INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team PRIMA

Perception, recognition and integration for interactive environments

Rhône-Alpes

THEME COG

Table of contents

1. Team

Head of the team
James L. Crowley [Professor INPG]

Professors
Augustin Lux [Professor INPG]
Patrick Reignier [Assistant professor UJF]
Dominique Vaufreydaz [Assistant professor UPMF]

Team assistant
Natacha Laugier
Caroline Ouari

Expert Engineers
Daniela Hall
Alba Ferrer-Biosca
Sebastien Pesnel
Alban Caporossi
Jean-Marie Vallet

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 [AMN, bourse MENRT]
Nicolas Gourier [Bourse INRIA]
Julien Letessier [Bourse INRIA]
Jerome Maisonnasse [INPG SA - Contrat France Telecom]
Oliver Brdiczka [Bourse INRIA]
Sofia Zaidenberg [BDI CNRS]
Remi Emonet [Bourse MENRT]

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. Robust view-invariant Computer Vision

Keywords: *Affine Invariance*, *Local Appearance*, *Receptive Fields*.

3.1.1. Summary

A long-term grand challenge in computer vision has been to develop a descriptor for image information that can be reliably used for a wide variety of computer vision tasks. Such a descriptor must capture the information in an image in a manner that is robust to changes the relative position of the camera as well as the position, pattern and spectrum of illumination.

Members of PRIMA have a long history of innovation in this area, with important results in the area of Multi-resolution pyramids, scale invariant image description, appearance based object recognition and receptive field histograms published during the period 1987 to 2000. During the period 2002 - 2004, the group has demonstrated several innovations in this area.

Key results in this area include

- 1) Fast, video rate, calculation of scale and orientation for image description with normalized chromatic receptive fields [42].
- 2) Real time indexing and recognition using a novel indexing tree to represent multi-dimensional receptive field histograms [60].
- 3) Robust visual features for face tracking [46], [45].
- 4) Affine invariant detection and tracking using natural interest lines [67].
- 5) Direct computation of time to collision over the entire visual field using rate of change of intrinsic scale [55].

3.1.2. Detailed Description

The visual appearance of a neighbourhood can be described by a local Taylor series. The coefficients of this series constitute a feature vector that compactly represents the neighbourhood appearance for indexing and matching. The set of possible local image neighbourhoods that project to the same feature vector are referred to as the "Local Jet". A key problem in computing the local jet is determining the scale at which to evaluate the image derivatives.

Lindeberg Lindeberg [49] has described scale invariant features based on profiles of Gaussian derivatives across scales. In particular, the profile of the Laplacian, evaluated over a range of scales at an image point, provides a local description that is "equi-variant" to changes in scale. Equi-variance means that the feature vector translates exactly with scale and can thus be used to track, index, match and recognize structures in the presence of changes in scale.

A receptive field is a local function G(s,x,y) defined over a region of an image [64]. We employ a set of receptive fields based on derivatives of the Gaussian functions as a basis for describing the local appearance. These functions resemble the receptive fields observed in the visual cortex of mammals. These receptive fields are applied to color images in which we have separated the chrominance and luminance components. Such functions are easily normalized to an intrinsic scale using the maximum of the Laplacian [49], and normalized in orientation using direction of the first derivatives [64].

The product of a Laplacian operator with the image that is a local maxima in x and y and scale of the Laplacian provides a "Natural interest point" [50]. Such natural interest points are salient points that may be robustly detected and used for matching. A problem with this approach is that the computational cost of determining intrinsic scale at each image position can potentially make real-time implementation unfeasible. We have recently achieved video rate calculation of intrinsic scale by interpolating pixels within a Binomial Pyramid computing using an O(N) algorithm [42].

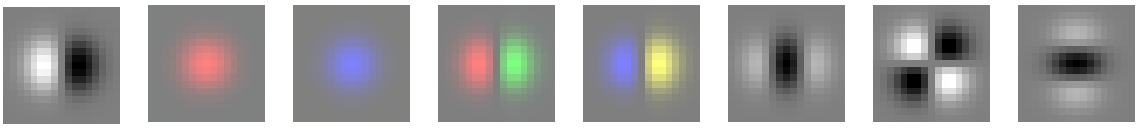

Figure 1. Chromatic Gaussian Receptive Fields ($G_x^L, G^{C_1}, G^{C_2}, G_x^{C_1}, G_x^{C_2}, G_{xx}^L, G_{xy}^L, G_{yy}^L$).

A vector of scale and orientation normalized Gaussian derivatives provides a characteristic vector for matching and indexing. The oriented Gaussian derivatives can easily be synthesized using the "steerability property" [43] of Gaussian derivatives. The problem is to determine the appropriate orientation. In earlier work by PRIMA members Colin de Verdiere [40], Schiele [64] and Hall [47], proposed normalising the local jet independently at each pixel to the direction of the first derivatives calculated at the intrinsic scale. This has provided promising results for many view invariant image recognition tasks. In particularly, this method has been used in the real time BrandDetect system described below. Lux and Hai have very recently developed an alternative method with provides a direct measurement of affine invariant local features based on extending natural interest points to "natural interest lines" [67]. The orientation of natural interest lines provides a local orientation in the region of an image structure. Early results indicate an important gain in discrimination rates.

Color is a powerful discriminator 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. 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 components C1 and C2 encodes the chromatic information in a Cartesian representation, while L is the luminance direction. Chromatic Gaussian receptive fields are computed by applying the Gaussian derivatives

independently to each of the three components, (L, C1, C2). Permutations of RGB lead to different opponent color spaces. The choice of the most appropriate space depends on the chromatic composition of the scene. An example of a second order steerable chromatic basis is the set of color opponent filters shown in figure 1.

3.2. Robust architectures for multi-modal perception

Keywords: *Autonomic Computing, MultiModal Perception, Process Architectures, Robust Perceptual Components.*

3.2.1. Summary

Machine perception is notoriously unreliable. Even in controlled laboratory conditions, programs for speech recognition or computer vision generally require supervision by highly trained engineers. Practical real-world use of machine perception requires fundamental progress in the way perceptual components are designed and implemented. A theoretical foundation for robust design can dramatically reduce the cost of implementing new services, both by reducing the cost of building components, and more importantly, by reducing the obscure, unpredictable behaviour that unreliable components can create in highly complex systems. To meet this challenge, we propose to adapt recent progress in autonomic computing to the problem of producing reliable, robust perceptual components.

Autonomic computing has emerged as an effort inspired by biological systems to render computing systems robust. Such systems monitor their environment and internal state in order to adapt to changes in resource availability and service requirements. Monitoring can have a variety of forms and raises a spectrum of problems.

An important form of monitoring relies on a description of the system architecture in terms of software components and their interconnection. Such a model provides the basis for collecting and integrating information from components about current reliability, in order to detect and respond to failure or degradation in a component or changes in resource availability (auto-configuration). However, automatic configuration, itself, imposes constraints on the way components are designed, as well as requirements on the design of the overall system [44].

Robust software design begins with the design of components. The PRIMA project has developed an autonomic software architecture as a foundation for robust perceptual components. This architecture allows experimental design with components exhibiting:

- Auto-criticism: Every computational result produced by a component is accompanied by an estimate of its reliability.
- Auto-regulation: The component regulates its internal parameters so as to satisfy a quality requirement such as reliability, precision, rapidity, or throughput.
- Auto-description: The component can provide a symbolic description of its own functionality, state, and parameters.
- Auto-Monitoring: the component can provide a report on its internal state in the form of a set of quality metrics such as throughput and load.
- Auto-configuration: The component reconfigures its own modules so as to respond to changes in the operating environment or quality requirements [59].

Maintenance of such autonomic properties can result in additional computing overhead within components, but can pay back important dividends in system reliability.

3.2.2. Detailed Description

As a general model, we propose a layered architecture for context aware services.At the lowest level, the service's view of the world is provided by a collection of physical sensors and actuators. This corresponds to the Sensor-actuator layer. This layer, which is dependent on the technology, encapsulates the diversity of sensors and actuators that may provide similar function. It provides logical interfaces, or standard API's that are function centered and device independent.

Services interact with humans through sensors and actuators. For performance reasons, the services may skip the situation layer and exchange information directly with perception-action or sensor and actuator layers. However, service provides semantics for interpretation. Hard-wiring service behavior to sensor signals can provide only simplistic services that are hardware dependent and of limited utility. Hardware independence and generality require abstract perception and abstract task specification. Perception interprets sensor signals by recognizing and observing entities. Abstract tasks are expressed in terms of a desired result rather than a blind action. Perception and action operate in terms of environmental state while sensors and actuators operate on device-specific signals.

The PRIMA software architecture for supervised autonomic perceptual components is shown in figure 2. Components are constructed as a cyclic detection and tracking process monitored and controlled by an autonomic supervisor.

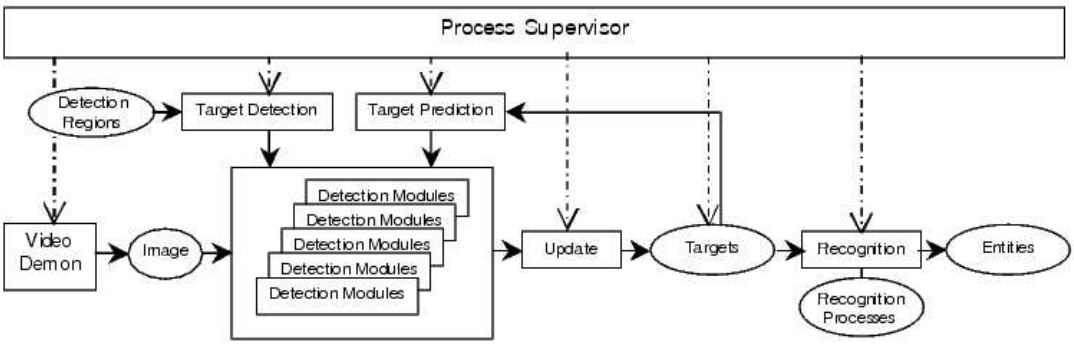

Figure 2. Architecture for an autonomic perceptual component

The supervisory controller provides five fundamental functions: command interpretation, execution scheduling, event handling, 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 handles event dispatching to other processes, and reacts to events from other processes. 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.

Real-time visual processing for the perceptual component is provided by tracking. Tracking conserves information about over time, thus provides object constancy. Object constancy assures that a label applied to a blob at time T1 can be used at time T2. Tracking enables the system focus attention, applying the appropriate detection processes only to the region of an image where a target is likely to be detected. Also the information about position and speed provided by tracking can be very important for describing situations.

Tracking is classically composed of four phases: Predict, observe, detect, and update. The prediction phase updates the previously estimated attributes for a set of entities to a value predicted for a specified time. The

observation phase applies the prediction to the current data to update the state of each target. The detect phase detects new targets. The update phase updates the list of targets to account for new and lost targets. The ability to execute different image processing procedures to process target information with an individual ROI is useful to simultaneously observe a variety of entities.

The PRIMA perceptual component architecture adds additional phases for interpretation, auto-regulation, and communication. In the interpretation phase, the tracker executes procedures that have been downloaded to the process by a configuration tool. These are interpreted by a RAVI interpreter [51] and may result in the generation of events or the output to a stream. The auto-regulation phase determines the quality of service metric, such as total cycle time and adapts the list of targets as well as the target parameters to maintain a desired quality. During the communication phase, the supervisor responds to requests from other processes. These requests may ask for descriptions of process state, or capabilities, or may provide specification of new recognition methods.

Homeostasis, or "autonomic regulation of internal state" is a fundamental property for robust operation in an uncontrolled environment. 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 component supervisor maintains homeostasis by adapting module parameters using the auto-critical reports from modules.

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 supervisor. Such descriptions are useful for both manual and automatic composition of federations of processes.

In the context of recent National projects (RNTL ContAct) and European Projects (FAME, CAVIAR, CHIL), the PRIMA perceptual component has been demonstrated with the construction of perceptual components for

- 1) Tracking individuals and groups in large areas to provide services,
- 2) Monitoring a parking lot to assist in navigation for an autonomous vehicle.
- 3) Observing participants in an meeting environment to automatically orient cameras
- 4) Observing faces of meeting participants to estimate gaze direction and interest.
- 5) Observing hands of meeting participants to detect 2-D and 3D gestures.

3.3. Context aware interactive environments

Keywords: *Context Aware Environments*, *Situation Modeling*, *Smart Environments*.

3.3.1. Summary

An environment is said to be "interactive" when it is capable of perceiving, acting, and communicating with its occupants. PRIMA explores multiple forms of interaction, including projected interaction widgets, observation of manipulation of objects, fusion of acoustic and visual information, and dynamic composition of systems that model interaction context in order to predict appropriate action by the environment.

The experiments in project PRIMA are oriented towards context aware observation of human activity. Over the last few years, the group has developed a technology for describing activity in terms of a network of situations. Such networks provide scripts of activities that tell a system what actions to expect from each individual and the appropriate behavior for the system. Current technology allows us to handcraft real-time systems for a specific service. The current hard challenge is to create a technology for automatically learning and adapting situation models with minimal or no disruption of users.

Over the last two years, PRIMA hase developed situation models based on the notion of a script. Atheatrical script provides more than dialog for actors. A script establishes abstract characters that provide actors with a space of activity for expression of emotion. It establishes a scene within which directors can layout a stage and place characters. Situation models are based on the same principle.

PRIMA has developed methods for defining situation models for driving context aware observation of activity. These methods have been applied to the develop of situation models for

- A context aware video acquisition system, used to record eight lectures of 3 hours at the Barcelona Forum of Cultures in July 2004
- Interpretation of interaction of groups in video surveillance
- Interpretation of activities in meetings.

3.3.2. Detailed Description

A script describes an activity in terms of a scene occupied by a set of actors and props. Each actor plays a role, thus defining a set of actions, including dialog, movement and emotional expressions. An audience understands the theatrical play by recognizing the roles played by characters. In a similar manner, a user service uses the situation model to understand the actions of users. However, a theatrical script is organised as a linear sequence of scenes, while human activity involves alternatives. In our approach, the situation model is not a linear sequence, but a network of possible situations, modeled as a directed graph.

Situation models are defined using roles and relations. A role is an abstract agent or object that enables an action or activity. Entities are bound to roles based on an acceptance test. This acceptance test can be seen as a form of discriminative recognition.

Currently situation models are constructed by hand. Our current challenge is to provide a technology by which situation models may be adapted and extended by explicit and implicit interaction with the user. An important aspect of taking services to the real world is an ability to adapt and extend service behaviour to accommodate individual preferences and interaction styles. Our approach is to adapt and extend an explicit model of user activity. While such adaptation requires feedback from users, it must avoid or at least minimize disruption.

The PRIMA group has refined its approach to context aware observation in the development of a process for real time production of a synchronized audio-visual stream based using multiple cameras, microphones and other information sources to observe meetings and lectures. This "context aware video acquisition system" is an automatic recording system that encompasses the roles of both the camera-man and the director. The system determines the target for each camera, and selects the most appropriate camera and microphone to record the current activity at each instant of time. Determining the most appropriate camera and microphone requires a model of activities of the actors, and an understanding of the video composition rules. The model of the activities of the actors is provided by a "situation model" as described above.

Version 1.0 of the video acquisition system was used to record 8 three-hour lectures in Barcelona in July 2004. Since that time, successive versions of the system have been used for recording testimonial’s at the FAME demo at the IST conference, at the Festival of Science in Grenoble in October 2004, and as part of the final integrated system for the national RNTL ContAct project. In addition to these public demonstrations, the system has been in frequent demand for recording local lectures and seminars. In most cases, these installations made use of a limited number of video sources, primarily switching between a lecturer, his slides and the audience based on speech activity and slide changes. Such actual use has allowed us to gradually improve system reliability. Version 2.0, released in December 2004, incorporated a number of innovations, including 3D tracking of the lecturer and detection of face orientation and pointing gestures. This version is currently used to record the InTech lecture series a the INRIA amphitheater. Discussions are underway to commercialise this technology with the creation of a start up company.

3.4. New forms of man-machine interaction based on perception

Keywords: *Augmented Reality, Projected interaction widgets, Steerable Camera Projector*.

Surfaces are pervasive and play a predominant role in human perception of the environment. Augmenting surfaces with projected information provides an easy-to-use interaction modality that can easily be adopted for a variety of tasks. Projection is an ecological (non-intrusive) way of augmenting the environment. Ordinary

objects such as walls, shelves, and cups may become physical supports for virtual functionalities [58]. The original functionality of the objects does not change, only its appearance. An example of object enhancement is presented in [32], where users can interact with both physical and virtual ink on a projection-augmented whiteboard.

Combinations of a camera and a video projector on a steerable assembly [33] are increasingly used in augmented environment systems [57] [62] as an inexpensive means of making projected images interactive. Steerable projectors [33] [58] provide an attractive solution overcoming the limited flexibility in creating interaction spaces of standard rigid video-projectors (e.g. by moving sub windows within the cone of projection in a small projection area [70]).

The PRIMA group has recently constructed a new form of interaction device based on a Steerable Camera-Projector (SCP) assembly. This device allows experiments with multiple interactive surfaces in both meeting and office environments. The SCP pair, shown in figure 3, is a device with two mechanical degrees of freedom, pan and tilt, mounted in such a way that the projected beam overlaps with the camera view. This creates a powerful actuator-sensor pair enabling observation of user actions within the camera field of view. This approach has been validated by a number of research projects as the DigitalDesk [72], the Magic Table [32] or the Tele-Graffiti application [66].

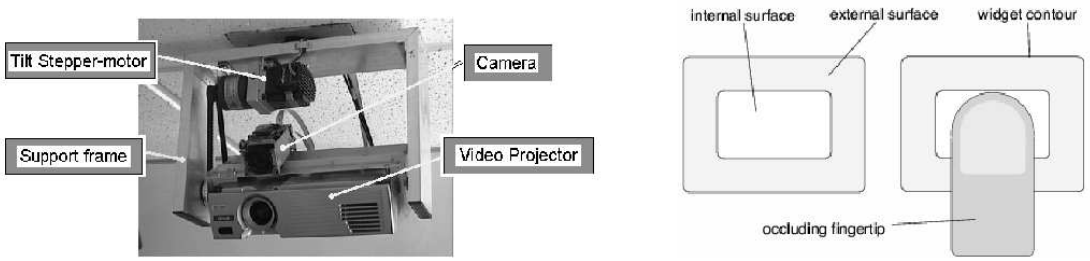

Figure 3. Steerable camera-projector pair (left) and surfaces defined to detect touch-like gestures over a widget (right)

For the user interaction, we are experimenting with interaction widgets that detect fingers dwelling over button-style UI elements, as shown to the right in figure 3.

4. Application Domains

4.1. The Augmented Meeting Environment

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

Participants: Patrick Reignier, Dominique Vaufreydaz, James L. Crowley, Oliver Brdiczka, Sofia Zaidenberg, Jerome Maisonnasse, Suphot Chunwiphat.

In order to test and develop systems for observation of human activity, Project PRIMA has constructed an "Augmented Meeting Environment", show in figure 4. 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.

Figure 4. 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.

Video interaction devices associate a camera with a video projector to provide new modes if 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, 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, Julien Letessier, Alban Caporossi, James L. Crowley.

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 proved 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. 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 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 [48], [65], [71]. 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 [68]. 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 an 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. 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 [54] and [58]. 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 [58] 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 [39], where users could interact with both physical and virtual ink on an projection-augmented whiteboard.

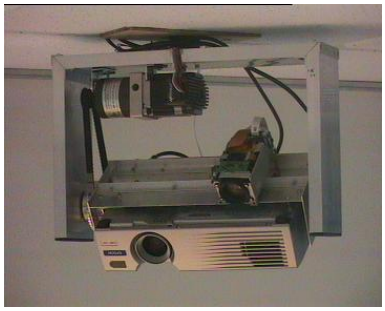

Figure 5. The Steerable Camera Projector

The Steerable Camera Projector (SCP) (figure 5) 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 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 [61].

Table 1. Rotation platform mechanical performance

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

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 6 shows automatically discovered planar surfaces within the AME. described below.

4.3. Context Aware Video Acquisition

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

Participants: Patrick Reignier, Dominique Vaufreydaz, 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

Figure 6. Planar surfaces in the environment

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 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 [41]. 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 for the federation supervisors has been developed and refined within the last two years.

In 2004, we have demonstrated a number of reals 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 "FÍte 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, Alban Caporossi, Daniela Hall, Thi-Thanh-Hai Tran, Remi Emonet.

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 [63] 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 [74]. The particular strength of the Ravi kernel comes from a combination of dynamic loading and automatic program generation within an interactive shell. This makes it possible to integrate external programs, up to the size of an entire library, in a completely automatic way.

Work on the Imalab system during the period 2004/05 mainly has concerned a new architecture of the main program, based on high-level system components (e.g. image processing machine, GUI)

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 Pyry Consulting, Helsinki, Finland
- Université de Liège, Belgium
- France Télécom R&D, Meylan France

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.

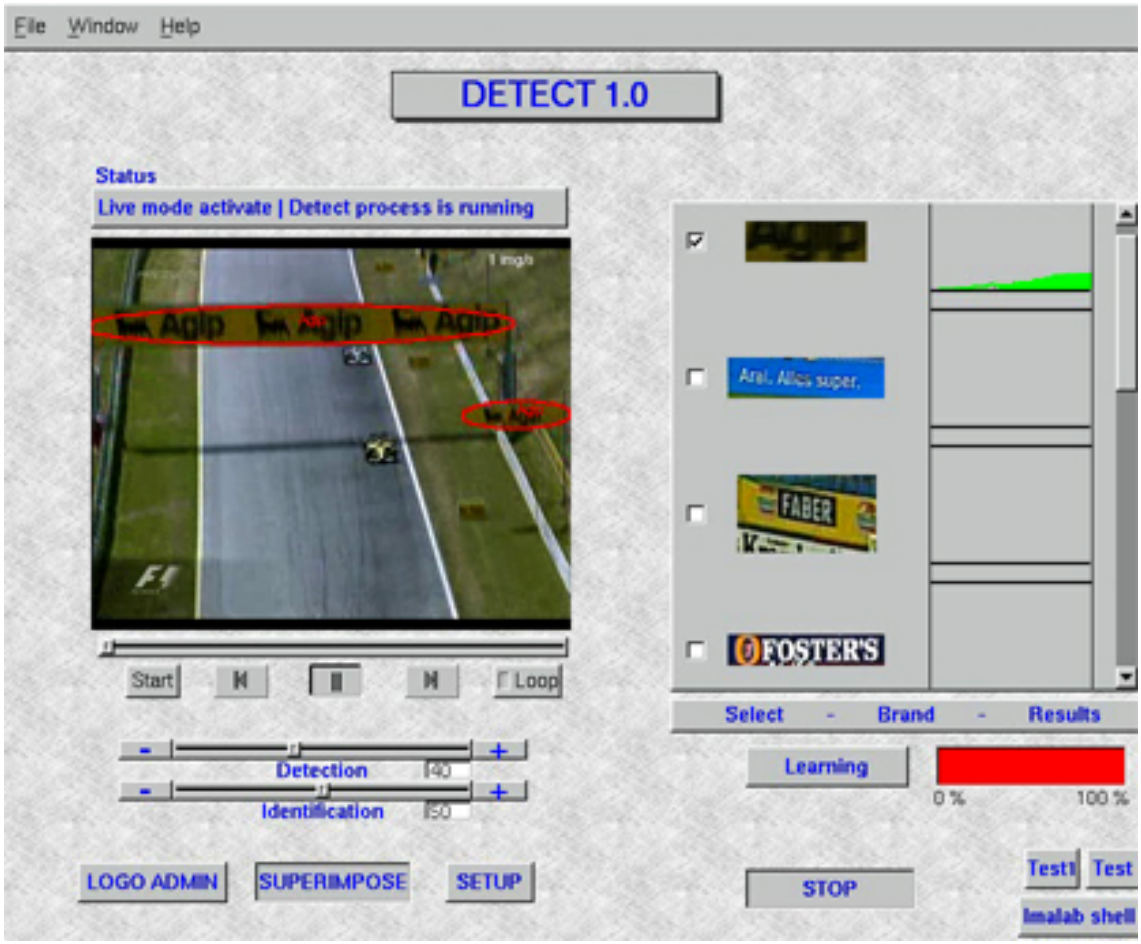

Figure 7. BrandDetect collects statistics on appearance of publicity panels in Broadcast video

In the period 2002 to 2004, in the context of European IST project DETECT, PRIMA has developed a software for detection, tracking and recognition of publicity in broadcast video of sports events. The result is a commercial software system named BrandDetect.

BandDetect 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 7.

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 for commercial exploitation has been negotiated with the Austrian company HSArt.

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.

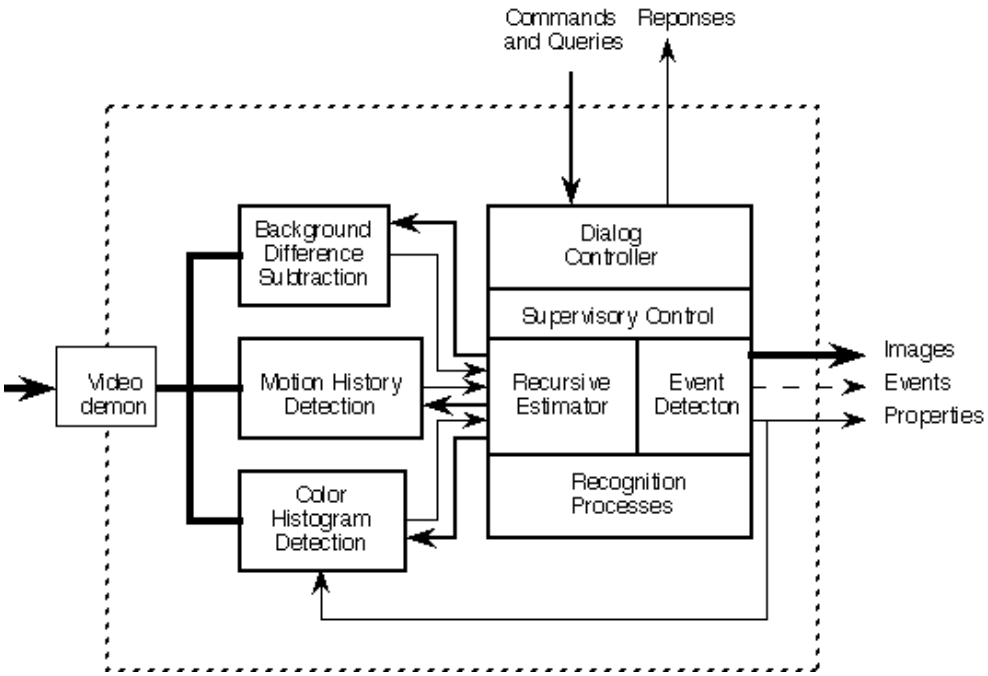

Figure 8. The CAR systems integrates several detection modules with a Kalman Filter for robust detection and tracking of entities

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.

5.4. PRIMA Automatic Audio-Visual Recording System

Keywords: *audio-visual recording system.*

Participants: Patrick Reignier, Dominique Vaufreydaz, Alba Ferrer-Biosca.

The PRIMA automatic audio-visual recording system controls a battery of cameras and microphones to record and transmit the most relevant audio and video events in a meeting or lecture. The system uses a can employ both steerable and fixed cameras, as well as a variety of microphones to record synchronized audio-video streams. Steerable cameras automatically oriented and zoomed to record faces, gestures or documents. At each moment the most appropriate camera and microphone is automatically selected for recording. System behaviour is specified by a context model. This model, and the resulting system behaviour, can be easily edited using a graphical user interface.

In video-conferencing mode, this system can be used to support collaborative interaction of geographically distributed groups of individuals. In this mode, the system records a streaming video, selecting the most appropriate camera and microphone to record speaking inviduals, workspaces, recorded documents, or an entire group. In meeting minute mode, the system records a audio-visual record of "who" said "what".

The system is appropriate for business, academic and governmental organizations in which geographically remote groups must collaborate, or in which important meetings are to be recorded for future reference.

The primary innovative features are: 1) Dynamic real time detection and tracking of individuals and workspace tools 2) Dynamic real time 3D modeling of the scene layout. 3) Dynamic recognition and modeling of human activity using stereotypical graphs of situations.

This system can dramatically improve the effectiveness of group video conferencing. It can eliminate the need for human camera operators and editors for recording public meetings. The product reduces energy consumption by allowing video conferencing to serve as an effective alternative to airline travel.

Market for this system is determined by 1) The number of "business meetings" between remote collaborators 2) The number of important meetings for which an audio-visual record should be maintained.

Rights to this system are jointly owned by INP Grenoble, UJF and INRIA. The system is currently undergoing Deposit at APP and will shortly be available for licensing. We are investigating plans to create a start up to commercially exploit this system.

5.5. APTE: Automatic Parameter Tuning and Error Recovery

Keywords: *Monitoring, Process optimization, Self-adaptive system, Video Surveillance*.

Participants: Daniela Hall, Remi Emonet.

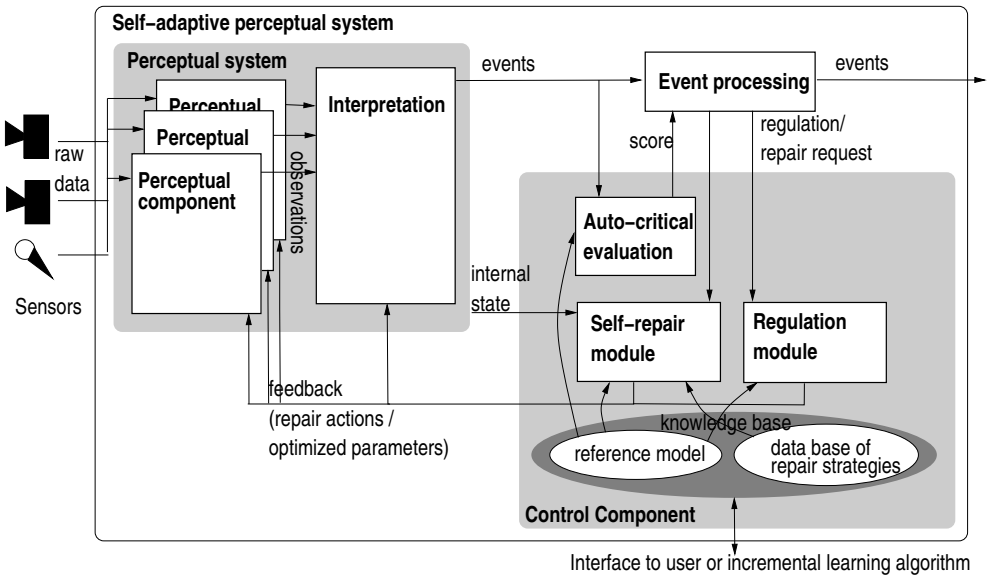

Figure 9. The APTE component endows a perceptual system with self-adaptive capabilities such as parameter optimisation and error recovery.

Perceptual systems observe the world, interpret the observations in form of input signals such as images, audio data, or laser range data and communicates the result as numeric, vectorial or symbolic events. There is a wide range of different perceptual systems such as tracking systems in video surveillance, expert systems in traffic control or image segmentation systems. Stability, reliability and robustness are required for perceptual systems to be exploited widely in commercial applications. To meet these constraints, developers commonly design simple systems whose parameters can be adapted manually. Such systems perform well as long as the environment stays constant. Unfortunately, in most real applications the environmental conditions perceived by the sensors frequently change, which often breaks the system and requires reinitialisation and new hand tuning of the parameters. This software provides a solution to this problem by enabling a system to automatically adapt its parameters to the environmental changes that would degrade the system performance.

Some error types can not be solved by parameter tuning. Examples are local illumination changes and person fragmentation. To cope with such kinds of errors, the software contains a system for automatic error detection, error classification and error repair. Error detection is performed by evaluating a probabilistic measure with respect to a model of “correct” system output. Classification requires an (incremental) acquisition of error classes. This acquisition requires a minimum amount of human supervision. The system provides a GUI that allows to manually process previously unclassified errors. The knowledge is incrementally incorporated in the system’s knowledge base. First tests showed that manual interaction of 15 min gives a usable error classification model.

This software has been filed with the APP "Agence pour la Protection des Programmes" under the Interdeposit Digital number IDDN.FR.001.480025.000.S.P.2005.000.10000.

6. New Results

6.1. Autonomic architecture for multi-modal tracking

Keywords: *Autonomic Perception Systems, Computer Vision Systems, Robust Tracking.*

Participants: Sebastien Pesnel, Daniela Hall, Remi Emonet, James L. Crowley.

Remi Emonet

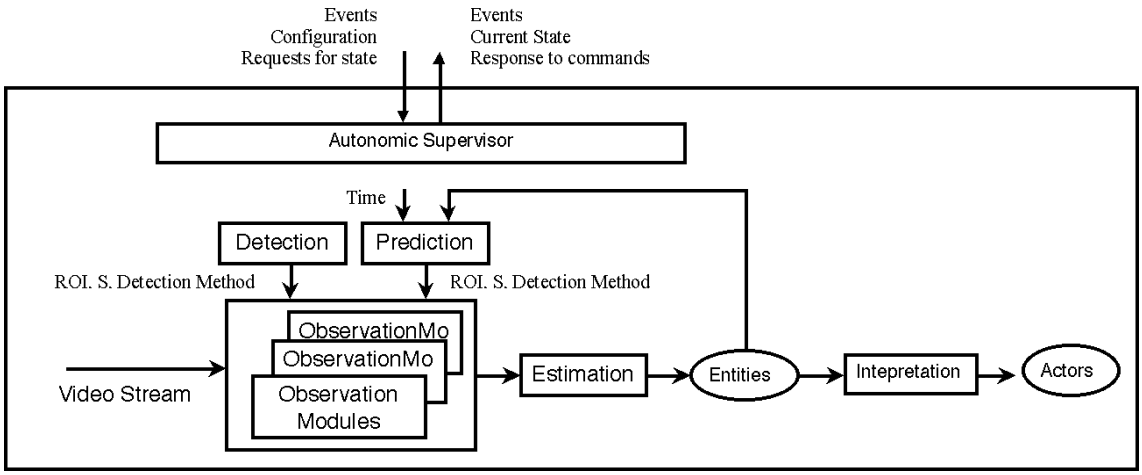

Figure 10. The components and architecture for an autonomic perceptual process. (Robust Tracker version 3.3)

As described above, PRIMA has filed an APP for a robust real time detection and tracking system under that name CAR. This system has been programmed in C++ using hard-coded control logic. Experience with robust tracking in a variety of applications has illustrated the importance of a technology for auto-configuration and auto-regulation for perceptual systems. In order to demonstrate and explore such system, PRIMA has implemented an autonomic perceptual component based on robust tracking.

The robust uses an autonomic supervisor to provide 1) Automatic parameter regulation 2) Automatic reconfiguration in order to maintain quality of service 3) Self description for capabilities and stated.

The system can operated within the imalab experimental environment, or can be compiled to run stand alone. This system has been registered with the APP as Prima Robust Tracker vesion 3.3.

6.2. Creation of the company Blue Eye Video

In 2003, with the assistance by INRIA Transfert and the GRAIN, the PRIMA group 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. Since mid 2004, the number of system installed by Blue Eye Video has been doubling roughly every 6 months. As of September 2005, Blue Eye Video has 9 employees, over 100 systems installed and a capital reserve of roughly 400 000 Euros.

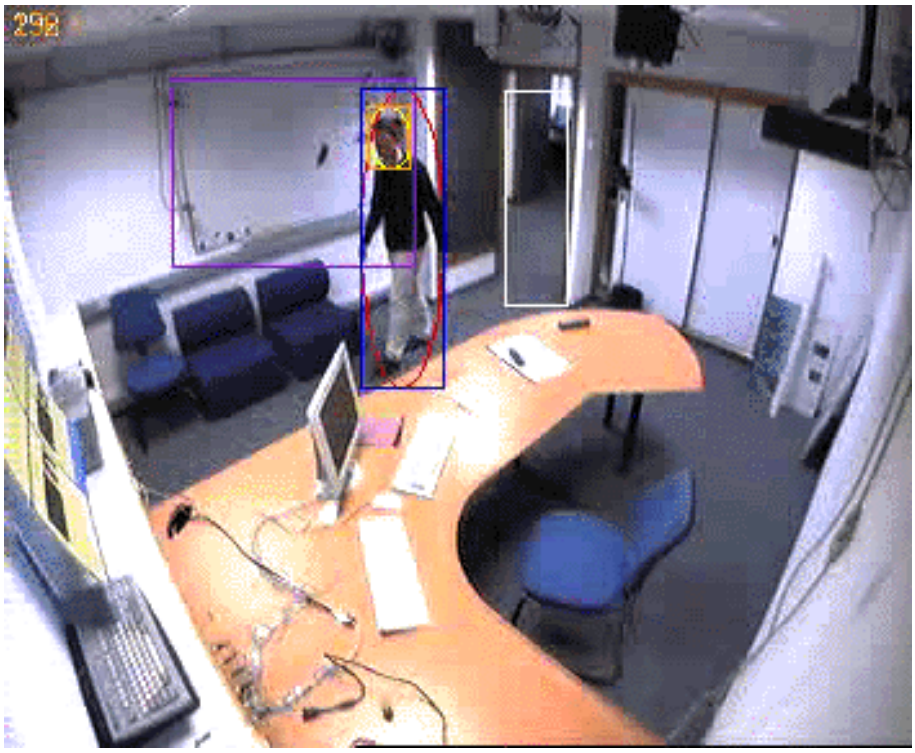

Figure 11. The new programmable robust tracker makes it possible to observe composite entities

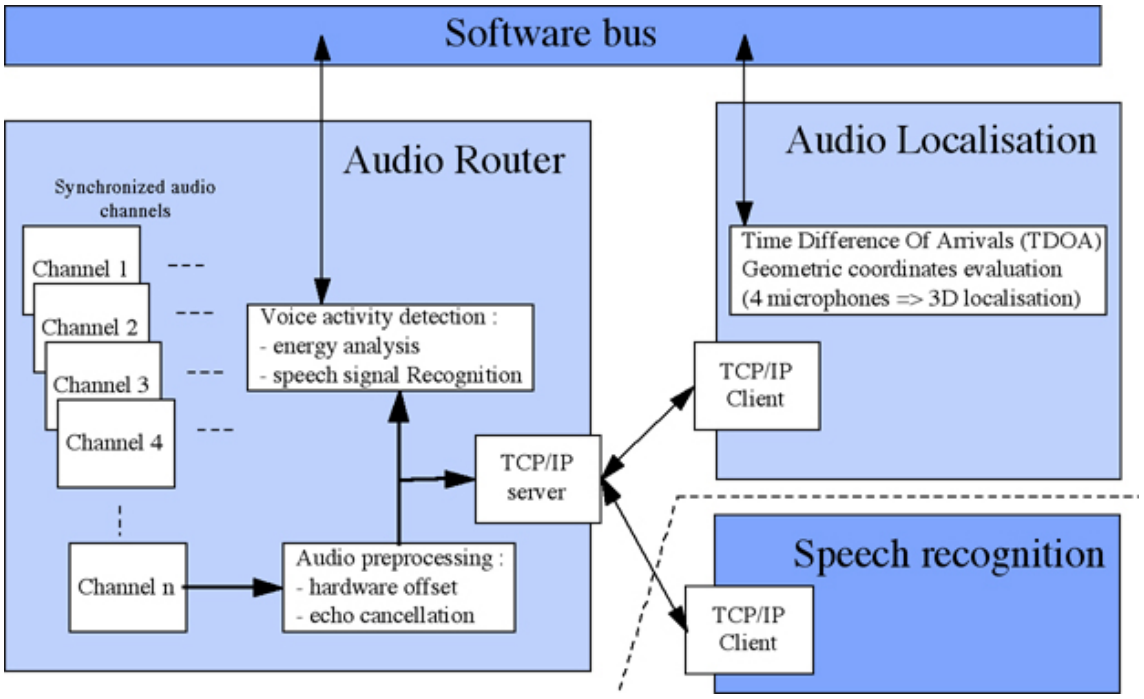

Figure 12. Processes for detection, recognition and tracking of Acoustic Sources

6.3. 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) [56]. 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) [69].

6.4. The Process Federation Tool

Keywords: *Distributed Computing, Middleware, Process Federations.*

Participants: Patrick Reignier, Dominique Vaufreydaz, Sebastien Pesnel, Alban Caporossi, Daniela Hall, Julien Letessier, Remi Emonet, James L. Crowley.

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. In order to experiment with assemblies of processes, we have constructed a middle-ware environment named "BIP".

BIP stands for Basic Interaction Protocol. BIP is designed to allow low-latency dynamic discovery and assembly of processes for assembling federations of perceptual processes. Perceptual processes and components advertise their abilities to provide services by publishing ontological descriptions in a Multi-cast Dynamic Name Service (DNS) Service Discovery (SD). BIP uses the Apple Rendezvous (DNS-SD) open standard to allow processes to describe and publish their perceptual abilities as software services.

In addition to facilities for publishing and discovery of services offered by perceptual processes, BIP provides primitives for systems level communication and control of software federations based on procedures for event flow services. BIP services exist for both JAVA and C++ and may be used to assemble federations of processes running under Linux, Win32 and Mac OSX.

Processes advertise their abilities by publishing a 64-byte event containing an IP address and port. Perceptual services and data structures are described using an XML protocol. An ontology server is provided organized as an inheritance hierarchy for software and perceptual service classes. This ontology can be searched using XPath to discover compatible services.

BIP is designed to meet a number of requirements encountered when constructing real time perceptual systems: low latency communication, reliability, function reuse and distribution.

6.4.1. 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.

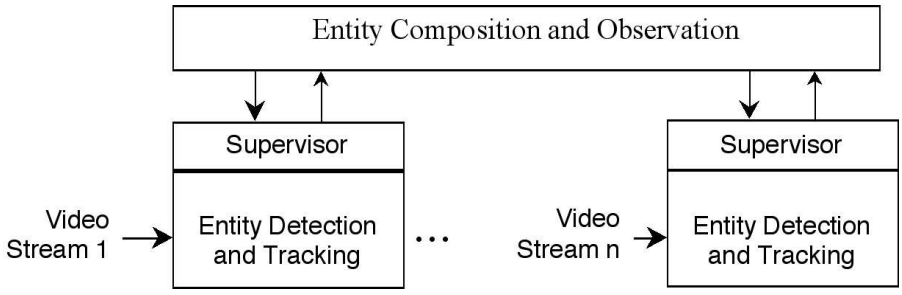

Figure 13. A simple process federation composed of entity detection process, and a composition process

6.5. 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.

6.6. Context model compiler

Participant: Patrick Reignier.

PRIMA has constructed a graphical interaction tool for designing situation graphs 14. 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.

6.6.1. 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 [31] : *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.

6.6.2. Synchronized Petri Nets

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

- The preconditions on places marks are verified.

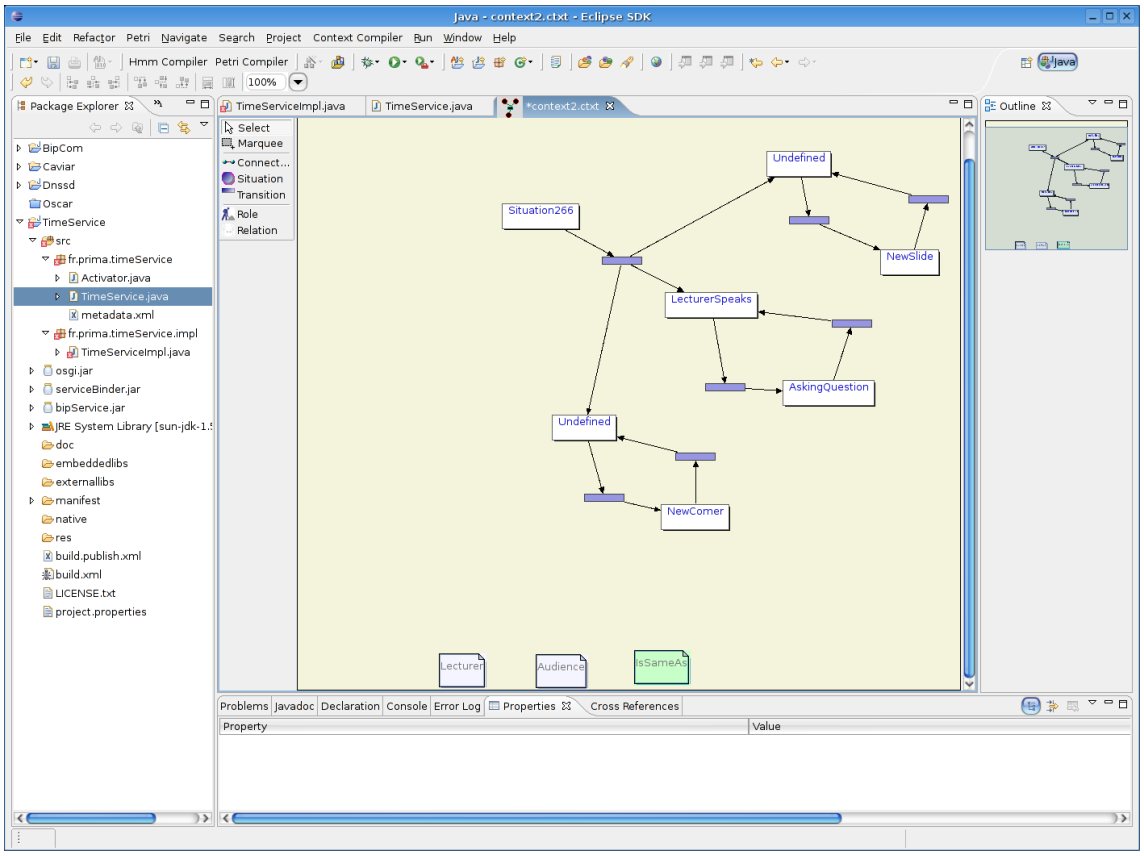

Figure 14. Screen copy of the context modelling tool

- 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.

6.6.3. 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 [30] 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 immplementing 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.

6.7. Automatic Acquisition of Context models

Participants: Patrick Reignier, Oliver Brdiczka, Sofia Zaidenberg, James L. Crowley.

James L. Crowley

Context-aware systems must adapt and develop while retaining continuity and stability for users. Adaptation is necessary to maintain consistent behavior while accommodating changes in the operating environment, task, user population, preferences, or some other factors. At the same time, context is too complex to be preprogrammed. A context model must develop through observation and interaction with users. Therefore, machine learning methods need to be used and evaluated.

We propose a first approach for adapting a given situation network representing context in [35] [36]. The approach takes user feedback on executed system services as input. The adaptations are based on the split of situations and the acquisition of new situation characteristics using decision trees. The results were good, adapting a given situation network to the needs of each particular user.

Further work concerned the non-supervised acquisition of situation networks. We represent the situation network by a HMM, each state of which corresponds to a particular situation. We propose a method based on an incremental algorithm creating a set of situation networks (represented by a set of HMM) for given sequences of perceptions. These perceptions are entities, their roles and relations describing human behavior in an environment. The method has been tested on the CAVIAR data sets shows promising results [73].

6.8. Interaction group detection for addressing services

Participants: Patrick Reignier, Oliver Brdiczka, Nicolas Gourier, Jerome Maisonnasse, James L. Crowley.

Patrick Reignier, James L. Crowley

In order to enable computer systems to sense and to respond to human activity, human actors need to be identified. In intelligent environments more and more devices are capable of perceiving user activity and offering services to the user. Offering services means to supply a system reaction or an interaction at the most appropriate moment, aligned with the activity of the users. Addressing the right user at the correct moment is essential. Thus we need to detect potential users and their connection while doing an activity.

The identification of the current configuration of interaction groups is necessary to analyze activity. In a physical environment, several individuals can form one group working on the same task, or they can split into subgroups doing independent tasks in parallel. The dynamics of group configuration, i.e. the split and merge of small interaction groups, allows us to perceive the appearing of new activities. We assume that a change in group configuration is strongly linked to a change in activity, at least to an interruption of the current activity. The fusion of several independent small groups is seen as important information for detecting a change of the current activity, on a local or global level.

To detect interaction groups, we take speech activity detection of individuals forming interaction groups as input. The approach is based on the assumption that conversational turn taking is synchronized inside groups. We propose two different real-time detectors constructed upon conversational hypotheses. One algorithmic detector and one detector based on HMM. Both detectors show good results [34], [52] and thus confirm our conversational hypotheses.

To detect different interaction groups and activities, conversational hypotheses need to be specified for each possible group configuration and activity. For many individuals, specifying hypotheses for all possible group configurations may be difficult or even impossible. We thus propose an unsupervised method for segmenting speech activity detection distributions in meetings. The method is based on detecting peaks of the Jeffrey divergence curve between sliding-window histograms. The methods shows promising results [37] [38], providing segments of (unknown) group configurations and activities that may be, in a second step, labeled by a supervisor.

Further work concerns an attentional model providing a priori information about the focus of attention of interacting individuals. Speech activity detection does not provide sufficient information about interactions and attention of people in a realistic environment. A multimodal approach needs to be envisaged. To provide a generic framework for a priori representation of attentional focus taking multimodal sensor information as input, we propose a model based on weighting perceptive salience of contextual objects [53].

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
- UniversitÈ Joseph Fourier, Laboratoire CLIPS, France, Prof. JoÎlle Coutaz
- Istituto Trentino di Cultura, Italy, Marcello Federico
- Universitat PolitËcnica de Catalunya Centre TALP, Spain, Prof. JosÈ B. MariÒo
- Sony International (Europe) GmbH, Germany, Ralf Kompe
- Applied Technologies on Language and Speech S. L., Germany, David Font

The goal of IST Project FAME has been to construct a context aware multi-modal system to facilitate communication among people from different cultures who collaborate on solving a common problem. This system provided 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 system serves as an information butler to aid multicultural communication in a transparent way. The system does not intervene in the conversation, but will remain in the background to provide the appropriate support. A public demonstration has been run at Barcelona Forum of Cultures for two weeks during July 2004.

7.1.2. IST 2001 37540 CAVIAR: Context Aware Vision using Image-based Active Recognition

European Commission project: IST 2001 37540
Starting Date: October 1, 2002 Finishing Date: Sept 30 2005 Duration: 36 Months
Key Action: Cognitive Vision
Consortium:

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

CAVIAR has addressed 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 studied were:

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. Behaviour of potential customers in a commercial settings. Marketing experts are in interested in the sequence of locations visited by potential customers, 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.

7.1.3. 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 fr Informations- und Datenverabeitung, Karlsruhe, Germany
- Universitt 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
- Insitut National Polytechnique de Grenoble, France
- Insituto Trentino di Cultura, Trento, Italy
- Kungl Tekniska Hgskolan (KTH), Stockholm, Sweden
- Centre National de la Recherche Scientifique, Orsay, France
- Technische Universiteit Eindhoven, Eindhoven, Netherlands
- Universitt 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 uses 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.

7.1.4. RNTL/Proact: ContAct Context management for pro-Active computing

Start Date February 2003.
Duration: 36 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 Pyry Consulting, Helsinki, Finland
- Ellipse, Helsinki, Finland

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

The aim of Project RNTL CONTACT was to explore novel approaches 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.

To achieve these results project CONTACT has included 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.

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

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 was a thematic network devoted to Cognitive Enabled Computer Vision Systems. ECVision served to unify the set of 8 IST projects funded in Framework V under the EC's Cognitive Vision program, including IST projects CAVIAR and DETECT.

The principal goal of ECVision has been 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 were 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 was coordinator of Research Planning for ECVision.

9. Dissemination

9.1. Contribution to the Scientific Community

9.1.1. Smart Objects and Ambient Intelligence, SOC-EUSAI '05

James L. Crowley was general chairman for the conference "Smart Objects and Ambient Intelligence", SOC-EUSAI, held in Grenoble in October 2005. Smart Objects and Ambient Intelligence provided a venue for the emerging multi-disciplinary community of researchers that work on Ambient Intelligence and smart objects. Ambient Intelligence represents a vision of the future where people are surrounded by sensitive and responsive electronic environments. Ambient intelligence technologies are expected to combine concepts of ubiquitous computing and intelligent systems putting humans in the center of technological developments.

9.1.2. ICMI 2005: International Conference on MultiModal Interaction

James L. Crowley was program co-chairman for the conference ICMI 2005: International Conference on MultiModal Interaction, held in Trento in October 2005. ICMI is the dominant conference for the field of multi–modal interaction, and rotates between North America, Europe and Asia. 1)

9.1.3. Participation on Conference Program Committees

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

- ICPR 2006, International Conference on Pattern Recognition, Hong Kong, Aug 2006
- CVPR 2006, IEEE International Conference on Computer Vision and Pattern Recognition, New York, June 2006
- ECCV 2006, European Conference on Computer Vision, Graz, Au, May 2006
- FG 2006, Automatic Face and Gesture Recognition, Southampton, UK, April 10-12 2006
- ICVS 2006, International Conference on Vision Systems, New York, January 2006
- ICCV 2005, IEEE International Conference on Computer Vision, Beijing, Oct 2005.
- IROS 2005, IEEE Conference on Intelligent Robotics and Systems, Juillet, 2005
- CVPR 2005, IEEE International Conference on Computer Vision and Pattern Recognition, 2005
- ICRA 2005, IEEE International Conference on Robotics and Automation, 2005

Daniela Hall served as a member of the program committee for the following conferences.

- HAREM 2005, International Workshop on human activity modelling and recognition, September 2005.
- VSPETS 2005, Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, October 2005
- ICPR 2006, International Conference on Pattern Recognition, Hong Kong, August 2006

9.1.4. Participation on Advisory Panels

During Novembre 2005, James L. Crowley has served on a proposal evaluation panel for the IST- FET Programme. During July 2005, James L. Crowley has served on a study panel to propose subjects for the EUY Framework Programme VI.

9.1.5. Invited Plenary Presentations at Conferences

- "Context Aware Observation of Human Activity", keynote Speaker at ASCI Symposium on Image Analysis, Boxmeer, Netherlands, June 2005.
- "Context Aware Observation of Human Activity", Invited Presentation at A*STAR Cognitive Science Symposium, Agency for Science, Technology and Research (A*STAR) of Singapore, Sept 2005.

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. COUTAZ, J. CROWLEY, S. DOBSON, D. GARLAN. *Contex is Key*, in "Communications of the ACM, Special issue on the Disappearing Computer", March 2005.

[3] 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.

[4] 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.

[5] J. CROWLEY, J. COUTAZ, G. REY, P. REIGNIER. *Using Context to Structure Perceptual Processes for Observing Activity*, in "UBICOMP, Sweden", September 2002.

[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] 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.

[8] C. LE GAL, J. MARTIN, A. LUX, J. L. CROWLEY. *Smart Office: An Intelligent Interactive Environment*, in "IEEE Intelligent Systems", July/August 2001.

[9] J. MAISONNASSE, N. GOURIER, O. BRDICZKA, P. REIGNIER. *Attentional Model for Perceiving Social Context in Intelligent Environments*, in "3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, Athens, Greece", to appear, June 2006.

[10] 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.

[11] 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 refereed journals and book chapters

[12] J. CROWLEY, P. REIGNIER, J. COUTAZ. *True Vision*, chap. Designing Context Aware Services for Ambient Informatics, 2005.

[13] D. HALL. *A system for object class detection*, in "Cognitive Vision Systems, Sampling the Spectrum of Approaches", H.-H. NAGEL, H. CHRISTENSEN (editors). , to appear, chap. Recognition and Categorization,

Springer Verlag, Heidelberg, 2005.

[14] J. TISSEAU, M. PARENTHOEN, C. BUCHE, P. REIGNIER. *Comportements perceptifs d'acteurs virtuels autonomes : Une application des cartes cognitives floues*, in "Technique et Science Informatique", to appear, 2006.

Publications in Conferences and Workshops

[15] O. BRDICZKA, J. MAISONNASSE, P. REIGNIER. *Automatic Detection of Interaction Groups*, in "Proceedings of International Conference on Multimodal Interfaces (ICMI)", ACM, october 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BMR05.

[16] O. BRDICZKA, P. REIGNIER, J. L. CROWLEY. *Automatic Development of an Abstract Context Model for an Intelligent Environment*, in "International Conference on Pervasive Computing and Communications (PerCom)", 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BRC05a.

[17] O. BRDICZKA, P. REIGNIER, J. L. CROWLEY. *Supervised Learning of an Abstract Context Model for an Intelligent Environment*, in "Proceedings of Smart Object and Ambient Intelligence Conference (sOc-EUSAI)", october 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BRC05.

[18] O. BRDICZKA, P. REIGNIER, J. CROWLEY, D. VAUFREYDAZ, J. MAISONNASSE. *Deterministic and Probabilistic Implementation of Context*, in "Proceedings of IEEE International Conference on Pervasive Computing and Communications Workshops", to appear, March 2006.

[19] O. BRDICZKA, P. REIGNIER, J. MAISONNASSE. *Unsupervised Segmentation of Small Group Meetings using Speech Activity Detection*, in "ICMI Workshop Proceedings", International Workshop on Multimodal Multiparty Meeting Processing (MMMP), october 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BRM05.

[20] O. BRDICZKA, D. VAUFREYDAZ, J. MAISONNASSE, P. REIGNIER. *Unsupervised Segmentation of Meeting Configurations and Activities using Speech Activity Detection*, in "3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI) 2006, Athens, Greece", to appear, June 2006.

[21] D. HALL, R. EMONET. *An automatic approach for parameter selection in self-adaptive tracking*, in "International Conference on Computer Vision Theory and Applications (VISAPP)", to appear, 2006.

[22] D. HALL. *Automatic parameter regulation for a tracking system with an auto-critical function*, in "International Workshop on Computer Architecture for Machine Perception, Palermo, Italy", July 2005, p. 39–45.

[23] D. HALL, J. NASCIMENTO, P. RIBEIRO, E. ANDRADE, P. MORENO, S. PESNEL, T. LIST, R. EMONET, R. FISHER, J. SANTOS VICTOR, J. CROWLEY. *Comparison of target detection algorithms using adaptive background models*, in "Performance Evaluation in Tracking and Surveillance", 2005.

[24] J. MAISONNASSE, O. BRDICZKA. *Détection automatique des groupes d'interactions*, in "Deuxièmes Journées Francophones: Mobilité et Ubiquité (UbiMob)", june 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/MB05.

[25] A. MATTHIEU, J. L. CROWLEY, V. DEVIN, G. PRIVAT. *Localisation intra-bâtiment multi-technologies: RFID, Wifi et vision*, in "Deuxièmes Journées Francophones: Mobilité et Ubiquité (UbiMob)", june 2005.

[26] F. METZE, P. GIESELMANN, H. HOLZAPFEL, T. KLUGE, I. ROGINA, A. WAIBEL, M. WOLFEL, J. CROWLEY, P. REIGNIER, D. VAUFREYDAZ, F. BERARD, B. COHEN, J. COUTAZ, S. ROUILLARD, V. ARRANZ, M. BERTRAN, H. RODRIGUEZ. *the "FAME" Interactive Space*, in "2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Edinburgh, UK", July 2005.

[27] T. T. H. TRAN, A. LUX, H. L. NGUYEN THI, A. BOUCHER. *A novel approach for text detection in images using structural features*, in "International Conference on Advances in Pattern Recognition (ICAPR), Bath, United Kingdom", 21-25 August 2005.

[28] T. T. H. TRAN, A. LUX, H. L. NGUYEN THI. *Towards a ridge and peak base symbolic representation for object recognition*, in "The 3rd International Conference in Computer Science, Can Tho University, VietNam", 21-24 Feb 2005.

[29] S. ZAIDENBERG, O. BRDICZKA, P. REIGNIER, J. CROWLEY. *Learning context models for the recognition of scenarios*, in "3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI) 2006, Athens, Greece", to appear, June 2006.

Bibliography in notes

[30] *Jess : the rule engine for the java*, http://herzberg.ca.sandia.gov/jess/.

[31] J. ALLEN. *Towards a general theory of action and time*, in "Artificial Intelligence", vol. 13, 1984.

[32] F. BERARD. *The magic table: computer-vision based augmentation of a whiteboard for creative meetings*, in "Workshop on Projector-Camera Systems", 2003.

[33] S. BORKOWSKI, O. RIFF, J. L. CROWLEY. *Projecting Rectified Images in an Augmented Environment*, in "PROCAMS'03 Workshop", 2003.

[34] O. BRDICZKA, J. MAISONNASSE, P. REIGNIER. *Automatic Detection of Interaction Groups*, in "Proceedings of International Conference on Multimodal Interfaces (ICMI)", ACM, october 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BMR05.

[35] O. BRDICZKA, P. REIGNIER, J. L. CROWLEY. *Automatic Development of an Abstract Context Model for an Intelligent Environment*, in "International Conference on Pervasive Computing and Communications (PerCom)", 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BRC05a.

[36] O. BRDICZKA, P. REIGNIER, J. L. CROWLEY. *Supervised Learning of an Abstract Context Model for an Intelligent Environment*, in "Proceedings of Smart Object and Ambient Intelligence Conference (sOcEUSAI)", october 2005, http://www-prima.inrialpes.fr/prima/pub/Publications/2005/BRC05.

[37] O. BRDICZKA, P. REIGNIER, J. MAISONNASSE. *Unsupervised Segmentation of Small Group Meetings using Speech Activity Detection*, in "ICMI Workshop Proceedings", International Workshop on Multimodal Multiparty Meeting Processing (MMMP), october 2005, http://www-

prima.inrialpes.fr/prima/pub/Publications/2005/BRM05.

[38] O. BRDICZKA, D. VAUFREYDAZ, J. MAISONNASSE, P. REIGNIER. *Unsupervised Segmentation of Meeting Configurations and Activities using Speech Activity Detection*, in "3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI) 2006, Athens, Greece", to appear, June 2006.

[39] 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.

[40] V. COLIN DE VERDIÈRE, J. CROWLEY. *Visual Recognition using Local Appearance*, in "ECCV98, Freiburg", June 1998, p. 640–654.

[41] J. CROWLEY, J. COUTAZ, G. REY, P. REIGNIER. *Using Context to Structure Perceptual Processes for Observing Activity*, in "UBICOMP, Sweden", September 2002.

[42] J. CROWLEY, O. RIFF. *Fast Computation of Scale Normalised Gaussian Receptive Fields*, in "International Conference on Scalespace theories in Computer vision, Skye, UK", June 2003, p. 584–598.

[43] W. FREEMAN, E. ADELSON. *The Design and Use of Steerable Filters*, in "Pattern Analysis and Machine Intelligence", vol. 13, nº 9, September 1991, p. 891–906.

[44] D. GARLAN, S. CHENG, A. HUANG, B. SCHMERL, P. STEENKISTE. *Rainbow: Architecture-based, self-adaptation with reusable infrastructure*, in "IEEE Computer", 2004.

[45] 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, p. 617-622.

[46] D. HALL, J. CROWLEY. *Détection du visage par caractéristiques génériques calculées à partir des images de luminance*, in "Reconnaissance des formes et intelligence artificelle, Toulouse, France", to appear, 2004.

[47] 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.

[48] B. JOHANSON, G. HUTCHINS, T. WINOGRAD, M. STONE. *PointRight: Experience with Flexible Input Redirection in Interactive Workspaces*, in "Proceedings of UIST-2002", 2002.

[49] T. LINDEBERG. *Feature Detection with Automatic Scale Selection*, in "International Journal of Computer Vision", vol. 30, nº 2, 1998, p. 79–116.

[50] D. LOWE. *Object Recognition from Local Scale-Invariant Features*, in "ICCV", 1999, p. 1150–1157.

[51] A. LUX. *The Imalab Method for Vision Systems*, in "ICVS03, Graz, Austria", April 2003.

[52] J. MAISONNASSE, O. BRDICZKA. *Détection automatique des groupes d'interactions*, in "Deuxièmes Journées Francophones: Mobilité et Ubiquité (UbiMob)", june 2005, http://www-

prima.inrialpes.fr/prima/pub/Publications/2005/MB05.

[53] J. MAISONNASSE, N. GOURIER, O. BRDICZKA, P. REIGNIER. *Attentional Model for Perceiving Social Context in Intelligent Environments*, in "3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, Athens, Greece", to appear, June 2006.

[54] 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.

[55] A. NEGRE, C. BRAILLON, J. CROWLEY. *Visual navigation from variation of intrinsic scale*, in "Submitted to ICRA", 2006.

[56] 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^o 3, 1997.

[57] G. PINGALI, C. PINHANEZ, A. LEVAS, R. KJELDSEN. *Steerable Interfaces for pervasive computing spaces*, in "IEEE PerCom", March 2003.

[58] C. PINHANEZ. *The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces*, in "Proceedings of Ubiquitous Computing 2001 Conference", September 2001.

[59] V. POLADIAN. *Dynamic Configuration of Resource-aware services*, in "Int. Conf. on Software Engineering", September 2001.

[60] F. PÉLISSON, D. HALL, O. RIFF, J. CROWLEY. *Brand identification using Gaussian derivative histograms*, in "International Conference on Vision Systems, Graz, Austria", April 2003, p. 492–501.

[61] R. RASKAR. *iLamps: Geometrically Aware and Self-Configuring Projectors*, in "ACM SIGGRAPH 2003 Conference Proceedings", 2003.

[62] 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.

[63] 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^o 1, World Scientific Press, 1994, p. 1-32.

[64] B. SCHIELE, J. CROWLEY. *Recognition without Correspondence using Multidimensional Receptive Field Histograms*, in "International Journal of Computer Vision", vol. 36, n^o 1, January 2000, p. 31–50.

[65] 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 Creativitiy and Innovation*, in "ACM Conference on Human Factors in Computing Systems", 1999.

[66] N. TAKAO, J. SHI, S. BAKER. *Tele Graffiti: a camera projector based remote sketching system with hand based user interface and automatic session summarization*, in "Int. Journal of Computer Vision", vol. 53, 2003, p. 115-133.

[67] T. T. H. TRAN, A. LUX. *A method for ridge extraction*, in "Asian Conference on Computer Vision, Jeju, Korea", 2004, p. 960–966.

[68] J. UNDERKOFFLERAND, B. ULLMER, H. ISHII. *Emancipated Pixels: Real-World Graphics in the Luminous Room*, in "Proceedings of ACM SIGGRAPH", 1999, p. 385-392.

[69] 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.

[70] F. VERNIER, N. LESH, C. SHEN. *Visualization Techniques for Circular Tabletop Interfaces*, in "Advanced Visual Interfaces", 2002.

[71] 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.

[72] P. WELLNER. *The DigitalDesk Calculator: Tactile Manipulation on a Desk Top Display*, in "ACM Symposium on User Interface Software and Technology (UIST '91)", November 1991, p. 27–33.

[73] S. ZAIDENBERG, O. BRDICZKA, P. REIGNIER, J. CROWLEY. *Learning context models for the recognition of scenarios*, in "3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI) 2006, Athens, Greece", to appear, June 2006.

[74] B. ZOPPIS. *Outils pour l'Intégration et le Contrôle en Vision et Robotique Mobile*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, June 1997.