (ROI) can greatly reduce the computational load for image analysis. The predicted position of a target determines the position of the ROI at which the target

should be found. The predicted size of the target, combined with the uncertainties of the size and position, can be used to estimate the appropriate size for the ROI. In the tracking literature, this ROI is part of the "validation gate", and is used to determine the acceptable values for properties.

Observation is provided by the observation and grouping modules described above. Processing is specific for each target. A call to a module applies a specified observation procedure for a target at a specified ROI in order to verify the presence of the target and to update its properties. When the detection confidence is large, grouping the resulting pixels provides the information to update the target properties.

The estimation process combines (or fuses) the observed properties with the previously estimated properties for each target. If the average detection confidence is low, the confidence in the existence of a target is reduced, and the predicted values are taken as the estimates for the next cycle. If the confidence of existence falls below a threshold, the target is removed from the target list.

The detection phase is used to trigger creation of new targets. In this phase, specified observation modules are executed within a specified list of "trigger" regions. Trigger regions can be specified dynamically, or recalled from a specified list. Target detection is inhibited whenever a target has been predicted to be present within a trigger region.

### 3.2.4. Process Federations

Perceptual processes may be organized into software federations [28][42]. A federation is a collection of independent processes that cooperate to perform a task. A process federation is assembled as a network of processes controlled by a federation supervisor. Federation supervisors invoke and configure processes to perform the transformations required to observe a context. The states of processes are monitored by the supervisory controller and process parameters are adapted in response to events. Supervisory controllers may be assembled into hierarchies in order to observe human activity. The exact assembly depends on the task that the system is to perform as described by a model of the user's task and context.

A crucial problem with this model is how to provide a mechanism for dynamically composing federations of supervisory controllers that observe the entities and relations relative to the user's context. Our approach is to propose a reflexive federation supervisor. A federation supervisor is designed for a specific context model. The federation supervisor maintains a model of the situation.

The federation supervisor can be seen as a form of reactive expert system. For each user context, it invokes and revokes the corresponding highest-level supervisory controllers. These controllers, in turn, invoke and revoke lower level controllers, down to the level of the lowest level observational processes. Supervisory controllers may evoke competing lower-level processes, informing each process of the roles that it may play. The selection of process for a role can then be re-assigned dynamically according to the quality of service estimate that each process provides for its parent controller.

## 3.3. Describing and Matching Local Appearance

**Keywords:** *Appearance Based Vision, Computer Vision, Object Recognition, Receptive Fields, Scale Invariance.*

The appearance of something is the set of possible visual stimuli that the thing may engender. For computer vision, the appearance function for an entity refers to the set of all possible images that may be observed for that entity. Appearance functions can be created for objects, activities, and scenes. Contrary to intuition, it is possible to capture a computer model for appearance functions. Such models can be used to provide efficient processes for detecting, tracking, and observing people and things in real world environments.

The members of project PRIMA have a long history in defining and developing methods for appearance based computer vision. Recently the project has demonstrated a variety of new computer vision methods based on the use of chromatic receptive fields.

Adopting the terminology of Schiele [58], a "receptive field" is a local function defined over some domain of image space. The term "receptive field", drawn from psychophysics, refers to the weighting functions used to encode visual stimuli in biological visual systems.

Although it is not our intention to propose a model for biological vision, we note in passing that Young [66] and others has used Gaussian derivatives as models for the simple cells in the early layers of the primate visual cortex [47]. PRIMA has developed a family of receptive field functions based on evaluating scale normalized Gaussian derivatives in a color opponent space. These functions provide a foundation for robust real time processes for observing objects and agents.

Receptive fields may be defined over image space, color, time, view-point or any other image formation parameter [37]. A projection of the appearance manifold onto a receptive field at a position in an image provides a scalar value which describes appearance at that position on the manifold. Projection to a vector of receptive fields provides a vector of features. Such a vector, provides a concise description of the appearance within the local neighborhood of the image.

Projection onto a receptive field is provided by an inner product.

Computing inner products at each image position is equivalent to a convolution of the receptive field with the image. When evaluated at every image point in an image, a receptive field is a form of linear filter. Gaussian derivative filters can be computed very rapidly using a variety of techniques including separable recursive filters [68] and Gaussian Pyramids [41].

The Hermite polynomial [31] of Gaussian derivatives has a variety of properties which make it ideally suited as a basis for image description using a Taylor series. Even ordered derivatives will respond to symmetric structures such as spots and bars while odd derivatives will respond to asymmetric structures, such as edges. In two dimensions, the Gaussian is the unique function which is both separable and circularly symmetric. Oriented derivatives can be defined as a convolution of separable components.

Changing viewing distance changes the scale of appearance. View invariant recognition requires estimating and normalizing such changes. The scale equivariant properties of Gaussian derivatives provide a simple method to estimate changes in scale. Such normalization provides both robustness to changes in distance, and adaptation of the receptive field to the most appropriate scale for describing the appearance at each image position.

Local normalization of the receptive fields requires a local estimate of the intrinsic scale and orientation. Intrinsic scale is determined by local maxima in the Laplacian with respect to change in scale. Intrinsic orientation is determined by the direction of the local gradient. With such normalization, the Laplacian of the Gaussian computed at a range of scales at each point provides a "Laplacian Profile" which is equivariant with scale [38].

Whenever the gradient is not close to zero, it is possible to estimate an intrinsic orientation using the arctangent of the ratio of first derivatives. The Gaussian derivatives vector provides a basis for synthesizing oriented derivatives [44]. This can be shown by expressing the 1st, 2nd and 3rd order Gaussian derivatives in polar coordinates and determining their Fourier transforms. Rotation is equivalent to a shift in phase. Thus oriented Gaussian derivatives can be synthesized at arbitrary angles using weighted sum of the Gaussian derivatives computed in the cardinal directions. The weights are provided by the direction cosines of the rotation.

Color is a very powerful discriminant for object recognition. Color images are commonly acquired in the Cartesian color space, RGB. The RGB color space has certain advantages for image acquisition, but is not the most appropriate space for recognizing objects or describing their shape. The HLS (Hue, Luminance, Saturation) color space is a commonly used representation which separates intensity and chrominance. In HLS, chrominance is represented by a polar coordinate representation in which hue is the angle and saturation is the radius. Projection of RGB onto a polar coordinate representation causes computational problems in the design of receptive fields. An alternative is to compute a Cartesian representation for chrominance, using differences of R, G and B. Such differences yield color opponent receptive fields resembling those found in biological visual systems.

The intensity component may be obtained by a weighted sum of the R, G, and, B. The exact weights depend on the camera and scene illumination and may be adapted to the RGB filters and the source illumination, by gain coefficients. The chromatic component may be obtained by differences. Such a separation may easily be performed by multiplying the color vector by a matrix [52].

The components C1 and C2 encodes the chromatic information in a Cartesian representation. Chromatic Gaussian receptive fields are computed by applying the Gaussian derivatives independently to each of the three components, (L, C1, C2). The result is a set of color opponent filters as shown in figure 4. Permutations of RGB lead to different opponent color spaces. The choice of the most appropriate space depends on the chromatic composition of the scene.

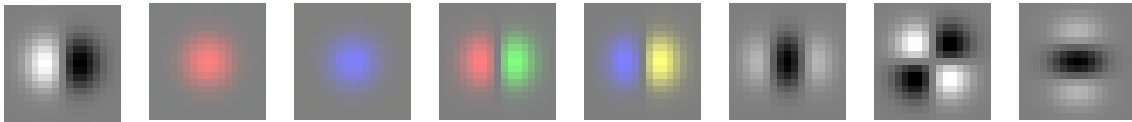


Figure 4. Chromatic Gaussian Receptive Fields ( $G_x^L, G^{C_1}, G^{C_2}, G_x^{C_1}, G_x^{C_2}, G_{xx}^L, G_{xy}^L, G_{yy}^L$ ).

### 3.4. Generic Features for Robust Tracking and Recognition

**Keywords:** *Computer Vision, Generic Features, Robust Matching.*

A successful detection, tracking and classification system must have two properties: it must be general enough to correctly assign instances of the same class despite large intra-class variability and it must be specific enough to reject instances that are not part of the class. Features robust to intra-class variability can be constructed by learning from examples. The result is a feature or part detector that can generalize from a small number of examples to new examples. Such a detector can provide a hypothesis about the presence of a class instance, but it is in general not specific enough for reliable detection and classification.

The relative position of distinct object features is important for classification and needs to be modeled. Current approaches [43][32] are computationally expensive. In the result section, we propose an efficient method for geometry verification. This enables a object detection and identification module which can be applied to various object types.

## 4. Application Domains

### 4.1. The Augmented Meeting Environment

**Keywords:** *Augmented Reality, Collaborative Work, Multi-modal Interaction.*

**Participants:** Patrick Reignier, Dominique Vaufreydaz, Christophe Le Gal, James L. Crowley.

In order to test and develop systems for observation of human activity, Project PRIMA has constructed an "Augmented Meeting Environment", show in figure 5. The PRIMA Augmented Meeting Environment is equipped with a microphone array, a fixed wide angle camera, five steerable cameras, three "video interaction devices". The microphone array is used as an acoustic sensor to detect, locate and classify acoustic signals for recognizing human activities. The wide-angle camera provides a field of view that covers the entire room, and allows detection and tracking of individuals. Steerable cameras are installed in each of the four corners of the room, and used to acquire video of activities from any viewing direction.

Video interaction devices associate a camera with a video projector to provide new modes of man-machine interaction. Such devices may be used for interaction, presentation or capture of information based on natural activity. Examples include selecting menus and buttons with a finger and capturing drawings from paper or a whiteboard. Fixed video interaction devices in the AME have been constructed for a vertical surface (a wall mounted white board) and a horizontal desk-top work-space. Recently a steerable interaction device has been constructed based on a tightly integrated steerable camera-projector pair (SCP). The SCP described below, allows any surface to be used for interaction with information. It also offers a range new sensing techniques,



*Figure 5. The augmented meeting environment is an office environment equipped with a microphone array, wireless lapel microphones, a wide angle surveillance camera, five steerable cameras, and three video-interaction devices.*

including automatic surveillance of an environment to discover the environment topology, as well as the use of structured light for direct sensing of texture mapped 3D models.

## 4.2. The Steerable Camera Projector

**Keywords:** *Interactive Environments, Man-Machine Interaction.*

**Participants:** Stan Borkowski, James L. Crowley, Olivier Riff.

Surfaces dominate the physical world. Every object is confined in space by its surface. Surfaces are pervasive and play a predominant role in human perception of the environment. We believe that augmenting surfaces with information technology will provide an interaction modality that will be easily adopted by humans.

PRIMA has constructed a steerable video interaction device composed of a tightly coupled camera and video projector. This device, known as a Steerable Camera-Projector (or SCP) enables experiments in which any surface in the augmented meeting environment may be used as an interactive display for information [35]. With such a device, an interaction interface may follow a user, automatically selecting the most appropriate surface. The SCP provides a range of capabilities (a) The SCP can be used as a sensor to discover the geometry of the environment, (b) The SCP can project interactive surfaces anywhere in the environment and (c) The SCP can be used to augment a mobile surface into a portable interactive display. (d) The SCP can be used to capture text and drawings from ordinary paper. (e) The SCP can be used as a structured light sensor to observe 3-D texture-mapped models of objects.

Current display technologies are based on planar surfaces. Recent work on augmented reality systems has assumed simultaneous use of multiple display surfaces [46][60][64]. Displays are usually treated as access points to a common information space, where users can manipulate vast amounts of information with a set of common controls. With the development of low-cost display technologies, the available interaction surface will continue to grow, and interfaces will migrate from a single, centralized screen to multiple, space-distributed interactive surfaces. New interaction tools that accommodate multiple distributed interaction surfaces will be required.

Video-projectors are increasingly used in augmented environment systems [56][61]. Projecting images is a simple way of augmenting everyday objects and offers the possibility to change their appearance or their function. However, standard video-projectors have a fairly small projection area which significantly limits their spatial flexibility as output devices in a pervasive system. A certain degree of steerability can be achieved for a rigidly mounted projector: In particular, a sub window can be steered within the cone of projection for a fixed projector [63]. However, extending and/or moving the display surface requires augmenting the range of angles to which the projector beam may be directed. If using fixed projectors, this means increasing the number of projectors which is relatively expensive. A natural solution is to use a Steerable projector-camera assembly [51] and [54]. With a trend towards increasingly small and inexpensive video projectors and cameras, this approach will become increasingly attractive. Additionally having the ability to modify the scene with projected light, projector-camera systems can be exploited as sensors, thus enabling to collect data that can be used to build a model of the environment.

Projection is an ecological (i.e. non-intrusive) way of augmenting the environment. Projection does not change the augmented object itself, only its appearance. This change can be used to supplement the functionality of the object and henceforth its role in the world. However, the most common consequence of augmenting an object with projected images is transforming the object into an access point to the virtual information space. In [54] ordinary artifacts such as walls, shelves, and cups are transformed into informative surfaces. Though the superimposed projected image enables the user to take advantage of the information provided by the virtual world, the functionality of the object itself does not change. The object becomes a physical support for virtual functionalities. An example of enhancing the functionality of an object was presented in [36], where users could interact with both physical and virtual ink on an projection-augmented whiteboard.

The Steerable Camera Projector (SCP) (figure 6) platform is a device that provides a video-projector with two mechanical degrees of freedom: pan and tilt. The mechanical performance of the SCP is presented in



Figure 6. The Steerable Camera Projector

Table 1. Rotation platform mechanical performance

|                    | Pan                 | Tilt               |
|--------------------|---------------------|--------------------|
| Rotation range     | $\pm 177^\circ$     | $+90^\circ$        |
| Angular resolution | $0.11^\circ$        | $0.18^\circ$       |
| Angular velocity   | $146 \frac{deg}{s}$ | $80 \frac{deg}{s}$ |
| Response time      | $\sim 2ms$          | $\sim 3ms$         |

Table 1. While somewhat bulky, our device anticipates the current trend of projectors to become portable devices, similar in shape to hand-held torch lamps [55].

Note that the SCP is not only a motorized video-projector, but a projector-camera pair. The camera is mounted in such a way that the projected beam overlaps with the camera-view. Equipping an SCP with a camera offers a number of interesting possibilities. User's actions can be observed within the field of view of the camera and interpreted as input information for the computer system. Additionally the system is able to provide visual feedback in response to users action. In other words association of a camera to a projector creates a powerful actuator-sensor pair.

The SCP can be used as a steerable structured light sensor to automatically discover surfaces that are suitable for interaction. Figure 7 shows automatically discovered planar surfaces within the AME. described below.



Figure 7. Planar surfaces in the environment

### 4.3. Context Aware Video Acquisition

**Keywords:** *Context Aware Systems, Intelligent Environments, Video Conferencing.*

**Participants:** Patrick Reignier, Dominique Vaufreydaz, Olivier Riff, Alban Caporossi, Alba Ferrer-Biosca, James L. Crowley.

Video communication has long been seen as a potentially powerful tool for communications, teaching and collaborative work. Continued exponential decreases in the cost of communication and computation (for coding and compression) have eliminated the cost of bandwidth as an economic barrier for such technology. However, there is more to video communication than acquiring and transmitting an image. Video communications technology is generally found to be disruptive to the underlying task, and thus unusable. To avoid disruption, the video stream must be composed of the most appropriate targets, placed at an appropriate size and position in the image. Inappropriately composed video communications create distraction and ultimately degrades the ability to communicate and collaborate.

During a lecture or a collaborative work activity, the most appropriate targets, camera angle, and zoom and target position change continually. A human camera operator understands the interactions that are being filmed and adapts the camera angle and image composition accordingly. However, such human expertise is costly. The lack of an automatic video composition and camera control technology is the current fundamental obstacle to the widespread use of video communications for communication, teaching and collaborative work. One of the goals of project PRIMA is to create a technology that overcomes this obstacle.

To provide a useful service for a communications, teaching and collaborative work, a video composition system must adapt the video composition to events in the scene. In common terms, we say that the system must be "aware of context". Computationally, such a technology requires that the video composition be determined



by a model of the activity that is being observed. As a first approach, we propose to hand-craft such models as finite networks of states, where each state corresponds to a situation in the scene to be filmed and specifies a camera placement, camera target, image placement and zoom.

A finite state approach is feasible in cases where human behavior follows an established stereotypical "script". A lecture or class room presentation provides an example of such a case. Lecturers and audiences share a common stereotype about the context of a lecture. Successful video communications require structuring the actions and interactions of actors to a great extent. We recognize that there will always be some number of unpredictable cases where humans deviate from the script. However, the number of such cases should be sufficiently limited so as to limit the disruption. Ultimately, we plan to investigate automatic techniques for "learning" new situations.

This system described above is based on an approach to context aware systems presented at UBICOMP in September 2002 [39]. The behavior of this system is specified as a situation graph that is automatically compiled into rules for a Java based supervisory process. The design process for compiling a situation graph into a rule based system for the federation supervisors has been developed and refined within the last two years.

In 2004, we have demonstrated a number of real systems based on this model. In the FAME project, we demonstrated a context aware video acquisition system at the Barcelona Forum of Cultures during two weeks in July 2004. This system was also demonstrated publicly at "Fête de la science" in Grenoble in October 2004, and exhibited at the IST Conference in Den Haag in November 2004. A variation of this system has been integrated into the ContAct context aware presentation composition system developed with XRCE (Xerox European Research Centre), and is at the heart of the CHIL Collaborative Workspace Service used in the IP Project CHIL. A context aware interpretation system for video surveillance is currently under development for the IST project CAVIAR.

## 5. Software

### 5.1. IMALAB

**Keywords:** *Computer Vision Systems, Software Development Environments.*

**Participants:** Augustin Lux, Olivier Riff, Alban Caporossi, Daniela Hall.

The Imalab system represents a longstanding effort within the Prima team (1) to capitalize on the work of successive generations of students, (2) to provide a coherent software framework for the development of new research, and (3) to supply a powerful toolbox for sophisticated applications. In its current form, it serves as a development environment for all researchers in the Prima team, and represents a considerable amount of effort (probably largely more than 10 man-years).

There are two major elements of the Imalab system: the PrimaVision library, which is a C++ based class library for the fundamental requirements of research in computer vision; and the Ravi system, which is an extensible system kernel providing an interactive programming language shell.

With respect to other well known computer vision systems, e.g. KHOROS [57] the most prominent features of Imalab are:

- A large choice of data structures and algorithms for the implementation of new algorithms.
- A subset of C++ statements as interaction language.
- Extensibility through dynamic loading.
- A multi language facility including C++, Scheme, Clips, Prolog.

The combination of these facilities is instrumental for achieving efficiency and generality in a large Artificial Intelligence system: efficiency is obtained through the use of C++ coding for all critical pieces of code; this code is seamlessly integrated with declarative programs that strive for generality.

Imalab's system kernel is built on the Ravi system first described in Bruno Zoppis's thesis [67]. The particular strength of this kernel comes from a combination of dynamic loading and automatic program generation within an interactive shell, in order to integrate new code, even new libraries, in a completely automatic way.

The Imalab system has, in particular, been used for the development of the BrandDetect software described below. The Imalab system has proven to be extremely efficient tool for the development of systems such as BrandDetect that extensive performance evaluation as well as incremental design of of a complex user interface.

We currently are in the process of registering of ImaLab with the APP (Agence pour la Protection des Programmes). Imalab has been distributed as share ware to several research laboratories around Europe. Imalab has been installed and is in use at:

- XRCE - Xerox European Research Centre, Meylan France
- JOANNEUM RESEARCH Forschungsgesellschaft mbH, Austria
- HS-ART Digital Service GmbH, Austria
- VIDEOCATION Fernseh-Systeme GmbH, Germany
- Univ. of Edinburgh, Edinburgh, UK
- Instituto Superior Tecnico, Lisbon, Portugal
- Neural Networks Research Centre, Helsinki University of Technology (HUT), Finland
- Jaakko Pöyry Consulting, Helsinki, Finland
- Université de Liège, Belgium
- France Télécom R&D, Meylan France

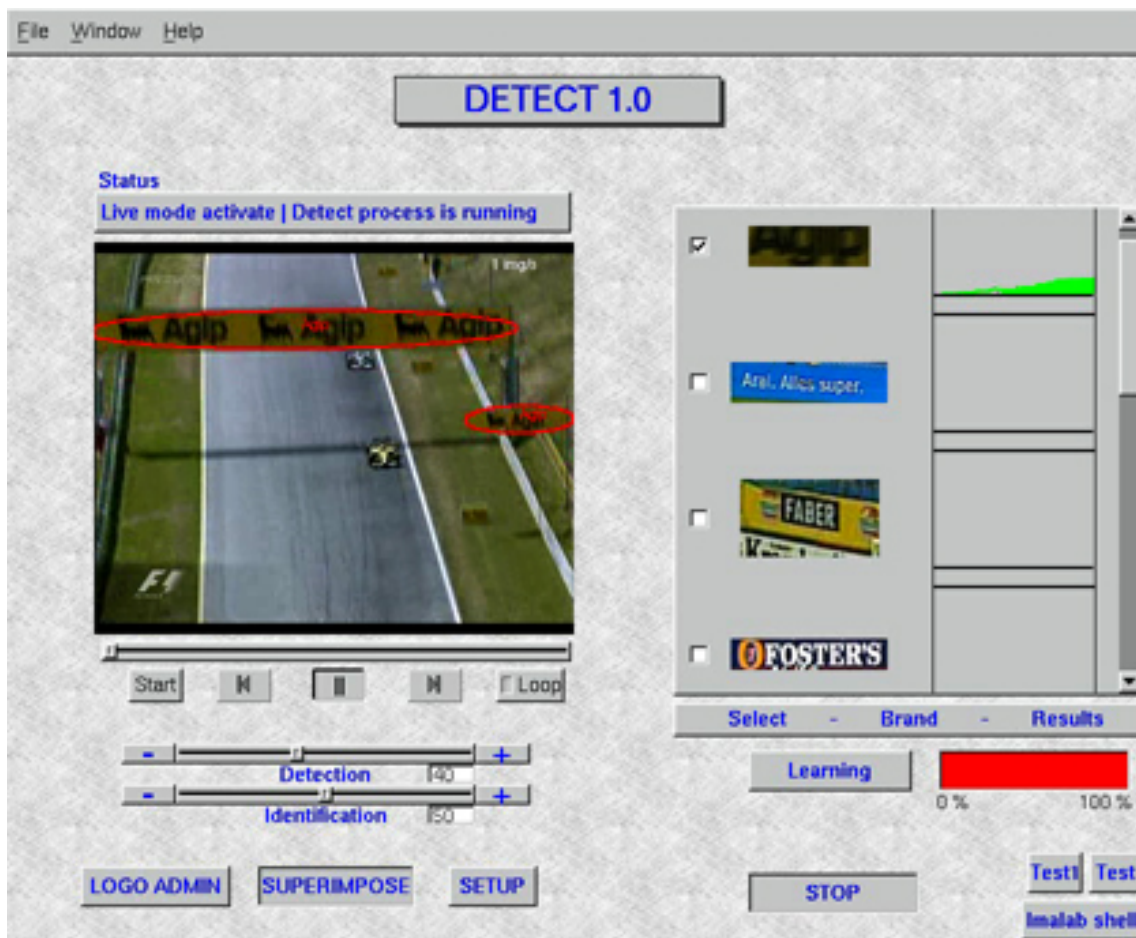


Figure 8. BrandDetect collects statistics on appearance of publicity panels in Broadcast video

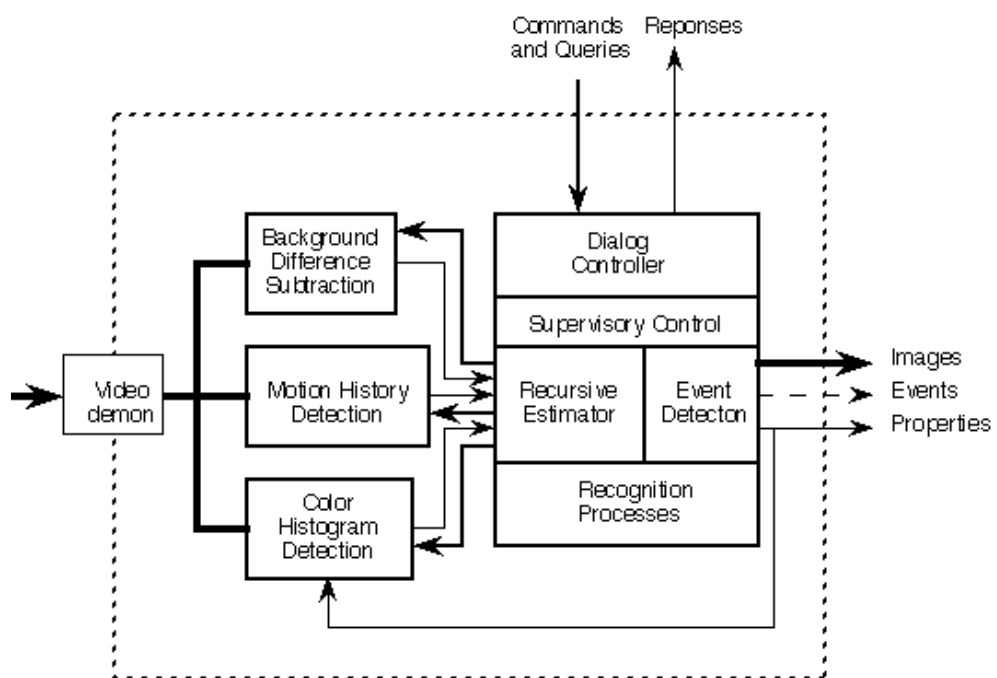


Figure 9. The CAR systems integrates several detection modules with a Kalman Filter for robust detection and tracking of entities

## 5.2. BrandDetect

**Keywords:** *Digital Television, Media Metrics, Video Monitoring.*

**Participants:** Augustin Lux, Olivier Riff, Alban Caporossi, Alba Ferrer-Biosca, James L. Crowley, Daniela Hall.

BrandDetect is a system for detection, tracking and recognition of corporate logos, commercial trademarks and other publicity panels in broadcast television video streams. BrandDetect collects statistics on the frequency of occurrence, size, appearance and duration of presentation of the publicity. It is especially designed for use in the production of broadcast video of sports events such as football matches and formula one racing.

The BrandDetect software can permanently monitor streaming video input from pre-recorded media (MPEG, AVI and other formats) as well as from real time video. BrandDetect looks for occurrences of a predefined set of publicity panels in a manner that is independent of size, rotation and position. Once detected, a publicity panel is tracked in order to collect statistics on duration, size, image quality, and position relative to the center of the screen. These statistics are used to produce an objective report that may be used to establish the potential impact and commercial value of publicity. An example screen image of BrandDetect is shown in figure 8.

BrandDetect has been filed with the l'APP (Agence pour la Protection des Programmes) the 07 Nov 03. (IDDN.FR.450046.000.S.P.2003.000.21000.). A license commercial exploitation has been negotiated with the Austrian company HS Art.

## 5.3. CAR: Robust Real-Time Detection and Tracking

**Keywords:** *Computer Vision Systems, Monitoring, Robust Tracking, Video Surveillance.*

**Participants:** James L. Crowley, Sebastien Pesnel, Alban Caporossi, Daniela Hall.

Tracking is a basic enabling technology for observing and recognizing human actions. A tracking system integrates successive observations of targets so as to conserve information about a target and its history over a period of time. A tracking system makes it possible to recognize an object using off-line (non-video rate) processes and to associate the results of recognition with a target when it is available. A tracking system makes it possible to collect spatio-temporal image sequences for a target in order to recognize activity. A tracking system provides a prediction of the current location of a target which can improve the reliability, and reduce the computational cost of observation.

Project PRIMA has implemented a robust real time detection and tracking system (CAR). This system is designed for observing the actions of individuals in a commercial or public environment, and is designed to be general so as to be easily integrated into other applications. This system has been filed with the APP "Agence pour la Protection des Programmes" and has Interdeposit Digital number of IDDN.FR.001.350009.000.R.P.2002.0000.00000. The basic component for the CAR systems is a method for robust detection and tracking of individuals [Schwerdt 00]. The system is robust in the sense that it uses multiple, complementary detection methods are used to ensure reliable detection. Targets are detected by pixel level detection processes based on back-ground subtraction, motion patterns and color statistics. The module architecture permits additional detection modes to be integrated into the process. A process supervisor adapts the parameters of tracking so as to minimize lost targets and to maintain real time response.

Individuals are tracked using a recursive estimation process. Predicted position and spatial extent are used to recalculate estimates for position and size using the first and second moments. Detection confidence is based on the detection energy. Tracking confidence is based on a confidence factor maintained for each target.

The CAR system uses techniques based on statistical estimation theory and robust statistics to predict, locate and track multiple targets. The location of targets are determined by calculating the center of gravity of detected regions. The spatial extent of a targets are estimated by computing the second moment (covariance) of detected regions. A form or recursive estimator (or Kalman filter) is used to integrate information from the multiple detection modes. All targets, and all detections are labeled with a confidence factor. The confidence factor is used to control the tracking process and the selection of detection mode.

outliers during estimation of the selecting and reinitializing the detection modes to reinitialize and adapt less

In 2003, with the assistance by INRIA Transfert and the GRAIN, Project PRIMA has founded a small enterprise, Blue Eye Video to develop commercial applications based on the CAR system. Blue Eye Video has been awarded an exclusive license for commercial application of the CAR tracker. In June 2003, Blue Eye Video was named Laureat of the national competition for the creation of enterprises.

## 6. New Results

### 6.1. A Programmable Robust Tracker

**Keywords:** *Computer Vision Systems, Event Detection, Monitoring, Robust Tracking, Video Surveillance.*

**Participants:** James L. Crowley, Alban Caporossi, Alba Ferrer-Biosca, Daniela Hall, Patrick Reigner.

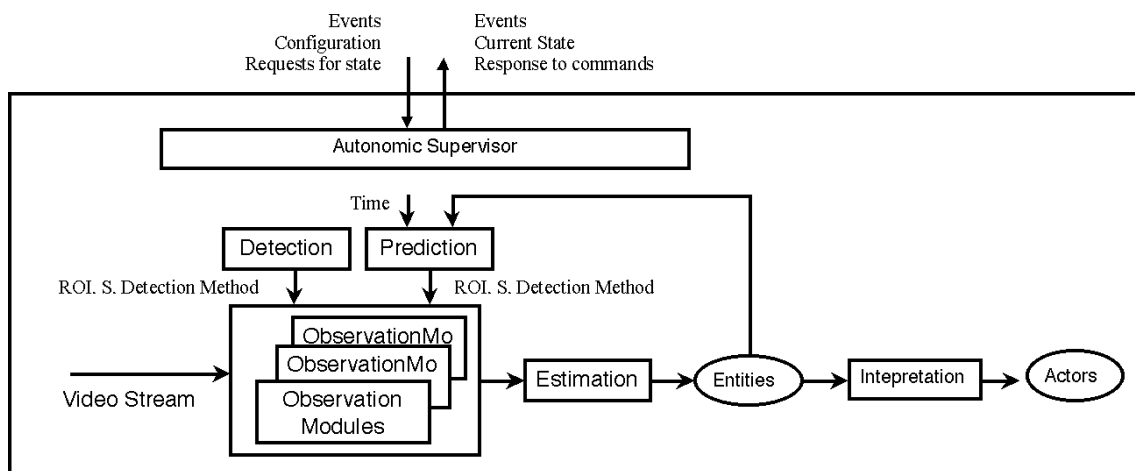


Figure 10. The components and architecture for the new agent detection and tracking process.

The CAR tracker, described in the section on Software Products was implemented in C++ with hard-wired control. In order to support experiments in observation of activity, a new programmable robust tracking project has been implemented in the ImaLab environment. The architecture for this process is shown in figure 10.

The visual observation process is organized as a set of modules controlled by an autonomic process supervisor. The supervisor provides four functional capabilities : module execution scheduling, parameter regulation, reflexive control, and interpretation and response to messages. The process supervisor is implemented using a Scheme interpreter. Scheme is a Lisp-like language with a very simple small and simple implementation. Interpreters for C++ and CLIPS rules have been added written in scheme for the process supervisor, making it possible to download and execute small snippets of code in C++, Lisp, or CLIPS in order to program new functions while the module is executed. This capability is currently used by the federation supervisor to configure modules for specific tasks at system set up.

The process supervisor iteratively executes seven phases of execution described in the following list. These phases provide the autonomic and reflexive control.

1. Acquisition: Get the next image from the video stream

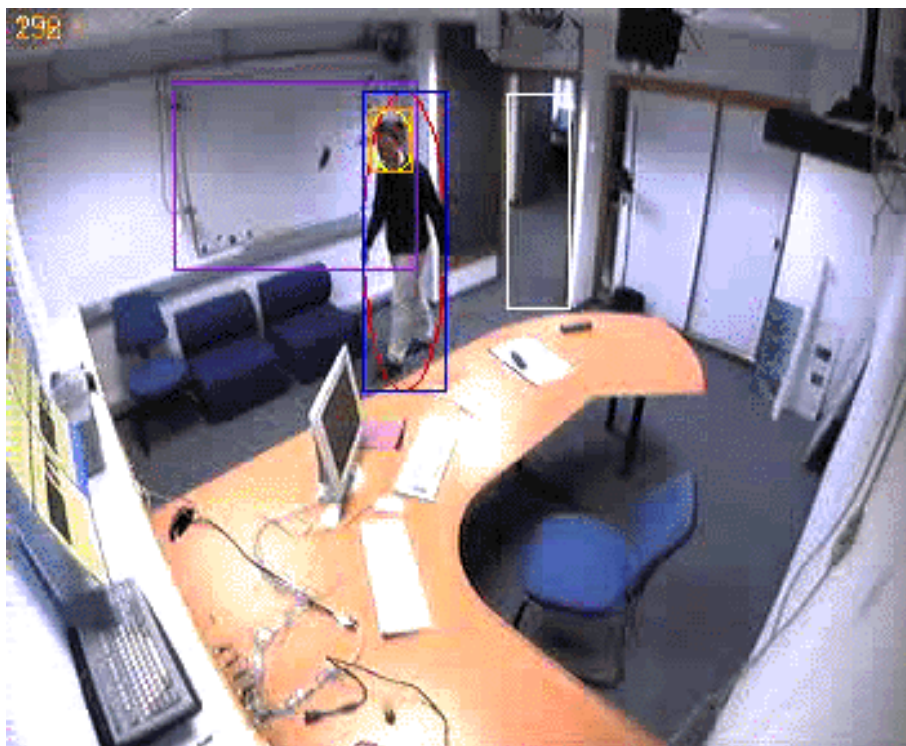


Figure 11. The new programmable robust tracker makes it possible to observe composite entities

2. Prediction and Observation: For each current target, predict a region of interest in the new image. Execute the detection module specified by the target in the region of interest, update the target parameters. This phase will delete targets whose confidence drops below a threshold.
3. Detection: For a subset of the list of "detection regions", execute the specified detection module. If a sufficient level of average detection energy is obtained in the region, calculate the position and size of the target using moments and add the target to the target list.
4. Regulation: Examine the time elapsed for each target and for each detection region during the current cycle. Regulate the precision or number of targets and detection regions so as to maintain a specified video rate or other quality of service such as target precision or priority.
5. Interpretation: Recognize configurations of targets as composite targets. For example, an agent is a composite target with a body and a face.
6. Event Detection: Generate events for targets and interpretations. Possible events include: detection and loss of a target, entry or exit to regions, entry or exit of the scene, overlap, split, and merge of agents or their components.
7. Respond to messages: Messages are text strings that may include requests for information about process state or may be new snippets of code to be added to the process supervision or interpretation phase.

During execution of each phase, the elapsed time is recorded. During the observation and detection phases, time is noted for each target or each detection region. During the regulation phase, if the elapsed time exceeds the available frame time, the resolution of tracking may be reduced to one pixel of  $N$ , the number of targets tracked may be reduced, or the number of detection regions may be reduced so as to maintain video rate.

The available detection modules include: adaptive background subtraction, color histogram tracking, image motion detection with hysteresis, and receptive field histograms. A new detection module based on object learned combinations of receptive fields will soon be added to the system.

Each target and each detection region includes a specification of the detection procedure that is to be used for detection and observation. This specification is a symbolic label that can be changed during the process regulation phase. As an example, figure 11 shows an example of a face and body detection. The body region was first detected based on motion energy in the detection region placed at the door (white rectangle). The body is tracked using subtraction from an adaptive background within the body ROI shown as a blue rectangle. The position and size of the body are shown as a red ellipse. When a body has been tracked for a sufficient number of frames, a detection region is created for the face at a position relative to the body. Skin color detection is run in this region. The interpretation process notes successful detection of the face and associates the body and face to create a composite entity for an actor (or person). Events are generated whenever actors are detected as well as when they enter and leave certain regions. Events are sent to the supervisor process and are used to signal changes in situation.



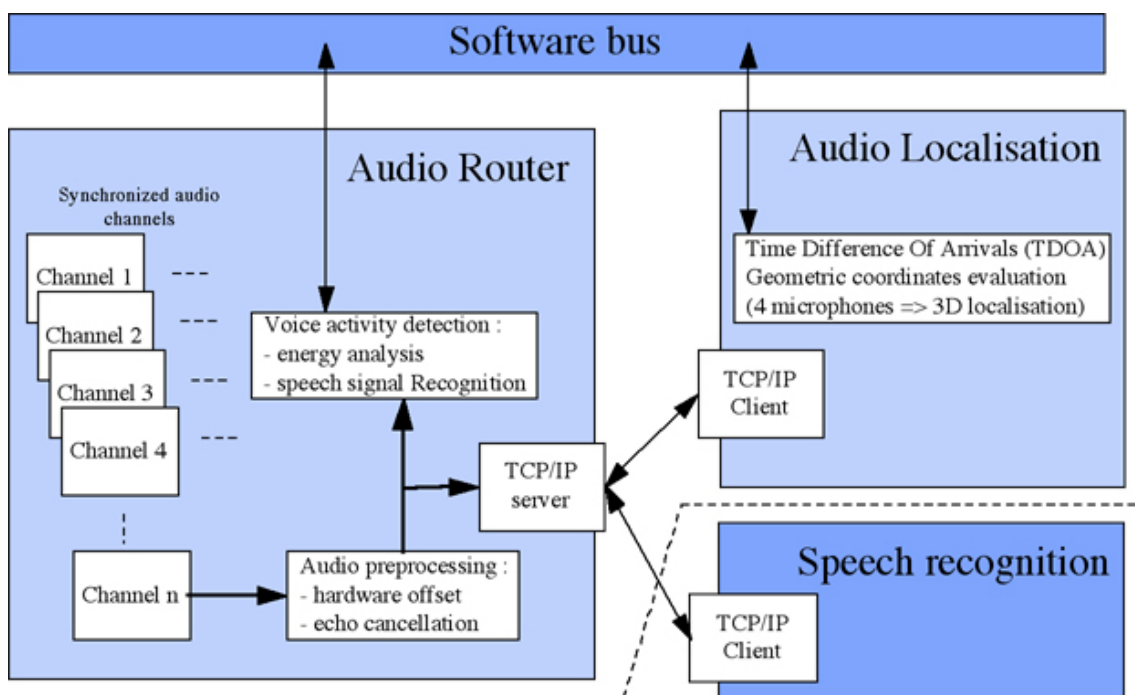


Figure 12. Processes for detection, recognition and tracking of Acoustic Sources

## 6.2. Audio Processes for Detection and Tracking

**Keywords:** *Acoustic Perception, Monitoring, Surveillance.*

**Participants:** Dominique Vaufreydaz, Patrick Reignier.

In addition to video tracking, Project PRIMA has also implemented processes for recognition and tracking of acoustic sources. Due to hardware compatibility, these processes are implemented under the MS Windows environment and communicate via the software bus. Acoustic perception is designed around a microphones array (with 4 or more microphones) and a set of lapel microphones. There are 4 modules included in AudioProcesses: "AudioRouter", "AudioLocalization", "SpeechRecognition" and "TopicSpotting".

AudioRouter is in charge of recording synchronously all the audio channels and to distribute audio data to other modules, and of some audio pre-processing: remove hardware recording offset and speech/non-speech classification. Speech classification techniques are used to detect speech activities on lapel or ambient microphones. Doing that, it is possible for example in the FAME context to determine if a lecturer or someone in his audience is speaking. According to the recognized context, acoustic signals tagged as speech can be sent to the SpeechRecognition module. The AudioRouter speech detection is based on a dynamic combination of 2 sub-modules: an energy detector and a neural network. The first one requires that the average signal energy over a specified (regulated) period be greater or smaller than a specified (regulated) threshold. All the periods and thresholds are established during system configuration and may be regulated by the supervisory controller. In parallel, the neural network is used to classify signals based on several temporal and spectral acoustic parameters (Linear Predictive Coding, zero-crossing, etc.). The neural network detects all voiced activity, i.e. sound that have been echoed in a human vocal track: plosive sounds are not recognized as speech but the following vowel is.

For AudioLocalization, the microphone array is composed of 4 microphones mounted at the corners of the presentation screen within the PRIMA Augmented Meeting Environment. 4 microphones are needed to do 3D sound localization. Relative phase information is used to recognize the source position for speech signals. Location estimation for an acoustic source is based on the Time Difference Of Arrivals (TDOA) [53]. The time lag between signals received at each microphones pair is determined using inter-correlation function between signal energy. The maximum of function between microphones provides a TDOA for each microphone pair. Then 2 methods are available for estimating the position of an acoustic source. The first method is a purely analytic approach. Given the relative position of microphones, each possible time delay corresponds to a sphere of positions whose distance correspond to the distance that sound travels during the delay. The relative TDOA of two microphones corresponds to a hyperbolic function that is the intersection of two spheres. Given three microphone pairs, one can compute the intersection of these hyperbolic functions to exactly predict the position of the acoustic source. Experience has shown that this intersection function is extremely unstable for most positions, due to echo. The second method is based on knowledge on a set of possible targets. It computes theoretical TDOAs using sources positions and calculates the distance with the estimated ones. The best target is then chosen with the minimal distance. In this case, we can use video targets' positions, given by the supervisory controller, to determine which system target is activated. Using threshold, it is possible to decide that a sound is not related to any known target. In this case, and under some assumptions, the controller can decide or not to launch a new video process in order to look after a new target.

The SpeechRecognition module uses state-of-the-art acoustic parameters (Mel-scaled Frequency Spectral Coefficient - MFCC -, energy, zero-crossing, variations and accelerations of these parameters). It is based on Hidden Markov Models for the acoustic module and on Statistical Language Model for the language modelling part. SpeechRecognition can recognize either lapel or ambient microphone signal. The TopicSpotting modules wait for messages from the SpeechRecognition. It can use 2 different approaches: a rule-based one using triggers and grammars, or a statistical one. In all case, using topic spotting information, the SpeechRecognition language models can be dynamically adapted to current interest of the speaker(s) [62].

## 6.3. The Process Federation Tool

**Participants:** Patrick Reignier, Sebastien Pesnel, Alban Caporossi, James L. Crowley.

The Process Federation Tool (PFT) has been defined to facilitate experiments in both manual and automatic configuration and supervision of perceptual processes and process federations. This tool allows processes and process federations to be specified, configured and monitored during execution, thus providing an important experimental tool for developing automatic methods for process regulation and federation configuration.

A process federation is a system of independent cooperating processes. A process federation provides a convenient mechanism to extract and integrate information from a network of sensors and cameras, without the need to communicate large volumes of high-bandwidth data. Federations can also be used to distribute processing over a network of computing devices in an ad-hoc manner in response to changes in operating context.

We have designed a middle ware environment that allows us to dynamically launch and connect process on different machines. In our system, processes are controlled by a federation supervisor or "federator". The federator configures a process by sending snippets of control script to be interpreted by the controller. Each control script defines a command that can be executed by a message from the federator. Processes may be interrogated by the federator to determine their current state and the current set of commands.

Federators can also launch and configure other federators so that federations can be built up hierarchically. Each federator invokes and controls lower level supervisors that perform the required transformation. At the lowest level are Perceptual processes that observe and track entities and observe the relations between entities. These are grouped into federations as required for to observe the situations in a context.

Design and debug of Federators is facilitated by the Process Federation Tool. The Process Federation Tool provides an interface that allows available federations to be displayed, configured, launched, monitored and saved. The tool provides a convenient means to launch, configure and monitor execution of individual processes or process federations. The user chooses the processes from a list of available types and defines values for parameters. The description results in the creation of a "ProcessDefinitions.scm" file. Processes and federations configured with the tool can be saved for further use in order to allow federations to be incrementally constructed and refined.

Processes available to the PFT are described by a process description file using an XML process description. Process descriptions are automatically imported from a selected process description directory, and displayed in a process selection window. The process selection window allows listed processes to be selected and instantiated for configuration. Instantiation allows processes to be configured and allocated to computing resources.

Process configuration requires knowledge of the set of available sensors on the selected matching, as well as knowledge of the process parameters that must be configured. The available sensors for each machine are described in a XML description that may be imported for each machine. As with processes, copies of these descriptions are kept in a predefined directory. A description of process parameters is obtained by sending a message to the newly instantiated process.

This auto-description feature for processes was designed to allow automatic configuration, but has been found to be very useful for manual configuration of distributed processes.

Instantiated processes may be connected by communication channels to form federations using the process federation window. Process federations may be specified from scratch using currently processes.

These descriptions can be named and saved using an XML format in a process federation directory. A list of previously specified federations is obtained by importing the XML descriptions from this directory. Selecting such a description triggers instantiation and default configuration for this federation.

The process federation specification window can display information about the process federation and its completeness. Incomplete federations are federations in which input channels lack processes.

The federation tool validates communication channels and indicates when a required communication channel is lacking.

### **6.3.1. An Example: Distributed Camera Net**

A simple example of a federation of perceptual processes is provided by a system that detects and tracks entities using a distributed network of cameras, each connected to a separate computer running a separate

robust tracking process, as shown in figure 13. Targets detected within each robust tracking process are reported to an entity composition process. The composition process assembles targets into composite entities and maintains a global history of the target evolution and trajectory. This system has been used to track vehicles and pedestrians within the INRIA CAVIAR parking lot test-bed combining results from up to 6 distributed surveillance cameras.

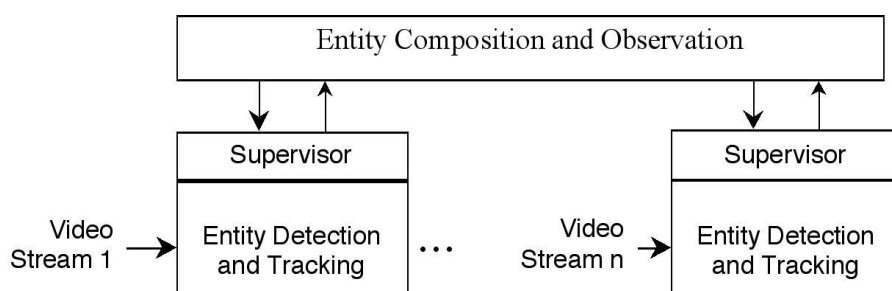


Figure 13. A simple process federation composed of entity detection process, and a composition process

## 6.4. Specifying a context model

**Keywords:** *Ambient Intelligence, Context Aware Systems, Context Modeling.*

**Participants:** Patrick Reignier, James L. Crowley.

A system exists to provide services. Providing services requires the system to perform actions. The results of actions are formalized by defining the output "state" of the system. Simple examples of actions for interactive environments include adapting the ambient illumination and temperature in a room. More sophisticated examples of tasks include configuring an information display at a specific location and orientation, or providing information or communications services to a group of people working on a common task.

The "state" of an environment is defined as a conjunction of predicates. The environment must act so as to render and maintain each of these predicates to be true. Environmental predicates may be functions of information observed in the environment, including the position, orientation and activity of people in the environment, as well as position, information and state of other equipment. The information required to maintain the environment state determines the requirements of the perception system.

The first step in building a context model is to specify the desired system behavior. For an interactive environment, this corresponds to the environmental states, defined in terms of the variables to be controlled by the environment, and predicates that should be maintained as true. For each state, the designer then lists a set of possible situations, where each situation is a configuration of entities and relations to be observed. Although a system state may correspond to many situations, each situation must uniquely belong to one state. Situations form a network, where the arcs correspond to changes in the relations between the entities that define the situation. Arcs define events that must be detected to observe the environment.

In real examples, we have noticed that there is a natural tendency for designers to include entities and relations that are not really relevant to the system task. Thus it is important to define the situations in terms of a minimal set of relations to prevent an explosion in the complexity of the system. This is best obtained by first specifying the environment state, then for each state specifying the situations, and for each situation specifying the entities and relations. Finally for each entity and relation, we determine the configuration of perceptual processes that may be used.

## 6.5. Context model compiler

**Participant:** Patrick Reignier.

PRIMA has constructed a graphical interaction tool for designing situation graphs. This tool allows situation graphs to be saved as an XML specification that is automatically transformed into a computer program that can observe and recognize situations and generate the desired actions.

### 6.5.1. Situation graphs and temporal relations

A context model is a graph of situations. Situations are connected by arcs, representing temporal constraints between them. They are decorated using the temporal operators defined by Allen [33] : *before*, *meets*, *overlaps*, *starts*, *equals*, *during*, *finished*.

The graph structure is given by the temporal relations. A path inside the graph is the result of the observation of the on-going situations.

A situation is a set of roles and relations. Based on the situation definition, we move from situation S1 to situation S2 if a role or a relation has changed in situation S1 (S1 is no more valid) and roles and relations are verified in situation S2. The transitions are event-driven. If we associate situations to places and events to transitions, the situation graph can be mapped on the *Synchronized Petri Nets* formalism. This Petri Net can then be transformed into a computer program.

### 6.5.2. Synchronized Petri Nets

A synchronized Petri Net is a Petri Net where transitions are associated to events. A transition can be fired if both :

- The preconditions on places marks are verified.
- The transition event has been received.

We have proposed for each Allen operator a corresponding Petri Net pattern. The synchronization events are automatically calculated based on the roles and relations of connected situations.

### 6.5.3. Jess rule generation

We have to program an event based system. One of the possible solution is to use a forward chaining rule programming environment. An event corresponds to a new fact in the facts database, triggering the corresponding rules.

We have selected the Jess expert system shell [29] for our rule based programming environment. The generated rules are separated in three groups :

- The rules implementing the structure of the Petri Net. They are a direct transcription of the Petri Net evolution function.
- The rules implementing the transition functions. They generate the synchronization events based on modification of roles and (or) relations.
- The rules implementing the control of the visual processes. These rules are based on the situation marks. When a situation mark is going to 0, we are not interested anymore in observing the entities playing the associated roles. We can shutdown the corresponding visual processes. When a situation mark is positive, we must configure the visual processes to search for entities playing the roles of all the connected situations. This is to be able to observe which situation will be the next one.

## 6.6. Generic Features for detection and classification

**Keywords:** *Computer Vision, Generic Features, Robust Matching, Robust Tracking.*

**Participants:** Daniela Hall, Nicolas Gourier.

Feature extraction is essential for detection and recognition tasks. In the case of detection and identification of agents in intelligent environments with multiple cameras, the use of raw Gaussian derivatives is going to fail due to the large variance introduced by changes in appearance, viewpoint and camera parameters. Appropriate features are robust to such changes by learning the expected variance from examples. In this section we propose an approach for the generation of such generic features and show how these low level features can be integrated in an architecture with two components (see Figure 14): a feature extraction module that provides features invariant to intra-class variability and a geometry verification module that introduces specificity, increases the reliability of the detection and rejects false positives.

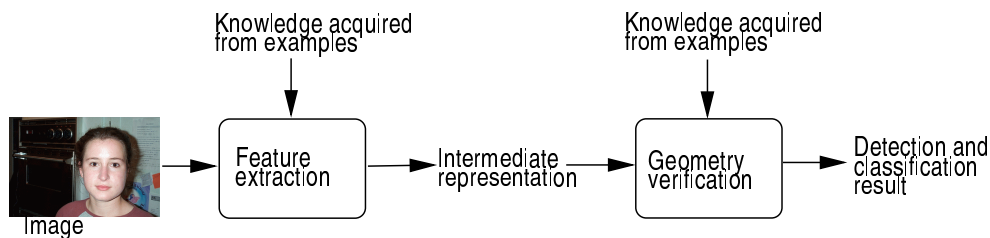


Figure 14. System architecture consisting of low level feature extraction and higher level geometry verification.

### 6.6.1. Computation of class-specific feature detectors

For the extraction of class-specific features, we learn the appearance of class-specific object parts from a dense, pixelwise, grid of features by clustering. Clustering of dense features is similar to Leung and Malik's approach for computation of generic features for texture classification [48]. Furthermore, k-means clustering produces statistically correct results, because a large amount of data points is used.

Figure 15 illustrates the feature extraction process. The top right graph shows an example of the probability maps (low probabilities are black, high probabilities are light grey). The probability maps can be condensed to a cluster map representation. This is an enormous data reduction, but at the same time, it preserves the information about class specific features and their location. This is the minimum information required for detection and classification and makes the cluster map an ideal input for the geometry verification module. For a detailed description of the experiments please see [45].

### 6.6.2. Modeling spatial relations

In this section we propose an automatic model generation that learns spatial relations of generic features. This model is inspired by Belongie's log-polar shape context [34]. A log-polar histogram has bins that are uniform in log-polar space. This makes the log-polar description appropriate for applications where the object undergoes affine transformations.

A region around the query position is transformed into log-polar representation using a lookup table for speedup. A bin of the log-polar histogram contains the ratio of the surface covered by the query pixel and the total surface of the histogram bin.

For learning the spatial relations of the target object, the user selects a reference position within the cluster map representation of a set of training images. This is the only user interaction required for training. A model histogram is constructed. For measuring the similarity between any query histogram  $Q$  and the model histogram  $H$  we use the  $\chi^2$  divergence measure.

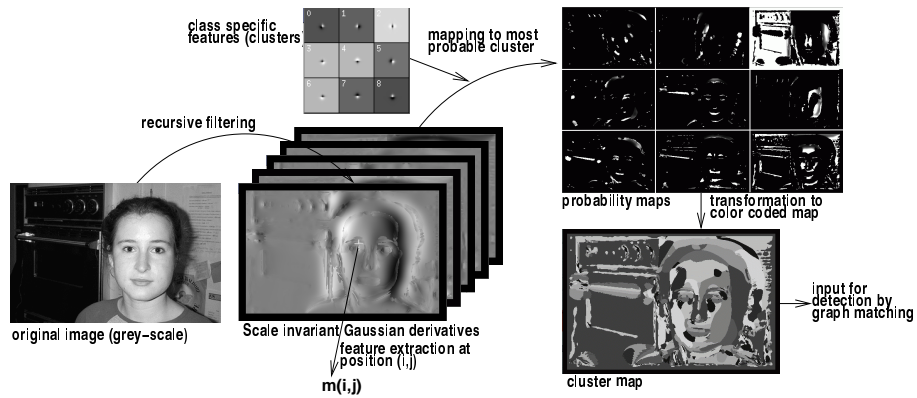


Figure 15. Algorithm for raw feature extraction and mapping to most probable class specific features. The probability maps are condensed to a single color coded cluster map, where color  $k$  marks points that are assigned to cluster  $k$ .



Figure 16. Detection example on unconstrained image. The training is performed on frontal faces from the AR database [50]. Detected faces marked by circles are characterized by the combined occurrence of facial features. No false detections are observed.

The log-polar implementation allows the modeling of spatial relations that is sufficiently discriminant to avoid false detections and is general enough to avoid over-fitting. In order to show the stability of our approach to images with cluttered background and different illumination conditions, we have performed a face detection experiment on the Caltech face database [65] on 435 images. We obtain a positive detection rate of 97.7% (425 out of 435 images are detected). In the example of unconstrained images in Figure 16 all faces are correctly detected despite significant differences between the training and observed faces, head pose, hairstyle and beard or glasses. We observe no false detections. This is a convincing result considering the variations in scale, head pose and lighting.

## 7. Contracts and Grants with Industry

### 7.1. European and National Projects

#### 7.1.1. IST-2000-28323 FAME: Facilitating Agent for Multi-Cultural Exchange

European Commission project IST-2000-28323

Starting Date : October 2001.

Duration: 40 months.

Key Action: MultiModal Interfaces

Consortium Members:

- Universitaet Karlsruhe (TH), Germany, Prof. Alex Waibel
- Laboratoire GRAVIR, UMR CNRS 5527, France, Prof. James L. Crowley
- Université Joseph Fourier, Laboratoire CLIPS, France, Prof. Joëlle Coutaz
- Istituto Trentino di Cultura, Italy, Marcello Federico
- Universitat Politècnica de Catalunya Centre TALP, Spain, Prof. José B. Mariño
- Sony International (Europe) GmbH, Germany, Ralf Kompe
- Applied Technologies on Language and Speech S. L., Germany, David Font

The goal of IST Project FAME is to construct an intelligent agent to facilitate communication among people from different cultures who collaborate on solving a common problem. This agent will provide three services: 1) facilitate human to human communication through multimodal interaction including vision, speech and object manipulation, 2) provide the appropriate information relevant to the context, and 3) make possible the production and manipulation of information blending both electronic and physical representations. The agent will serve as an information butler to aid multicultural communication in a transparent way. The agent will not intervene in the conversation, but will remain in the background to provide the appropriate support. A public demonstration is planned for the Barcelona Cultural Fair 2004.

management (WP1) and dissemination (WP11). Two to three demonstrators of increasing complexity will be designed and the interfaces between the components of the system will be specified (WP2). These demonstrators will concern scenarios for seminar style discussion, round-table group problem solving, and collaborative problem solving involving remote groups. Each scenario will include assistance for communication between people from different cultures using different languages. A test-bed will be constructed to verify compliance with the specifications and to provide an environment for evaluating the performance and usability of software components (WP3). At the end of the project, a robust demonstrator will be exhibited at the Barcelona Cultural Fair (WP2). Constructing a facilitating agent for multicultural communication requires technological innovations in several areas: speech understanding to control the acquisition and presentation of video (WP4) and audio (WP5) for communication. 2)Active control of acquisition to allow users to speak and move freely about the room unencumbered by microphones or wires. 3)Conversational speech recognition to recognise unplanned speech acts in order to feed the conversation context model with the meaning



of utterances (WP6). 4)Dialog modeling (WP10) to maintain a task model that integrates information from perceptual awareness in order to inform the information retrieval process. 5)An information retrieval process to provide information relevant to the current communication and problem solving context (WP7). 6)Natural interaction, blending physical and electronic representations of information, to enable people to speak, write and manipulate physical objects as part of the problem solving process (WP8). 7)Automatic speech translation to enable communication across cultural and language barriers (WP9).

### **7.1.2. IST-2001-32157 DETECT: Real Time Detection of Motion Picture Content in Live Broadcasts**

European Commission project IST-2001-32157

Start Date: Novembre 2001

Duration: 27 Months

Project: IST-2001-32157

Key Action: Cognitive Vision.

- JOANNEUM RESEARCH Forschungsgesellschaft mbH, Austria
- DUVIDEO II - Profissionais de Imagem S.A., Portugal
- Taurus Media Technik GmbH, Germany
- HS-ART Digital Service GmbH, Austria
- VIDEOCATION Fernseh-Systeme GmbH, Germany
- INRIA Rhône Alpes, INRIA-GRAVIR-UMR5527, France
- Institut National Polytechnique de Grenoble, INPG-GRAVIR-UMR5527, France
- Centre National de la Recherche Scientifique, CNRS-GRAVIR-UMR5527, France
- Université Joseph Fourier, UJF-GRAVIR-UMR5527, France

The principle goal of the DETECT project is to implement a general platform for real time detection of semantic blocks and regions within digital video streams. These detected regions are then be subject to further analysis and processing. The project will focus on the applications problem of detecting and delimiting predefined static and dynamic objects. This issue has currently a very large demand for both cultural and economic reasons. DETECT is an industrial driven project, although its nature is R&D. The project will result in a prototype, which can be turned into a product after the project. For this reason the main modules are implemented as sample applications (ProcessingUnits) for the categories which are of high commercial interest (e.g.: identification of race-cars and soccer-players).

DETECT provides a general platform for a real time analysis for streaming video input and supports three different types of ProcessingUnits (modules) which are as follows:

#### 7.1.2.1. *Detection of semantic blocks.*

A semantic block covers the temporal domain only. Thus a typical semantic block-detector just indicates, whether the streamed content is of a certain type or not. Within DETECT the semantic block-concept will be implemented for commercial/advertising blocks, as they appear frequently in television-broadcast. Depending on the outcome of the semantic block detection, a specific detection of regions of interest (further on ROI) can be applied:

#### 7.1.2.2. *Detection of ROI in order to identify static and dynamic objects.*

ROI is a specific local region which is due to the nature of the streaming input also related to the temporal domain. Such ROI, whenever identified in real time can be used to restrict further analysis or to simply recognize predefined objects, which match with the ROI. Within DETECT the ROI-concept will be applied to sports-applications (soccer and formula 1) and therein to locally moving objects like soccer-player and race-car, but also to static objects like hoardings. Each detected ROI can be further analyzed with pattern recognition tools depending on the type of the ROI.

#### 7.1.2.3. *Motion picture analysis.*

The main objective herein is, to detect predefined (company) logos stored in a central reference database. Those logos can be trademarks like in the DETECT sample application, but could also be of nearly any other type. As the size of the reference-system has to be scalable, the logo-detection analysis will be done off-line not as a real time application

### 7.1.3. **IST 2001 37540 CAVIAR: Context Aware Vision using Image-based Active Recognition**

European Commission project: IST 2001 37540

Starting Date: October 1, 2002

Durations: 36 Months

Key Action: Cognitive Vision

Consortium:

- Univ. of Edinburgh (United Kingdom)
- Instituto Superior Tecnico, Lisbon, Portugal
- INRIA Rhône Alpes, France
- Institut National Polytechnique de Grenoble, France
- Université Joseph Fourier, Grenoble, France
- CNRS, France

The main objective of the CAVIAR is to address the scientific question: Can rich local image descriptions from fovea and other image sensors, selected by a hierarchal visual attention process and guided and processed using task, scene, function and object contextual knowledge improve image-based recognition processes? This is clearly addressing issues central to the cognitive vision approach.

The two applications that the project will address are:

1. City centre surveillance: Many large cities have nighttime crime and antisocial behaviour problems, such as drunkenness, fights, vandalism, breaking and entering shop windows, etc. Often these cities have video cameras already installed, but what is lacking is a semi-automatic analysis of the video stream. Such analysis could detect unusual events, such as patterns of running people, converging people, or stationary people, and then alert human security staff.
2. Marketers are interested in the behaviour of potential customers in a commercial setting, such as what sequence of locations do they visit, how long they stop at particular locations, what behavioural options do typical customers take, etc. Automatic analysis of customer behaviour could enable evaluation of shop layouts, changing displays and the effect of promotional materials.

#### 7.1.4. IST 506909 CHIL: *Computers in the Human Interaction Loop*

European Commission project IST 506909 (Framework VI - Call 1)

Strategic Objective: Multi-modal Interaction

Start Date 1 January 2004.

Duration 36 months (renewable).

CHIL is an Integrated Project in the new Framework VI programme.

Participants

- Fraunhofer Institut für Informations- und Datenverarbeitung, Karlsruhe, Germany
- Universität Karlsruhe (TH), Interactive Systems Laboratories, Germany
- Daimler Chrysler AG, Stuttgart, Germany
- ELDA, Paris, France
- IBM Czech Republic, Prague, Czech Republic
- Research and Education Society in Information Systems, Athens, Greece
- Institut National Polytechnique de Grenoble, France
- Instituto Trentino di Cultura, Trento, Italy
- Kungl Tekniska Högskolan (KTH), Stockholm, Sweden
- Centre National de la Recherche Scientifique, Orsay, France
- Technische Universiteit Eindhoven, Eindhoven, Netherlands
- Universität Karlsruhe (TH), IPD, Karlsruhe, Germany
- Universitat Politecnica de Catalunya, Barcelona, Spain
- Stanford University, Stanford, USA
- Carnegie Mellon University, Pittsburgh, USA

The theme of project IP CHIL is to put Computers in the loop of humans interacting with humans. To achieve this goal of Computers in the Human Interaction Loop (CHIL), the computer must engage and act on perceived human needs and intrude as little as possible with only relevant information or on explicit request. The computer must also learn from its interaction with the environment and people. Finally, the computing devices must allow for a dynamically networked and self-healing hardware and software infrastructure. The CHIL consortium will build prototypical, integrated environments providing:

**Perceptually Aware Interfaces:** Perceptually aware interfaces can gather all relevant information (speech, faces, people, writing, and emotion) to model and interpret human activity, behaviour, and actions. To achieve this task we need a variety of core technologies that have progressed individually over the years: speech recognition and synthesis, people identification and tracking, computer vision, automatic categorization and retrieval, to name a few. Perceptually aware interfaces differ dramatically from past and present approaches, since the machine now observes human interaction rather than being directly addressed. This requires considerably more robust and integrated perceptual technology, since perspectives, styles and recording conditions are less controlled and less predictable, leading to dramatically higher error rates.

**Cognitive Infrastructure:** The supporting infrastructure that will allow the perceptual interfaces to provide real services to the users needs to be dramatically advanced. Cognitive and Social modeling to understand human activities, model human workload, infer and predict human needs has to be included in the agent and middleware technology that supports CHIL. Further, the network infrastructure has to be dynamic and reconfigurable to accommodate the integration of a variety of platforms, components, and sensory systems to collaborate seamlessly and on-demand to satisfy user needs.

**Context Aware Computing Devices:** CHIL aims to change present desktop computer systems to context aware computing devices that provide services implicitly and autonomously. Devices will be able to utilize the

advanced perceptual interfaces developed and the infrastructure in CHIL to free the user and allow him instead of serving the device to be served and supported in the tasks and human-to-human interactions he needs to focus. Further, human centered design, where the artistic value, appeal, and look & feel, become important in taking computing devices and human environments to the next level.

Novel services: The above innovations and advances in perceptual interfaces, cognitive infrastructure and context aware computing devices are integrated and showcased in novel services that aim at radically changing the way humans interact with computers to achieve their tasks in a more productive and less stressful way. These services are based on a thorough understanding of the social setting, the task situation, and the optimal interaction that maximizes human control while minimizing workload. Furthermore, some issues of privacy and security are to be addressed since the change human-computer interaction introduced by CHIL also touches a lot of the ways information in which is shared and communicated.

New measures of Performance: The resulting systems should reduce workload in measurable ways. To achieve these breakthroughs in a number of component technologies, the integrated system and a better understanding of its new use in human spaces are needed. Evaluation must be carried out both, in terms of performance and effectiveness to assess and track progress of each component, and the "end to end" integrated system(s). This will be carried out by an independent infrastructure that would also allow any third party to benchmark its findings against the project results after the end of the project.

### **7.1.5. RNTL/Proact: ContAct Context management for pro-Active computing**

Start Date February 2003.

Duration: 24 months

The consortium consists of five partners:

- Xerox Research Centre Europe (Project coordinator)
- Project PRIMA, Laboratoire GRAVIR, INRIA Rhone Alpes
- Neural Networks Research Centre, Helsinki University of Technology (HUT), Finland
- Jaakko Pöyry Consulting, Helsinki, Finland
- Ellipse, Helsinki, Finland

Project Contact is one of three RNTL projects that have been included in the French-Finland scientific program: ProAct.

The aim of Project RNTL CONTACT is to explore novel approaches, based mainly on neural nets<sup>2</sup>, to the detection and manipulation of contextual information to support proactive computing applications, and to make the results available as part of a more extensive toolkit for ubiquitous and proactive computing. The project will investigate the effectiveness of neural networks as part of a "contextual middleware". In particular the project will address two levels of context manipulation:

Support for developing and adapting neural network classifiers to compute context variables. Potentially, an end user will be able to specify the learning phase of the network and associate it with a specific new context attribute to be used in the application. The provision of example classifiers already trained. In this case some samples of such attributes will be developed and provided as a library. To develop these functions a specific scenario will be used: people planning, coordinating and reflecting on their activities in an organization in an augmented smart building (equipped with large screens, wireless networks, video cameras, identification sensors, and mobile devices). The attributes will be articulated at two different levels of abstraction: sensor oriented and application/user oriented

To achieve these results project CONTACT will cover four major activities: Definition of an ontology that describes context variables both at the user and at the sensor level. Definition of a platform providing formalism and an appropriate architecture to learn and combine context attributes. Definition of a library of context attributes, general enough to be reusable in support of different scenarios than the one used in the project. Validation of the contextual middleware on a pilot case. The chosen application of personal time management will help guide the development of the middleware and also to conduct an evaluation of our technology using a real-world problem.

## 8. Other Grants and Activities

### 8.1. European Research Networks

#### 8.1.1. *IST-2001-35454 ECVision: European Research Network for Cognitive AI-enabled Computer Vision Systems*

Project Acronym: ECVision

Project Full Title: European Research Network for Cognitive AI-enabled Computer Vision Systems

Start Date: March 2002

Duration: 36 months

ECVision is a thematic network devoted to Cognitive Enabled Computer Vision Systems. ECVision serves to unify the set of 8 IST projects funded in Framework V under the EC's Cognitive Vision program.

The principal goal of ECVision is to promote research, education, and application systems engineering in cognitive AI-enabled computer vision in Europe through focussed networking, multi-disciplinary peer-interaction, targeted identification of priority issues, and wide-spread promotion of the area's challenges and successes within both the academic and industrial communities.

The project goal can be realized by achieved by setting up and running a research network with the following objectives: These objectives will be accomplished through four main operational goals:

Research Planning - identify key challenges, problems, and system functionalities so that the community and the EC can target the critical areas efficiently and effectively. In doing so, ECVision will develop a 'research roadmap' which will identify the key challenges and priority topics, together with plans and time scales for attacking them.

Education and Training - identify and develop courses, curricula, texts, material, and delivery mechanisms; promote excellence in education at all levels, and foster exchange of ideas through inter-institutional interaction of staff and students. Information Dissemination - promote the visibility and profile of cognitive vision at conferences and in journals by organizing special sessions, workshops, tutorials, summer schools, short courses, and by providing links to the work of those in the AI & Robotics communities. Industrial Liaison - identify application drivers and highlight any successes, promote research trials, addressing all types of industries: games, entertainment, white goods manufacturers (e.g. vigilant appliances), construction (e.g. smart buildings), medicine (e.g. aids for the disabled), etc.

In addition, the network will include two support activities:

- Provision of an Information Infrastructure for both computer-supported cooperative work, e.g. discussion forums and email distribution lists, and for web-based dissemination of all material generated under the four areas identified above.
- Operational management by a Network Coordinator and Area Leaders in each of the four areas above; these people will constitute the ECVision Executive Committee.

James Crowley of Project prima is coordinator of Research Planning for ECVision.

### **8.1.2. IST-2000-26434 FGnet: Face and Gesture Recognition Working Group**

Start Date: 15 October 2001

Duration: 36 months

- University of Manchester, UK
- Technical University of Munich, Germany
- Computer Vision & Media Technology Lab, Aalborg University, Denmark
- Laboratoire GRAVIR, INRIA Rhône Alpes, France
- The Dalle Molle Institute for Perceptual Artificial Intelligence, Switzerland
- Dept. of Computer Science, Cyprus College, Nicosia, Cyprus

FGnet is a thematic network devoted to visual techniques for detection, tracking and recognition of faces and gestures. The aim of this project is to encourage technology development in the area of face and gesture recognition. The precise goals are: (1) to act as a focus for the workers developing face and gesture recognition technology (2) to create a set of foresight reports defining development road maps and future use scenarios for the technology in the medium (5-7 years) and long (10-20 years) term (3) to specify, develop and supply resources (e.g. image sets) supporting these scenarios (4) to use these resources to encouraging technology development. The use of shared resources and data sets to encourage the development of complex process and recognition systems has been very successful in the speech analysis and recognition field, and in the image analysis field in the specific cases where it has been applied. The basis of project, is that when properly defined and collects, such resources would also be of benefit in the development of wider problems in face and gesture recognition.

Project PRIMA is responsible for organizing dissemination workshops for FGnet, as well as contributing to the collection of benchmark data sets for performance evaluation. FGnet has provided resources for organizing the PETS series of workshops (Performance Evaluation for Tracking and Surveillance).

## **9. Dissemination**

### **9.1. Contribution to the Scientific Community**

#### **9.1.1. EUSAI '04: European Symposium on Ambient Intelligence**

James L. Crowley was program co-chair of the 2nd European Symposium on Ambient Intelligence, held in Eindhoven in the Netherlands, on November 8-10, 2004.

EUSAI provides a venue for an emerging multi-disciplinary community of researchers that work on Ambient Intelligence. Ambient Intelligence represents a vision of the future where we shall be surrounded by electronic environments, sensitive and responsive to people. Ambient intelligence technologies are expected to combine concepts of ubiquitous computing and intelligent systems putting humans in the center of technological developments.

Ambient Intelligence represents a long-term objective for European research bringing together researchers across multiple disciplines: computer science, electronics and mechanical engineering, design, architecture, social sciences, software engineering.

EUSAI '04 presented research in the areas of Ubiquitous computing, Context Awareness, Intelligence and Natural user-system interaction. Conference Proceedings are available from Springer Verlag Lecture Notes in Computer Science.

#### **9.1.2. PETS '04: Performance Evaluation for Tracking and Surveillance**

James L. Crowley organized the

The Sixth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2004) held in Prague, 10 May 2004, in collaboration with ECCV '04.

The theme for PETS '04 has been the automatic observation of human activity in public places. Six scenarios were recorded with a wide-angle camera lens in the entrance lobby of the INRIA Rhône-Alpes research laboratory in Montbonnot France, using members of the CAVIAR project as actors. Activities included a person walking in a straight line (3 sequences), a person browsing at information displays (5 sequences), behaviours while seated in a chair (3 sequences), persons abandoning packages (5 sequences), groups of people encountering (6 sequences), and people fighting (4 sequences). For each scenario, a ground-truth file has been constructed to indicate a bounding box for each individual, activity labels for each individual (appear, disappear, occluded, inactive, active, walking, running), a scenario label for each individual (fighter role, browser role, left victim role, leaving group role, walker role, left object role) and a situation label for each frame: (moving, inactive, browsing) and a scenario label for each frame (browsing, immobile, walking, drop down). These ground truth files have been made public for half of the sequences for use as training data. The PETS challenge was to demonstrate automatic labeling for the remaining sequences.

Authors were asked to describe the tracking and recognition methods, estimate or measure computational costs, and present error rates obtained with the published ground truth. Authors are also invited to propose new performance evaluation metrics that might be of interest.

The workshop was attended by 27 registered participants, assisted by several unregistered "walk-ins". The program was composed of 11 presentations, selected from 12 papers submitted to the review process. The PETS '04 workshop has helped to establish the CAVIAR labeled data sets as a known and widely used benchmark data set for the community.

Datasets are available on the FGNet web server, maintained by PRIMA ([www-prima.inrialpes.fr/FGnet](http://www-prima.inrialpes.fr/FGnet)).

### **9.1.3. Pointing '04: International Workshop on Deictic Gestures**

Project PRIMA organized "Pointing 2004: International Workshop on Visual Observation of Deictic Gestures" in Cambridge, UK, on the 22 August 2004, in conjunction with the International Conference on Pattern Recognition (ICPR). The workshop was attended by 29 participants and included 8 formal presentations and extensive informal discussion after each presentation.

The theme for the POINTING '04 workshop was pointing (or deictic) gestures. Project PRIMA has prepared video images and sequences of people pointing at targets with their hands and faces. Data has been prepared by projecting a target on a wall using a steerable video projector. Subject were seated in a room at a distance of 3 meters from the wall and asked to point at the target with their right hand and with their face. Video sequences were recorded from four cameras placed at different positions around the subject. Still images were also obtained for each gesture. For each gesture a ground truth was automatically recorded in the form of the known target position.

Participants have been invited to compete in recognizing or estimating the target position with four categories of data: Stereo image pairs of hand pointing. Monocular images sequences of hand pointing, monocular images of face pointing. A third data set has been included in which pointing gestures are observed from a hat mounted camera.

Two sets of sequences and still image sets have been published on the workshop web site. Ground truth data has been provided for the first set, but withheld from the second. For each image or sequence, participants have been invited to determine: precision of target estimation and probability of failure. The workshop has been organized with three objectives:

1. To provide encourage comparative evaluation of vision techniques.
2. To encourage the definition and use of new metrics for performance evaluation.
3. To provide a publicly available benchmark by which future algorithms may be compared.

The data sets will remain available via the FGNET web site maintained by Project PRIMA.

#### **9.1.4. RFIA '04: Reconnaissance de Forme et Intelligence Artificielle**

James L. Crowley was a member of the editorial Board for RFIA 2004 : the “14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle”, 28 - 30 January 2004, in Toulouse

#### **9.1.5. Participation on Conference Program Committees**

James L. Crowley served as a member of the program committee for the following conferences.

- CVPR 2005, IEEE International Conference on Computer Vision and Pattern Recognition, 2005
- ICRA 2005, IEEE International Conference on Robotics and Automation, 2005
- EUSAI 2004, European Symposium on Ambient Intelligence, Eindhoven, Oct. 2004
- ICMI 2004, IEEE Conf. on Multimodal Interaction, Penn State, Oct. 2004
- ICPR 2004, International Conference on Pattern Recognition, Cambridge, UK, August 2004
- CVPR 2004, IEEE Conf. on Computer Vision and Pattern Recognition, Washington, June 2004
- FG 2004, International Conference on Automatic Face and Gesture Recognition, June 2004.
- ECCV 2004, European Conference on Computer Vision, Prague, June 2004.
- IAS 2004, Intelligent Autonomous Systems, March 2004.
- RFIA 2004, Reconnaissance des Formes et Intelligence Artificielle, Jan 2004

#### **9.1.6. Participation on Advisory Panels**

During April and May 2004, James L. Crowley has served evaluation panel on FET proactive Programme "Presence".

#### **9.1.7. Invited Plenary Presentations at Conferences**

- "What does it mean to see? Recent progress in Computer Vision", Plenary presentation at the Polish National Robotics Conference, Poland, June 2004
- "Context Driven Observation of Human Activity", Présentation au colloque : PSIPS Processing Sensory Information for Proactive Systems, Oulu June 2004.
- "Context Driven Observation of Human Activity", Présentation invite à France-Taiwan Colloque sur la Technologie Informatique, Ecole Polytechnique, Avril 2004.



## 10. Bibliography

### Major publications by the team in recent years

- [1] O. CHOMAT, V. COLIN DE VERDIÈRE, D. HALL, J. CROWLEY. *Local Scale Selection for Gaussian Based Description Techniques*, in "European Conference on Computer Vision, Dublin, Ireland", June 2000, p. I 117–133.
- [2] J. CROWLEY, F. BÉRARD. *Multi-Modal Tracking of Faces for Video Communications*, in "IEEE Conference on Computer Vision and Pattern Recognition, CVPR '97, San Juan, Puerto Rico", June 1997, p. 640–645.
- [3] J. L. CROWLEY, J. COUTAZ, F. BERARD. *Things that See: Machine Perception for Human Computer Interaction*, in "Communications of the A.C.M.", vol. 43, n° 3, March 2000, p. 54-64.
- [4] J. CROWLEY, J. COUTAZ, G. REY, P. REIGNIER. *Using Context to Structure Perceptual Processes for Observing Activity*, in "UBICOMP, Sweden", September 2002.
- [5] J. L. CROWLEY. *Integration and Control of Reactive Visual Processes*, in "Robotics and Autonomous Systems", vol. 15, n° 1, December 1995.
- [6] J. L. CROWLEY. *Vision for Man machine interaction*, in "Robotics and Autonomous Systems", vol. 19, n° 3-4, April 1997, p. 347-359.
- [7] C. L. GAL, J. MARTIN, A. LUX, J. L. CROWLEY. *Smart Office: An Intelligent Interactive Environment*, in "IEEE Intelligent Systems", July/August 2001.
- [8] D. HALL, V. COLIN DE VERDIÈRE, J. CROWLEY. *Object Recognition using Coloured Receptive Fields*, in "European Conference on Computer Vision, Dublin, Ireland", June 2000, p. I 164–177.
- [9] B. SCHIELE, J. CROWLEY. *Recognition without Correspondence using Multidimensional Receptive Field Histograms*, in "International Journal of Computer Vision", vol. 36, n° 1, January 2000, p. 31–50.
- [10] K. SCHWERDT, J. CROWLEY. *Robust Face Tracking using Color*, in "International Conference on Automatic Face and Gesture Recognition, Grenoble, France", March 2000, p. 90–95.

### Articles in referred journals and book chapters

- [11] J. COUTAZ, J. CROWLEY, S. DOBSON, D. GARLAN. *Context is Key*, in "Communications of the ACM, Special issue on the Disappearing Computer", to appear, March 2005.
- [12] J. CROWLEY, J. PIATER. *Introduction to the special issue: International Conference on Vision Systems*, in "Machine Vision and Applications", vol. 16, n° 1, 2004, p. 4–5.
- [13] D. HALL, F. PÉLISSON, O. RIFF, J. CROWLEY. *Brand Identification using Gaussian Derivative Histograms*, in "Machine Vision and Applications", vol. 16, n° 1, 2004, p. 41–46.

- [14] A. LUX. *The Imalab Method for Vision Systems*, in "Machine Vision and Applications", vol. 16, n° 1, 2004, p. 21–26.

## Publications in Conferences and Workshops

- [15] S. BORKOWSKI, J. LETESSIER, J. CROWLEY. *Spatial Control of Interactive Surfaces in an Augmented Environment*, in "European Conference on Human Computer Interaction, EHCI 04", July 2004.
- [16] S. BORKOWSKI, S. SABRY, J. CROWLEY. *Camera-projector sensing and actuating abilities*, in "Polish National Robotics Conference", June 2004.
- [17] O. BRDICZKA, P. REIGNIER, J. CROWLEY. *Automatic Development of an Abstract Context Model for an Intelligent Environment*, in "International Conference on Pervasive Computing, PerCom 05, Kuai, Hawaii", to appear, March 2005.
- [18] A. CAPOROSI, D. HALL, P. REIGNIER, J. CROWLEY. *Robust visual tracking from dynamic control of processing*, in "Performance Evaluation of Tracking and Surveillance, Prague, Czech Republic", May 2004, p. 23–31.
- [19] J. CROWLEY. *Context Aware Observation of Human Activity*, in "PSIPS, Oulu, Finland", 2004.
- [20] N. GOURIER, D. HALL, J. CROWLEY. *Estimating Face orientation from Robust Detection of Salient Facial Structures*, in "FG Net Workshop on Visual Observation of Deictic Gestures (POINTING)", August 2004.
- [21] N. GOURIER, D. HALL, J. CROWLEY. *Facial feature detection robust to pose, illumination, and identity*, in "International Conference on Systems, Man and Cybernetics, Special track on Automatic Facial Expression Analysis", October 2004.
- [22] D. HALL, J. CROWLEY. *Détection du visage par caractéristiques génériques calculées à partir des images de luminance*, in "Reconnaissance des formes et intelligence artificielle, Toulouse, France", January 2004.
- [23] J. LETESSIER, F. BERARD. *Visual tracking of bare fingers for interactive surfaces*, in "17th annual ACM symposium on User interface software and technology (UIST'04), Santa Fe, USA", A. PRESS (editor), October 2004, p. 119–122.
- [24] J. LETESSIER, J. CROWLEY. *Video sequence database for 3D finger pointing tasks*, in "ICPR Workshop on Pointing Gestures (Pointing'04), Cambridge, United Kingdom", September 2004.
- [25] T. TRAN, A. LUX. *A method for ridge extraction*, in "Asian Conference on Computer Vision", 2004.
- [26] L. VILLASEÑOR-PINEDA, M. MONTES-Y-GÓMEZ, D. VAUFREYDAZ, J.-F. SERIGNAT. *Experiments on the Construction of a Phonetically Balanced Corpus from the Web*, in "Conference on Intelligent Text Processing and Computational Linguistics CICLing-2004, Seoul, Korea", Lecture Notes in Computer Science, vol. 2945, Springer, February 2004.

## Bibliography in notes

- [27] J. L. CROWLEY, H. I. CHRISTENSEN (editors). *Vision as Process*, Springer Verlag, 1993.
- [28] A. FINKELSTEIN, J. KRAMER, B. NUSEIBEH (editors). *Software Process Modeling and Technology*, Research Studies Press, John Wiley and Sons Inc, 1994.
- [29] *Jess : the rule engine for the java*, <http://herzberg.ca.sandia.gov/jess/>.
- [30] *JORAM*, <http://joram.objectweb.org/>.
- [31] M. ABRAMOWITZ, I. STEGUN. *Handbook of Mathematical Functions*, MIT Press, 1965.
- [32] S. AGARWAL, D. ROTH. *Learning a sparse Representation for Object Detection*, in "European Conference on Computer Vision", 2002, p. 113-130.
- [33] J. ALLEN. *Towards a general theory of action and time*, in "Artificial Intelligence", vol. 13, 1984.
- [34] S. BELONGIE, J. MALIK, J. PUZICHA. *Shape Matching and Object Recognition Using Shape Context*, in "Pattern Analysis and Machine Intelligence", vol. 24, n° 4, April 2002, p. 509–522.
- [35] S. BORKOWSKI, O. RIFF, J. L. CROWLEY. *Projecting Rectified Images in an Augmented Environment*, in "PROCAMS'03 Workshop", 2003.
- [36] F. BÉRARD. *The Magic Table: Computer-Vision Based Augmentation of a Whiteboard for Creative Meetings*, in "Proceedings of the ICCV Workshop on Projector-Camera Systems", IEEE Computer Society Press, 2003.
- [37] O. CHOMAT, J. CROWLEY. *Probabilistic Recognition of Activity using local appearance*, in "Computer Vision and Pattern Recognition, Fort Collins, USA", June 1999, p. 104–109.
- [38] O. CHOMAT, V. COLIN DE VERDIÈRE, D. HALL, J. CROWLEY. *Local Scale Selection for Gaussian Based Description Techniques*, in "European Conference on Computer Vision, Dublin, Ireland", June 2000, p. I 117–133.
- [39] J. CROWLEY, J. COUTAZ, G. REY, P. REIGNIER. *Using Context to Structure Perceptual Processes for Observing Activity*, in "UBICOMP, Sweden", September 2002.
- [40] J. CROWLEY. *Integration and Control of Reactive Visual Processes*, in "Robotics and Autonomous Systems", vol. 15, n° 1, 1995.
- [41] J. CROWLEY, O. RIFF, J. PIATER. *Fast Computation of Characteristic Scale using a Half Octave Pyramid*, in "International Workshop on Cognitive Computing, Zurich, Switzerland", September 2002.
- [42] J. ESTUBLIER, P. Y. CUNIN, N. BELKHATIR. *Architectures for Process Support Interoperability*, in "ICSP5", 1997.

- 
- [43] R. FERGUS, P. PERONA, A. ZISSERMAN. *Object Class Recognition by Unsupervised Scale-Invariant Learning*, in "Computer Vision and Pattern Recognition, Madison, USA", 2003.
- [44] W. FREEMAN, E. ADELSON. *The Design and Use of Steerable Filters*, in "Pattern Analysis and Machine Intelligence", vol. 13, n° 9, September 1991, p. 891–906.
- [45] D. HALL, J. CROWLEY. *Détection du visage par caractéristiques génériques calculées à partir des images de luminance*, in "Reconnaissance des formes et intelligence artificielle, Toulouse, France", to appear, 2004.
- [46] B. JOHANSON, G. HUTCHINS, T. WINOGRAD, M. STONE. *PointRight: Experience with Flexible Input Redirection in Interactive Workspaces*, in "Proceedings of UIST-2002", 2002.
- [47] J. J. KULIKOWSKI, P. O. BISHOP. *Fourier Analysis and Spatial Representation in the visual Cortex*, in "Experientia", vol. 37, n° 1, 1981, p. 160–163.
- [48] T. LEUNG, J. MALIK. *Recognizing Surfaces using Three-dimensional Textons*, in "International Conference on Computer Vision, Corfu, Greece", September 1999.
- [49] A. LUX. *The Imalab Method for Vision Systems*, in "ICVS03, Graz, Austria", April 2003.
- [50] A. MARTINEZ, R. BENAVENTE. *The AR Face Database*, Technical report, n° 24, CVC, June 1998.
- [51] N. NAKAMURA, R. HIRAIKE. *Active Projector: Image correction for moving image over uneven screens*, in "Companion of the 15th Annual ACM Symposium on User Interface Software and Technology", October 2002, p. 1–2.
- [52] Y. OHTA, Y. HAMASHI. *Recovery of Illuminant and Surface Colors from Images Based on the CIE Daylight*, in "European Conference on Computer Vision", J.-O. EKLUNDH (editor), 1994.
- [53] M. OMOLOGO, P. SVAIZER. *Use of the Crossposwer-Spectrum Phase in Acoustic Event Location*, in "IEEE Transaction on Speech and Audio processing", vol. 5, n° 3, 1997.
- [54] C. PINHANEZ. *The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces*, in "Proceedings of Ubiquitous Computing 2001 Conference", September 2001.
- [55] R. RASKAR. *iLamps: Geometrically Aware and Self-Configuring Projectors*, in "ACM SIGGRAPH 2003 Conference Proceedings", 2003.
- [56] R. RASKAR, G. WELCH, M. CUTTS, A. LAKE, L. STESIN, H. FUCHS. *The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays*, in "Proceedings of the ACM SIGGRAPH'98 Conference", 1998.
- [57] J. RASURE, S. KUBICA. *The Khoros Application Development Environment*, in "Experimental Environments for Computer Vision and Image Processing", J. CROWLEY, H. CHRISTENSEN (editors), Machine Perception Artificial Intelligence Series, vol. 11, n° 1, World Scientific Press, 1994, p. 1-32.

- 
- [58] B. SCHIELE, J. CROWLEY. *Recognition without Correspondence using Multidimensional Receptive Field Histograms*, in "International Journal of Computer Vision", vol. 36, n° 1, January 2000, p. 31–50.
- [59] M. SHAW, D. GARLAN. *Software Architecture: Perspectives on an Emerging Disciplines*, Prentice Hall, 1996.
- [60] N. A. STREITZ, J. GEISSLER, T. HOLMER, S. KONOMI, C. MÜLLER-TOMFELDE, W. REISCHL, P. REXROTH, P. SEITZ, R. STEINMETZ. *i-LAND: An interactive Landscape for Creativity and Innovation*, in "ACM Conference on Human Factors in Computing Systems", 1999.
- [61] J. U. B. ULLMER, H. ISHII. *Emancipated Pixels: Real-World Graphics in the Luminous Room*, in "Proceedings of ACM SIGGRAPH", 1999, p. 385-392.
- [62] D. VAUFREYDAZ. *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*, Ph.D. thesis in Computer Sciences, University Joseph Fourier, Grenoble (France), January 2002.
- [63] F. VERNIER, N. LESH, C. SHEN. *Visualization Techniques for Circular Tabletop Interfaces*, in "Advanced Visual Interfaces", 2002.
- [64] S. VOIDA, E. MYNATT, B. MACINTYRE, G. CORSO. *Integrating virtual and physical context to support knowledge workers*, in "Proceedings of Pervasive Computing Conference", IEEE Computer Society Press, 2002.
- [65] M. WEBER. *Frontal face dataset*, 2003, <http://www.vision.caltech.edu/html-files/archive.html>, internet.
- [66] R. YOUNG. *The Gaussian Derivative Theory for Spatial Vision: Analysis of Cortical Cell Receptive Field Line-Weighting Profiles*, Technical report, General Motors Research Laboratories, May 1985.
- [67] B. ZOPPIS. *Outils pour l'Intégration et le Contrôle en Vision et Robotique Mobile*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, June 1997.
- [68] L. VAN VLIET, I. YOUNG, P. VERBEEK. *Recursive Gaussian Derivative Filters*, in "International Conference on Pattern Recognition", August 1998, p. 509–514.