SIXTH FRAMEWORK PROGRAMME
PRIORITY 2
INFORMATION SOCIETY TECHNOLOGY

Contract for:

SPECIFIC TARGETED RESEARCH PROJECT
STREP

Annex I – Description of Work

Project acronym: **POP**
Project full title: **Perception On Purpose**
Proposal/Contract no.: 027268

Contents

1 Project summary

The ease with which we make sense of our environment belies the complex processing required to convert sensory signals into meaningful cognitive descriptions. Computational approaches have so far made little impact on this fundamental problem. Visual and auditory processes have typically been studied independently, yet it is clear that the two senses provide complementary information which can help a system to respond robustly in challenging conditions. In addition, most algorithmic approaches adopt the perspective of a static observer or listener, ignoring all the benefits of interaction with the environment. This project proposes the development of a fundamentally new approach, perception on purpose, that puts forward the modelling of perception (visual and auditory) as a complex attentional mechanism that embodies a **decision taking** process. The task of the latter is to find a trade-off between the reliability of the sensorial stimuli (bottom-up attention) and the plausibility of prior knowledge (top-down attention). The scientific and methodological developments are based on five principles. First, visual and auditory information should be integrated in both space and time. Second, active exploration of the environment is required to improve the audiovisual signal-to-noise ratio. Third, the enormous potential sensory requirements of the entire input array should be rendered manageable by multimodal models of attentional processes. Fourth, bottom-up perception should be stabilized by top-down cognitive function and lead to purposeful action. Finally, all parts of the system should be underpinned by rigorous mathematical theory, from physical models of low-level binocular and binaural sensory processing to trainable probabilistic models of audiovisual scenes. These ideas will be put into practice through behavioural and neuroimaging studies as well as in the construction of testable computational models. A demonstrator platform consisting of a mobile audiovisual head will be developed and its behaviour evaluated in a range of application scenarios. Project participants represent expertise in computational, behavioural and cognitive neuroscientific aspects of vision and hearing needed both to carry out the POP manifesto and to contribute to the training of a new community of scientists.

2 Project objectives

2.1 Objectives, approach, and originality

Whenever researchers attempt to investigate and build an intelligent system, they are confronted with the problem of modelling the interactions between the system and its physical environment. The world that surrounds artificial and/or biological agents has three dimensions and its geometrical/physical structure varies over time. Therefore it is not surprising that one of the primarily goals of perception is to create an internal spatio-temporal representation of the external environment. One reason why both the spatial and temporal dimensions of sensory processing are so important is because internal representations of space and time are needed to guide behaviour such as movements and therefore they allow the system to interact with the physical world.

In the recent past, visual and auditory processes have been studied almost independently. Cross-modal integration of visual and auditory data and interaction between auditory and visual processes are beneficial because each modality provides partial information about different aspects of world objects and events, and combination of these modalities can be very important in understanding how they complement each other to provide unambiguous world information, how they transform their inputs into knowledge and meaning, and how they control behaviour.

> *The objective of this project is to put forward the modelling of perception (visual and auditory) as a complex attentional mechanism that embodies a **decision taking** process. The task of the latter is to find a trade-off between the reliability of the sensorial stimuli and the plausibility of prior knowledge.*

For example, the system should be able to localize in space, identify, and track over time an object that can be seen and heard simultaneously, i.e., crossmodal integration. From the point of view of auditory analysis, **binaural cue detection** may be used for sound localization and for building a spatial map of auditory sources. From the point of view of visual analysis, visual cues such as **stereoscopic vision** may be used for producing an observer-centered depth map of the 3-D layout. However, crossmodal integration cannot be performed at that level because there is no obvious way to associate depth and sound sources. Temporal processing of both sound and vision enhance these spatial descriptions. Sequential processing of audio data may **group events into streams** and produce sets of distinct sounds perceived as arising from a single source. Similarly, **optical flow** may be extracted from image sequences and the visual data can be segmented into static and dynamic, into fast and slow motions, and into rigid and deformable motions. Therefore we have additional audio and visual descriptors, yet this is not sufficient to associate audio and visual events. Crossmodal integration, followed by decision and action, necessitates a higher level association process, beyond the physical and geometrical formats that describe sound and light stimuli. We firmly believe that the perception-action cycle needs some form of cognitive modelling.

The integration of visual and auditory cues with control of behaviour will be one of our central topics of investigation. *In particular, we believe that the interactions between the sensory level processes (bottom-up) and the cognitive level processes (top-down) reside at the core of a cognitive system and these interactions distinguish the latter from a classical computer system.*

The approach that will be advocated in this project and implemented within its associated work-packages relies on the following principles:

- **Auditory and visual cues must be integrated both in space and time**. We will encapsulate various perceptual processes for enabling the fusion of the processed data. We will develop

theoretical and algorithmic models based on image and signal processing, on geometric fusion, on probabilistic modelling of both the input and the processed data, and on statistical modelling.

- **Physical inputs must be confronted with prior knowledge**. This data-to-meaning association problem will be thoroughly addressed and its computational complexity will be modelled and investigated. Probabilistic models (Bayesian decision theory and hidden Markov models, in particular) will be combined with efficient optimization techniques.

- **Visual sensors must actively explore the world.** The field of view of a visual sensor is inherently limited. Therefore a static sensor cannot build a complete and reliable representation of the world. Furthermore, when the attention of an observer must focus onto a specific region, the narrowness of the field becomes a desirable feature. Therefore, active vision plays a crucial role at all levels of information processing: there is a strong link between attention and sensory-motor control. We will thoroughly investigate these issues from the view point of control theory and in synergy with neurophysiology and psychophysics.

- **Computational models for active listening will be thoroughly investigated.** The freedom to orient and move the location of the auditory sensors has tremendous potential for improving the signal-to-noise ratio of the attended sound source. Orientation allows for directional listening and blocking of unwanted sources by head-shadowing, while movement enables the listener to avoid hard reflective surfaces, move closer to the target, and get an unobstructed view of information-bearing features such as lip movements.

- **We will build a proof-of-concept robotic platform** which will include the following elements: the design of an active audiovisual head mounted onto a mobile robot, algorithms and software that implement visual and audio cue extraction, cross-modal integration, end sensory-motor decisional loops.

- **Biological attention mechanisms will be quantitatively studied** using the most modern eye-tracking devices in conjunction with advanced brain imaging techniques. In particular we will study the relation between cross-modal stimulus integration and attentional selection at a behavioural level. Key areas in the brain where crossmodal integration occurs must be identified. These experiments will guide the implementation of the active audio-visual robotic head.

- **Neurophysiological and psychophysical findings will be cast into an abstract computational model that is implementable with today's computers**. The task of modelling the brain from a computational point of view is tremendous. It is possible to model groups of neurons as a dynamic system using partial derivative equations. But these models cannot yet be generalized to complex brain functions such as those involved in perception. We believe that an intermediate reachable goal is to consider an abstract model based on modern statistics. This type of approach has been successful in many cognitive-system related fields such as artificial intelligence, computer vision, speech recognition, etc.

- In the past, **computational models for vision and hearing have been studied independently** by distinct research communities with almost no interaction. **We will adopt a multidisciplinary approach** involving a broad range of methodologies from computational and biological vision, robotics and control, computational auditory scene analysis, speech recognition, psychophysics, and psychology.

- **The simultaneous study of visual and auditory processes and their relationship with attention and motor control** is a novel scientific endeavour. We feel that the time to combine research results from all these disciplines has come.

- More precisely, we will address the difficult problems of **integrating spatial and temporal audio-visual stimuli** using a geometrical and probabilistic framework and we will attack the problem of **associating sensorial descriptions with representations of prior knowledge**. Moreover, we will **design an audio-visual active robotic head** and we will **build models for sensory-motor control.**

- **The fact that computer scientists are trying to build an artificial system that models biological systems made out of neurons is often ignored**, and until recently there was no answer to the question of biological plausibility: how a given algorithm may be implemented in a biological brain? Equally important is the issue of computational neuroscience: how one could implement with hardware and software recent neuro-physiological findings which model the way our brain functions?

- Nevertheless, **the structure of today's computers and their common programming languages is very far removed from the brain's architecture**. Moreover, there is a lack of formalism and of common theoretical and practical methodologies allowing the above mentioned disciplines to work together and to produce common research results. This gap (and its impact onto both computational and biological models) may be reduced in the near future but it is obvious that this cannot be done within a single research project.

- **The contribution of the POP project will be to address a specific research topic and to set up a common set of formal models** (upon which both communities agree), and **to combine a broad range of methodologies** in order to both (i) *implement these models with computers* and (ii) *verify them through the measuring and interpretation of human-brain activity.*

3 Participant list

Part. role	Part. no.	Participant name	Part. Sh. name	Ctry	Date enter Project	Date exit Project
CO	1	Institut National de Recherche en Informatique et Automatique	INRIA	F	month 1	month 36
CR	2	University of Osnabrück	UOFM	D	month 1	month 36
CR	3	University Hospital Hamburg-Eppendorf	UKE	D	month 1	month 36
CR	4	Fac. de Ciencias et Tec. Univ. Coimbra - ISR	FCTUC	P	month 1	month 36
CR	5	The University of Sheffield	USFD	UK	month 1	month 36

Table 1: *List of participants*

4 State-of-the-art and scientific and technology baseline

The primarily objective of the POP project is to address a specific research issue, namely the problem of modelling, implementing, and testing an artificial system that gathers sensorial information on purpose – according to the task at hand – and that controls its attention and behaviour. Therefore we will address scientific and technological issues associated with an *artificial cognitive system* that interacts with the physical world. In our specific case these interactions will be materialized through the use of sensors and actuators.

We will consider both visual and auditory sensors and we will address the problems of how these sensorial modalities integrate and complement each other, how they interact with task-dependent higher-level knowledge, and how they are combined with actuators within sensory-motor feedback loops. The development of systems able to transform perception of sensorial stimuli into meaning is at the core of future cognitive systems. The in depth understanding of the processes allowing to map light and acoustic **physical signals** into **symbolic descriptions** is the keystone for the development of intelligent agents able to communicate with people and to take decisions.

Neuroscience researchers have recognized for a long time that perception is a high-level cognitive function. Although we have already accumulated a great deal of knowldege about the brain's anatomy, physiology, and neural mechanisms, this knowledge is not nearly enough to determine analytic equations that describe large systems of neurons. Even if precise knowledge of neural dynamics is available, it would yield only partial understanding of how brain functions. Therefore, direct approaches based on studying the brain **must be augmented** with information processing approaches that attempt to model perception (and intelligence in general) at a more abstract level.

The POP project will therefore contribute to establish a unifying theory of auditory and visual perception that is quantifiable (i.e., mathematical formulation, computer algorithms and software, and thorough experimental validation).

When viewing and analyzing complex natural scenes, humans do not process all the sensorial information homogeneously and simultaneously. Attention is rapidly directed to small parts of the scene that are processed in greater detail. This attentive behaviour, however, goes well beyond simple stimulus/actuator loops. A decision has to be made concerning which part of the environment should be attended to, and subsequently, appropriate movements of the sensors have to be made (these movements raise difficult inverse-kinematic and dynamic issues that are worth to be studied in their own right). Thus, attention involves a complete perception-decision-action cycle in closed loop with both the internal representations and the environment.

In contradiction with the biological process just described, current computerized visual and auditory scene analysis systems operate in a much more static way. **Snapshot vision has been the major paradigm studied worldwide** and researchers exhaustively investigated geometric, photometric, and statistical methods for recovering world information from a set of fixed cameras. Similarly, **computational hearing has been dominated by speech recognition**, which has typically been studied in the absence of competing sources that characterize everyday listening situations.

The POP project will fill in an important gap by attempting to reduce the distance that separates current audio/visual processing paradigms, on one side, from higher-level skills, on the other side, such as the ability of an artificial system to take decisions, to interact with the world, and on a longer term, to communicate with people at an abstract level.

4.1 The state of the art

In the past, neurobiological studies recognized the importance of multisensory (or cross-modal) integration and its links to attention. Substantial knowledge is available, within individual sensory channels, regarding how the brain extracts information from environmental events and converts it to perceptions, memories, and actions. Events and objects in the the outside world are often jointly detected by two or more sensory systems and the perceptual and behavioural consequences are not just the sum of either sensory component alone [104].

Two main views exist about how multisensory integration may be accomplished: (i) multisensory interactions could rest on a convergence of different pathways in higher-order association areas, possibly with feedback connections to earlier sensory areas [43], or (ii) multisensory integration could be achieved by dynamic interactions between sensory modules. The role of dynamic binding of by changes in the coherence of neural activity within sensory areas has been shown to be important in unimodal feature binding [102],[46] and is yet underexplored in the context of multisensory perception.

While substantial evidence is available supporting the notion of temporal binding within individual modalities, nothing is known about the potential relevance of such mechanisms for dynamic cross-modal integration. The prediction that can be derived from the temporal binding model is that neural modules involved in the processing of different sensory aspects (e.g., visual and auditory features of an object) should change their dynamic interaction and, specifically, their coherence, depending on the cross-modal relation of object features. As one possible consequence of such an enhanced coherence, higher-order multimodal areas receiving synchronized inputs might then become more strongly activated. These hypotheses will be tested in the project suggested here.

While bottom-up factors certainly play a key role in determining such dynamic interactions, top-down factors also can be expected to be of crucial importance [45]. These findings lead, in the context of assembly dynamics, to the hypothesis that attention facilitates the joint processing of multi-modal stimuli by synchronizing responses across separated uni-modal brain areas in a similar way as has been observed in the unimodal (visual) case for spatially separate neuronal populations [53].

Mechanisms of attentional gating have so far mostly been investigated in visual areas and revealed that the strength of neuronal responses is increased when the receptive field of the cells under study are encompassed by the focus of attention [84], [85], [97]. For cross-modal shifts of attention, no particular mechanism has been identified, but the general hypothesis thus far had been that enhanced response amplitudes expressed as increased firing rates at the neuronal level serve to increase the saliency of certain representations which enables them to take control over behaviour or induce memory formation. There is evidence from ERP studies [82] that acoustically-induced feedback from multimodal superior temporal cortex influences visual processing some 15 to 25 ms later. Although the latter example is reminiscent of reflex behavior, it represents some kind of expectancy of the system which has in the motor-cortex been observed at the neuronal level [95]. One of the first studies that clearly went beyond the concept of response enhancement as a mechanism for attention, described that changing of the task-relevant sensory input from somatosensory to visual information reduces spike synchrony in secondary somatosensory cortex [105]. The concept of biased competition for explaining the processes underlying selective attention within the visual domain [93] has led to the direct investigation of neuronal synchronization as a possible mechanism [53]. A recent textbook [98] dedicated to the computational neuroscience of vision proposes a mathematical model for visual attention based on the concept of biased competition.

Computer vision researchers have been deeply influenced and inspired by the work of Marr [81] and totally ignored the role (and hence the modelling) of attentional mechanisms in vision [51]. A

recent European workshop dedicated to visual attention gathered an interesting collection of papers that propose several computational approaches to visual attention [56], [109], [2], [38], [60], and others.

In an attempt to model neurobiological findings and psychophysical experiments related to cross-modal integration [26], [108], [18], [101], more abstract models have been suggested that are based on Bayesian decision theory [15], [78], [70], [40], to cite just a few. A recent paper [79] reviews various attempts for modelling the computations in the early visual cortex.

It is only very recently that scientists have started to address the problems of audio-visual integration and perceptual attention from an algorithmic point of view. Several authors proposed interesting solutions to combine a monocular visual tracker with speech localization [89], [106], [58, 57]. It is interesting to notice that the vast majority of these approaches and methods consider monocular visual cues only. The fact that gaze control may be combined with stereopsis and that the active exploration of depth cues may influence auditory analysis has been completely ignored [108].

There exists a well developed account of auditory perceptual organisation, 'auditory scene analysis' (ASA), which describes how listeners build an internal model of their acoustic environment from the raw acoustic data arriving at the ear using a mixture of primitive and schema driven processes. The ASA account has been built using 'listening experiments' (i.e. with no visual input). As a result most computational models of ASA are currently blind to the fact that hearing is profoundly influenced by visual information [16], [83], [42]. There have been recent attempts integrate visual input with models of audio scene analysis, [103], but there is little human behavioural data on which to base such models. Although much attention has been paid to the influence of vision on the perception of speech, the role of visual information has not been systematically studied.

State of the art audio-visual speech recognition systems [92] use the visual input primarily to provide support for discriminating between speech units when the acoustic information becomes unreliable. They do not use the visual information in the 'POP' sense, that is, they do not use the visual information as a cue to aid the segregation of the competing sound sources, and hence to improve the reliability of the acoustic information. Existing model-based ASA approaches to speech recognition [3] can be readily extended to incorporate visual speech information in a manner that would exploit both its value as an organisational cue and its phonetic (visemic) content.

4.2 The state of the art of the POP teams

Partner 1 (INRIA) developed various statistical methods based on HMM (Hidden Markov Models) and on EM (expectation-maximization). In particular, the problem of image segmentation by unsupervised learning was addressed. [50], [27], [28], [49]. INRIA researchers developed methods for vision-based robot control using calibrated or uncalibrated cameras [65], [99], [39], [77], [100]. The problem of image matching was addressed using graphs [67] and using points of interest [44]. Bottom-up attention was addressed as a figure-ground discrimination problem [62], [63], or within an active stereo system able to focus on a single object [66]. Other computer vision issues addressed were 3-D reconstruction [29], interpretation of surface contours [64], and top-down recognition of objects using graph spectra [68].

Active vision issues such as sensory-motor control, active stereo sensing, and panoramic vision were addressed in the past by partner 4 (FCTUC) [12], [11], [13], [41], [7], [14], [8], [9], [10], [90], [6], [91], [5], [30], [4].

Partner 5 (USFD) is the POP's expert in computationally auditory scene analysis and speech recognition and USFD researchers developed methods able to recognize speech in the presence of other sources of sound and with missing or uncertain data [3], [33], [34], [32], [31], [35], binaural

interaction [61], and methods for handling convolutional distorsion [88].

Neurophysiological models were developed by partners 2 (UOFM) and 3 (UKE). Their join work on oscillatory neural synchronization is wellknown. This basic model lead UOFM and UKE researchers to establish cognitive and neuronal models for stimulus attention, selective visual attention, overt attention, top-down processing of information, [59], [47], [86], [54], [96], [20], [21], [71], [72], [19], [46], [48], [52], [73], [55], [74], [36], [75], [87], [22], [37], [76], [94], [17], [24], [69], [1], [107], [23], [25], [80].

5 Potential impact

From a scientific point of view, the main challenge of POP is to address the specific problem of modelling attentional mechanisms by integration of three ingredients: interactions between prior knowledge and sensorial data, cross-modal integration of visual and auditory stimuli, and the coordination between sensorial information processing and motor control.

The POP partners will be committed to build a common technological platform. The latter will basically consist in both hardware and software components. It will gather an audiovisual active head – a stereoscopic camera system and a binaural microphone system – mounted onto a mobile robotic platform, as well as software that will implement both existing algorithms and algorithms to be developed by the POP partners. This platform will be designed by the POP partners. The software libraries and packages will be open source code. We will exploit both our hardware design and our open-source software developments (through patents and software licenses) and we will disseminate them such that they impact onto technology transfer. Moreover, the raw and processed data sets produced by this platform will be disseminated as well in order to allow other researchers to test their own methods and/or for bench-marking purposes.

POP's outcomes will also be likely to have **an important impact on a wide spectrum of application developments** and this impact is detailed below.

There are increasing needs for **human activity recognition** based on video and audio analysis. These needs span from **security and safety applications** (traffic, airport, and building surveillance) to **medical monitoring** of people with special needs. In order to cope with some legal regulations and to protect the private life of people under monitoring, it is crucial not to communicate images but higher-level abstract descriptions. Therefore, a cognitive approach is crucial.

Another challenging application domain on which POP will have impact is the domain of **human-to-human and human-to-machine communication**. Video conferencing systems are prominent examples of systems that heavily rely on visual and audio processing and interpretation. These systems are in their primitive stage of development. Current prototypes and commercial products use fixed cameras and microphones. In the near future we predict that the audio-visual capturing device will (and should) act like a *autonomous and intelligent cameraman robot* that selects a speaker among several people, and purposively moves towards this speaker to avoid occlusions, to increase the signal to noise ratio for reliable speech recognition, and to zoom onto the speaker's face from the right viewpoint.

Another example where an audio-visual cognitive system that selectively processes and interprets the sensorial information is important, is the example of a **driving assistant**. In the long term, all European car manufacturers foresee that visual sensors and their associated software will become part of standard equipment. Cameras' fields of view will span both outside and inside the car. The system will be able to detect and recognize obstacles and to verbally alert the driver. Other audio and visual sensors will be dispatched inside the car for the recognition of the driver's state and behaviour.

The methodological and algorithmic developments to be carried out within POP's work-packages will allow integration of vision and hearing in complex situations. It will be possible, for example, to observe a scene with several people (moving around and speaking) and to associate a voice signal to each one of the persons in the scene. This will have a **direct and practical impact onto the development of hearing aids**. State-of-art hearing aids give good understanding of individual speakers but fail in cluttered and multi-source environments. For this reason, the acceptance of hearing aids by patients is still poor. The ability to focus the auditory attention onto a target that is visually selected could solve this fundamental problem.

The **European impact of POP** will be effective through collaborations with other projects.

5.1 Contribution to standards

Not relevant to this project

5.2 Contribution to policy developments

Not relevant to this project

5.3 Risk assessment and related communication strategy

Not relevant to this project

6 Project management and exploitation dissemination plan

6.1 Project management

The management will be organized as follows. First of all a **Steering Committee** will be created (SC). The SC will be the consortium's main decision making and arbitration body. The SC will have one representative from each contractor and will be chaired by a representative of the coordinator – **The project coordinator**.

The Steering Committee is the decision making body of the project. All contractual issues, changes in technical specifications of workpackages, IPR issues, risk management, etc. will be discussed and decided by the Steering Committee. The decisions will be taken with a majority vote, one vote per partner.

The SC will hold a video-conference meeting every month and a physical meeting every 6 months in order to monitor the progress of the project, anticipate difficulties, and take decisions, through the following tasks:

- Appoint the **site-** and **workpackage managers**. Provisionally, the site managers will the following persons:
  - INRIA – Radu Horaud.
  - UOFM – Peter König.
  - UKE – Andreas Engel.
  - FCTUC – Helder Araujo.
  - USFD – Martin Cooke.
- Decide revisions and their use of project resources for work-packages;
- Be in charge of risk management, monitoring the progress of the project and ensuring that any changes are discussed and implemented in a timely fashion;
- Take actions in case a contractor makes default;
- Anticipate and suggest solutions in case of a conflict;
- Elaborate the rules governing the Intellectual Property Rights (IPR);

Under the direction of the **Steering Committee** there will be the **site managers** and the **work-package managers** (or work-package leaders). Since each partner will have the leadership of at least one workpackage, the site managers will be the same physical persons as the work-package managers.

The **Project coordinator** appointed by the coordinating partner and by the SC will be responsible for the overall day-to-day project administration and financial management. He will:

- Organise the relation between the Contractors and the European Commission acting as the contact-point for the project;
- Collect, monitor and integrate all the technical, administrative, and financial data from the partners and prepare appropriate documents for the European Commission and for the auditors: management reports, progress reports, final report, cost statements, financial statements, audit reports, etc.;

- Organise the technical audits, and;
- Chair the **site managers'** and **work-package managers'** sessions and meetings;
- Play a pivotal role between the **Steering Committee** on one side and the project researchers and engineers on another side, both top-down and bottom-up.

The **Work-package managers** will be responsible of the scientific and technical advancement of the project. They will:

- Hold a video-conference meeting on a monthly basis or as often as it is required for advancing the work;
- Hold a technical meeting every 6 months;
- Coordinate the work carried out within the individual work-packages and tasks and ensure its adherence to the pre-defined work-plan and timetable;
- Anticipate deviations from the planned work and monitor actions to correct these deviations;
- Report to the **Project coordinator** any foreseen problems and suggest practical solutions to overcome these problems in order to allow smooth execution of the planned tasks;
- Coordinate the preparation of project deliverables and assist the **Project coordinator** for the preparation of the technical documents to be delivered to the European Commission;

Risk, problem, and conflict management. Under the leadership of the project coordinator, the work-package managers and the site managers will be responsible for smoothly carrying out the work during the duration of the project, at both the work-package level and at the site level. Whenever needed they will take the initiative to organize management meetings at the work-package or site levels in order to anticipate any risks, problems and/or conflicts:

- Detect scientific and technical difficulties in achieving a work-package or a work-task and provide alternatives and solutions to the problems encountered.
- Anticipate problems with the termination of a work-package and offer suitable solutions before the situation results in delays or dead-ends.
- Foresee fluid information sharing between the partners (background research, outcomes of a work-task, etc.) and avoid technical and bureaucratic barriers to information access.
- Anticipate the end of contract of a researcher, changes in personnel, reallocation in human resources by a partner, leave of absence of a researcher, temporarily absence of a researcher, conflicts between researchers, etc.

If a satisfactory solution cannot be found on a rapid and cordial basis, the work-package manager will refer to the project coordinator and to the site managers who will ask the Steering Committee to meet and to take a decision.

6.2 Plan for using and disseminating knowledge

The manager of work-package 5 will act as the **Exploitation-Dissemination Manager** and will coordinate all the exploitation, training, and dissemination activities of the project. His specific responsibilities include:

- Coordinate the promotion of the results of the project through the **POP website**;
- Promote the preparation and presentation of POP papers at international journals, conferences, and specialized workshops.
- Organize tutorials and thematic schools both at the project level and in collaboration with other parties (University departments, European networks, Marie-Curie actions, etc.);
- Promote and establish bi-lateral contacts with other IST projects and instruments relevant to POP;
- Assist project researchers for the organisation of workshops and tutorials aimed at promoting the project and at collaborating with other similar European consortia;
- Establish and maintain a list of European companies interested in the outcomes of the project.
- Periodically advertize the project achievements;

The POP partners will be committed to launch an ambitious interdisciplinary training programme. The POP themes and teams will be attractive to talented doctoral and postdoctoral researchers. However, very few will possess the required interdisciplinary skills from the outset. The application of bio-inspired approaches to engineering problems requires the integration of ideas from a number of disciplines, ranging from cognitive science and psychology to computation, mathematics and neurobiology. Few individuals can claim expertise in more than one of these areas, and even within a single discipline such as neurobiology, far too little work crosses traditional boundaries. Auditory and visual processing are typically tackled in isolation by different communities, and hardly any engineering work straddles the boundary between hearing and vision. A solution to the problems which will be tackled in POP will require integration of work in hearing, vision and other modalities at every level of the project, from low-level salience to high-level scene understanding. Consequently, **all project researchers will be given the opportunity for cross-training in disciplines in which project partners have expertise**. They will be encouraged to develop the technical and collaborative skills needed to solve complex problems. The training impact of POP will be amplified through the European networking that is planned (see below).

The POP project will play an inter-disciplinary and training role. Indeed, none of some related projects (VISITOR, VISIONTRAIN, and AMI) covers all the scientific disciplines spanned by POP. For instance, there is no robotics, computer vision, and neuroscience research in AMI and there is no robotics, neuroscience, speech and auditory research in VISIONTRAIN. **POP will act as a link between these large projects and will contribute to their training and dissemination activities**. Moreover, it will add a cognitive and neuroscience dimension to these IST projects.

Management of knowledge and intellectual property rights. Intellectual Property Rights (IPR) will be handled in line with the general policy of the European Commission regarding ownership and exploitation of rights and confidentiality. Management of IPR issues will have two aspects: management of IPR internal to the Consortium and management of IPR in relation with external actors. While

the protection of IPR against external actors is in most of the cases the competence of the individual organisations, the Consortium will offer several initiatives in that respect. The **Steering Committee** will be responsible for the design of the overall IPR protection policy of the project. The general policy will include, among others: rules on security, procedures for information exchange, recommendations for safe dissemination of results, advice for IPR protection, definition of use cases and procedures, etc. The **Steering Committee** will advice individual partners on the best IPR protection method at every stage of the project development. The Committee will also take the responsibility to lead and coordinate protection of results jointly owned by several of the partners in the POP consortium.

In a first approach, **background information** and **background patents** will be made available to the consortium members on favorable conditions, if they are necessary to perform the research in this project and if no major business interest of the owner of the background information opposes the disclosure or grant of licences for such patents or information. **Foreground information** and **foreground patents** are owned by the contractor generating such information. Each contractor shall make available its foreground information, on a royalty-free basis, to other contractors to the extent that such information is necessary for the execution of their own research within the project. If proprietary information is made available, the information shall be duly marked as confidential and the recipient will preserve its confidentiality. Non-disclosure agreements will be prepared if so requested.

IPR issues will be detailed in the Consortium Agreement. Particular emphasis will be put on the **coordination of dissemination and IPR policies** in order to make compatible ample diffusion of the project results with the necessary protection of rights.

In the particular case of the POP project, the consortium is formed of academic partners. Therefore, there will be no IPR restrictions concerning the publications of scientific articles and reports.

6.3 Raising public participation and awarness

The POP partners have their own “science and society” programmes aimed at disseminating academic research results. The teams involved in POP will contribute to these programmes during the lifetime of the project.

7 Workplan – for the whole duration of the project

7.1 Introduction – general description and milestones

The work will be carried out in six work-packages. There will be four scientific and technological work-packages, one work-package dedicated to training and to the exploitation and dissemination of the project's results, and one management work-package. In order to maximize interactions between disciplines and partners, all work-packages are a "mixture" of computational/algorithmic and cognitive/biological approaches and methods.

WP1 – Cognitive mechanisms of attention, will address the problem of attention of perception at a fundamental level, using the most modern neurobiological and psychophysical investigation techniques, and proposing computational models.

WP2 – Integration of visual and auditory cues, will study and implement methods, algorithms and software for cross-modal integration, as well as the experimental testing of neurobiological hypotheses.

WP3 – Sensory-motor coordination, will address the link between perception and action from biological, computational, and algorithmic point of view. The WP's output will be both methodological and practical, i.e., software libraries.

WP4 – Development of methodological and experimental platforms, will build and integrate software and hardware components within a audiovisual robot platform that will be used for tests, validations, and applications scenarios.

WP5 – Exploitation, training, and dissemination of results, will coordinate a number of activities aimed at training and at promoting the results of the project.

WP6 – Management, will coordinate the management activities to be jointly carried out by the partners.

Milestone	Month	Description
M1	1	Interdisciplinary tutorial session **Expected results:** Bring together background knowledge and establish bilateral collaborations
M6	6	Tutorial session will include first results of WP1. Workshop on audiovisual attention **Expected results:** Bring together POP and other researchers and communicate the first results.
M12	12	Neurophysiological bases for audio-visual mechanisms of attention Computational models based on neurophysiological/psychophysical findings will be proposed by POP partners. **Expected results:** scientific reports on preliminary results: cognitive and computational models described in detail. First prototype of audio-visual head (hardware) available.
M18	18	Methods and algorithms for cross-modal integration (audio and video) based on stimulus (sound and light) processing First prototypes of algorithms for sensory-motor control **Expected results:** Algorithms are demonstrated using data gathered with the audiovisual head.
M24	24	Second prototype of audiovisual head (hardware/software) The coupling between recorded psychophysical data and an audiovisual robotic head. Computational models of attention fully completed. **Expected results:** Scientific publications, First version of demonstrator available. A workshop will disseminate both methodological and experimental results.
M30	30	Final results on neurophysiological and computational models. An abstract model for cross-modal integration based on statistical methods Final results on sensory-motor control. **Expected results:** Description of algorithms, Description of methods. Description of software packages.
M36	36	Final demonstrator available: Audiovisual head selects a person in a crowded room based on POP findings. **Expected results:** Description of hardware and software, publications based on interdisciplinary investigations, a workshop will disseminate the final results.

Table 2: *The milestones of the POP project.*

7.2 Work planning and timetable

The work planning and the timetable are detailed on figures 1 and 2.

7.3 Graphical presentations of workpackages

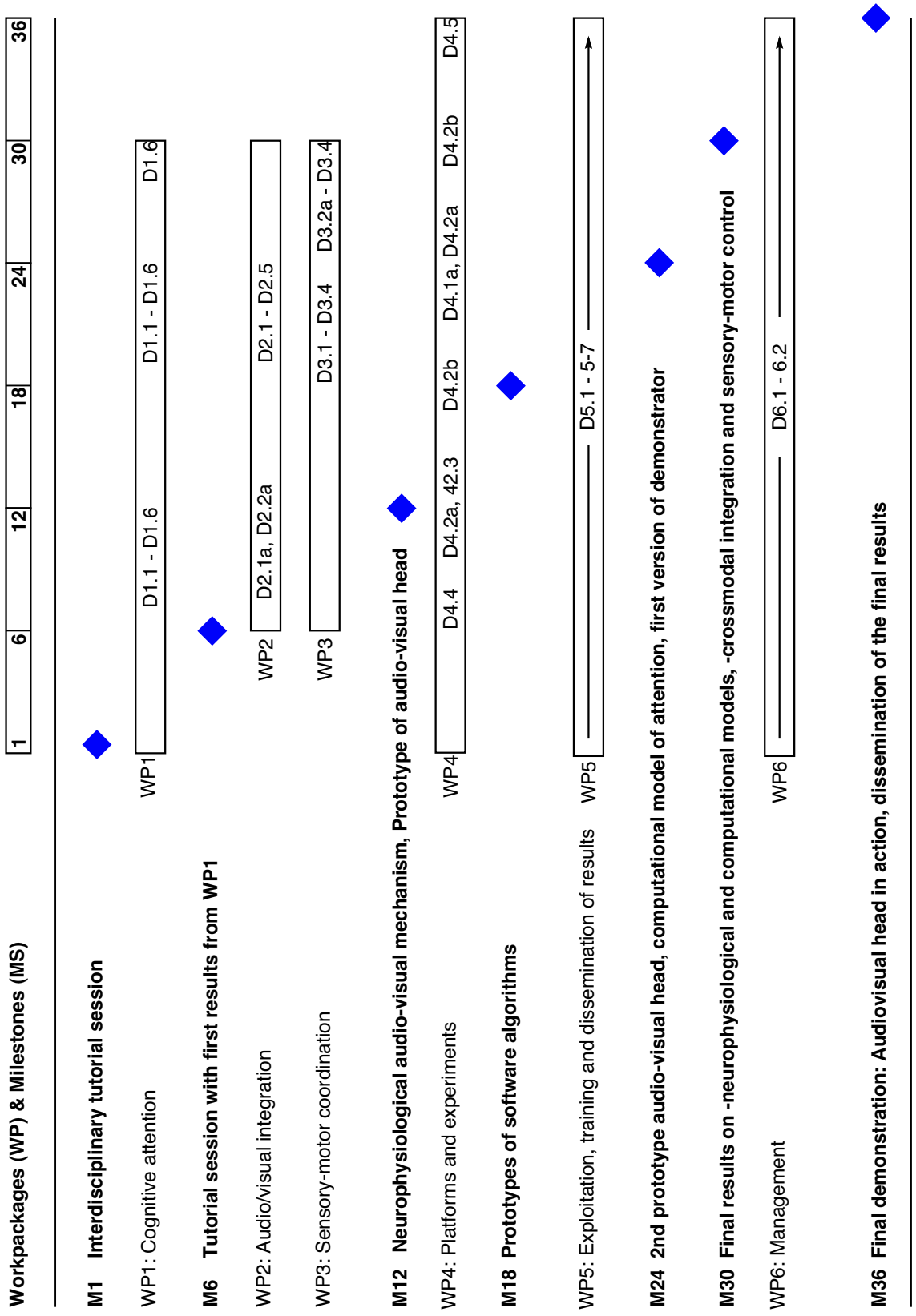

Figure 1: *Twe work-plan gant chart.*

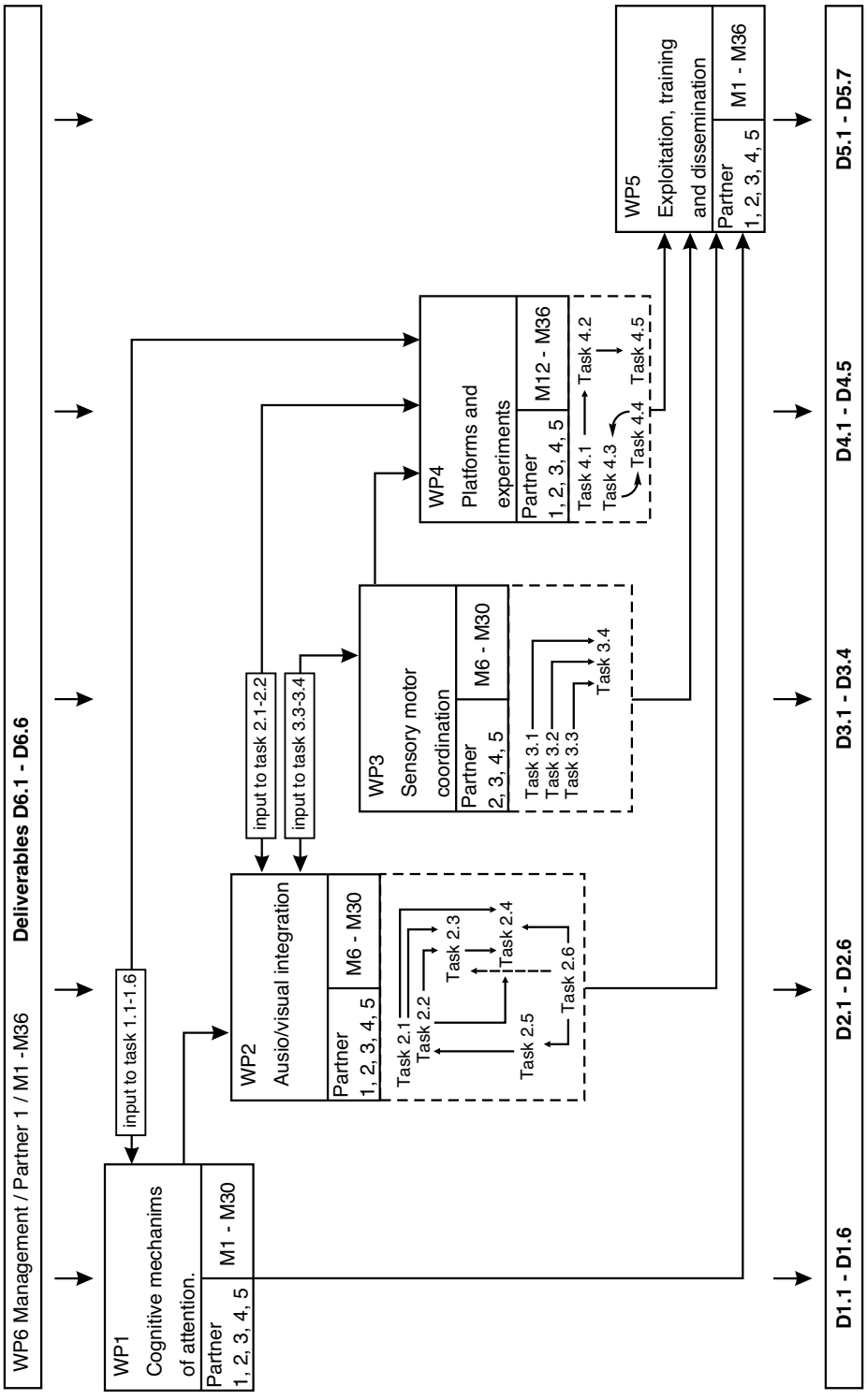

Figure 2: *The work-plan pert chart.*

7.4 Workpackage list/overview

WP No.	Workpackage title	Leader	Person-months	Start	End
WP1	Cognitive attention	UKE (3)	112	1	30
WP2	Audio/visual integration	USFD (5)	94	6	30
WP3	Sensory-motor coord.	UOFM (2)	72	6	30
WP4	Platforms and experiments	FCTUC (4)	84	1	36
WP5	Expl., train., diss.	USFD (5)	17	1	36
WP6	Management	INRIA (1)	17	1	36
	Total		396		

Table 3: *Workpackage list for the full duration of the project.*

7.5 Deliverables list

WP	Del. no.	Date	Name	Lead	PM	Type/Diss.
WP1	D1					
T1.1	D1.1	12,24	Bottom-up attention	UOFM	24	R/PU
T1.2	D1.2	12,24	Multi-feature interactions	UOFM	12	R/PU
T1.3	D1.3	12,24	Attention on purpose	INRIA	7	R/PU
T1.4	D1.4	12,24	Multimodal interactions	UKE	11	R/PU
T1.5	D1.5	12,24	Overt attention	UKE	13	R/PU
T1.6	D1.6	12,24	fMRI analysis	UKE	10	R/PU
T1.7	D1.6	12,24,30	Computational models	USFD	35	R/PU
WP2	D2					
T2.1	D2.1.a	12,24	Algorithms/software	INRIA	14	P/PU
	D2.1.b	24	Extraction of visual cues	INRIA	2	R/PU
T2.2	D2.2.a	12,24	Algorithms/software	USFD	7	P/PU
	D2.2.b	24	Extraction of auditory cues	USFD	1	R/PU
T2.3	D2.3.a	24	Algorithms/software	USFD	7	P/PU
	D2.3.b	24	Spatial fusion	USFD	1	R/PU
T2.4	D2.4.a	24	Algorithms/software	UKE	6	P/PU
	D2.4.b	24	Temporal fusion	UKE	6	R/PU
T2.5	D2.5.a	24	Algorithms/software	INRIA	18	P/PU
	D2.5.b	24	Cross-modal integration	INRIA	14	R/PU
T2.6	D2.6.	30	Biological models	UOFM	18	P/PU
WP3	D3					
T3.1	D3.1	12	Algorithms/software	FCTUC	6	P/PU
T3.2	D3.2.a	24,30	Algorithms/software	UOFM	6	P/PU
	D3.2.b	24,30	Oculomotor control	UOFM	13	R/PU
T3.3	D3.3.a	24,30	Algorithms/software	USFD	20	P/PU
	D3.3.b	24,30	Audiomotor control	USFD	8	R/PU
T3.4	D3.4.a	24,30	Algorithms/software	UOFM	6	P/PU
	D3.4.b	24,30	Crossmodal control	UOFM	13	R/PU
WP4	D4					
T4.1	D4.1.a	24	First Demonstrator	INRIA	1	O/PU
T4.2	D4.1.b	36	Final Demonstrator	INRIA	2	O/RE
	D4.2.b	12	Hardware components	FCTUC	30	P/RE
	D4.2.b	24,36	Software packages	FCTUC	46	P/RE
T4.3	D4.3	12	Experimental platform	UOFM	3	P/RE
T4.4	D4.4	6	Annotated datasets	UOFM	2	O/PU
T4.5	D4.5	36	Experimental platform	UOFM	2	P/PU

Table 4: *The list of project deliverables: WP1, WP2, WP3, and WP4.*

WP	Del. no.	Date	Name	Lead	PM	Type/Diss.
WP5	D5					
T5.1	D5.1	1–36	Web pages	INRIA	2.5	O/PU
T5.3	D5.3	32	Industrial workshop	USFD	0.5	O/PU
T5.6	D5.6a	1,6	Tutorials	USFD	6	O/PU
	D5.6.b	12,24	Thematic schools	USFD	6	O/PU
T5.7	D5.7	6–36	Workshops	USFD	2	O/PU
WP6	D6					
T6.1	D6.1a	12,24,36	Periodic and final reports	INRIA	4	R/RE
T6.1	D6.1b	6,18,30	Interim reports	INRIA	1	R/RE
T6.2	D6.2	12,36	Audit certificates	INRIA	1	O/RE

Table 5: *The list of project deliverables: WP5 and WP6.*

7.6 Workpackage descriptions

WP1: Cognitive mechanisms of attention

WP1	INRIA	UOFM	UKE	FCTUC	USFD	Total
Person-months	25	25	30	12	20	112
Start/End	1/30					
Leader	UKE (partner 3)					

Description of work. The general objective of this workpackage is to study a model of perceptual attention at a fundamental level (behaviroual and cognitive) using the most modern investigation techniques and to propose a computational paradigm for audiovisual attention. More precisely, we will investigate the contribution of major visual and auditory features, interaction of features, cross-modal interactions and the nitration of bottom-up and top-down signals in the control of overt attention. Using the combined expertise in the consortium we will apply psychophysical, electrophysiological, and imaging methods to the same paradigms and use theoretical methods to arrive at a quantitative computational model of this central human cognitive ability.

Task 1.1. Attentional selection in the visual and auditory domain by single features. [INRIA 10PM, **UOFM** 4PM, USFD 10PM]

This task will characterize the contribution of visual (luminance, colour, motion, orientation, spatial frequency and stereoscopic depth) and auditory features in guidance of overt attention. This covers all visual features where strong experimental evidence is available that they contribute to the guidance of attention in humans performing pop-out tasks. In most models of overt attention luminance contrast is a major contribution to saliency. With respect to colour careful calibration and relevant natural stimuli are an important aspect. Although the outstanding importance of motion cues in capturing attention it is well recognized, little data is available. Here, we will use natural motion stimuli and apply the algorithms developed in WP2 for determination of motion vectors. The feature of orientation will be investigated with well established procedures. All these features will be considered at different spatial scales in a resolution hierarchy. Binocular disparity is an important cue for depth estimation and particularly effective at close distances, i.e. in the range where object can be manipulated. A comparison with the selection of fixation points under normal conditions allows separating the contribution of disparity information and pictorial depth cues. Importantly, absolute disparity, disparity contrast and depth information available from the laser scan data will be evaluated and compared. Very little behavioural or computational work has investigated the nature of attention in audition. We will investigate whether attentional processes in the two modalities behave similarly, and what the relevant auditory features are guiding overt attention.

Task 1.2. Multi-feature interactions during attentional tasks. [INRIA 2PM, **UOFM** 4PM, FCTUC 6PM]

The interaction of different features is investigated by information theoretic methods. We will determine the influence of any single feature on the relation of another feature to gaze movements in terms of the conditional probabilities and mutual information. This model free approach facilitates comparison to experimental data (Tasks 1.5 and 1.6) and technical implementations (WP4). We will factor out correlations of features in natural stimuli by computing the partial correlations. Finally,

matching the psychophysical data against the image statistics will lead to a quantitative description of the probability to fixate a region as a function of local features.

Task 1.3. Attention on purpose. [UOFM 4PM, **INRIA** 3PM]

This task investigates the context dependence of overt attention and its relation to conscious perception. Firstly we study the modulation of bottom-up signals to the guidance of overt attention by a set of different task instructions. We have already identified a number of different tasks that result consistent behaviour across subjects. We will compare global statistics of eye movements, such as average saccade latency and length and area covered, and the relation to local features in different contexts. An emphasis will be place on the role of depth cues like binocular disparity in the interaction with task instructions of manipulating objects within reach. This will elucidate the degree of semantic processing before visual information feeds a saliency map. Secondly we exploit the setup for stereoscopic presentation to induce binocular rivalry with simultaneous measurements of eye movements. We will determine the contribution of the suppressed image to the selection of fixation points and compare this with physical superpositions of images. In this way we determine the degree of semantic processing of visual stimuli when contributing to the control of overt attention.

Task 1.4. Multimodal interactions in overt attention. [UOFM 3PM, **UKE** 8PM]

Here we study the integration of simultaneously presented auditory and visual cues for the control of human overt attention. We will present multi-modal stimuli developed in WP4 in congruent, concongruent and isolated conditions. Simultaneously we record eye traces of participants by a high precision optically based system (EyeLink II). Upon combined presentation of auditory and visual cues the spatial distribution of fixation points are expected to be biased to the side of the auditory cue. However, does the pattern of fixation points adhere to the spatial structure of the salient visual stimulus? Or colloquially, when I look to the side of an auditory stimulus, do I look at visually salient regions there? We will compare fits of the data to weighted sums of patterns of fixation points recorded with isolated auditory and visual stimuli, and match this to a multiplicative model. The amount of variance explained leads to a quantitative model of the dominance (purely auditory, $A(x)$, dominated distribution of fixation points and purely visually, $V(x)$, dominated), or integration (weighted sum of auditory and visual patterns ($a_1A(x) + a_2V(x)$)) and multiplicative models ($a_3A(x) * V(x)$). Hence, these experiments allow comparing and evaluating different hypotheses pertaining to the kind of interaction of multi-modal signals (additive, multiplicative, maximum). Integrating the results we will construct the first multimodal model of attention. The combination of auditory cues to new events with visual cues will help to produce a more robust attentional component which is capable of behaviours such as preemptive visual orienting based on unseen but heard auditory cues as well visual identification of novel auditory stimuli.

Task 1.5. Neuronal interactions during overt attention. [UOFM 3PM, **UKE** 10PM]

Here we study the integration of simultaneously presented auditory and visual cues for the control of human overt attention. Attentional effects will be tested using 128-channel EEG recordings or, alternatively 275 channel MEG which is currently being installed by partner UKE. Data analysis will put particular emphasis on so-called induced rhythms, i.e., electric or magnetic activity components which are not phase-locked to stimulus transients.

We will compare two different hypotheses of cross-modal interactions during perception and attentive stimulus processing. First, visual - auditory interaction could occur in a distributed fashion and be mediated by synchronization between modal-specific areas. This leads to the prediction that congruent and incongruent stimulation do not differ fundamentally in regional activation, instead upon congruent multimodal stimulation the coherence between low-level and mid-level modal specific areas should be increased. The second hypothesis predicts a localized integration of sensory information with no major effect of modal specific preprocessing. This would lead to changes in the distribution of cortical activity, but not to large scale changes in interaction between early and midlevel sensory areas. These two opposing predictions can be evaluated by the experimental approach described below.

Auditory and visual stimuli will be implemented using commercially available stimulus-presentation packages. Prior to physiological measurements, perceptual and attentional effects of the stimuli will be tested in behavioural measurements. While in part of the experiments, subjects will view stimuli with central fixation, in later series of measurements we will allow free viewing and perform a saccade-triggered analysis of the data. This is of particular interest when moving from more simple to complex visual stimuli. This will require methodological preparatory steps, since a combined setup allowing EEG recording and simultaneous eye tracking needs to be developed (WP 4). In detail, the experiments planned are the following: Experiment 1: Images of natural objects and natural sounds, both without background; successive presentation of visual and auditory stimuli; subjects have to decide by button-press about semantic congruency or categorize the second stimulus. Experiment 2: Natural objects and sounds appear embedded in complex backgrounds in different spatial locations; subjects have to maintain fixation; covert attention is directed by stimulus saliency (bottom-up) or by spatial cueing (top-down); subjects have to decide by button-press about semantic congruency or spatial congruency. Experiment 3: Natural objects and sounds appear embedded in complex backgrounds in different spatial locations; subjects are allowed to make eye movements (overt attention); otherwise identical to exp. 2. Experiment 4: Cross-modal matching of dynamic moving stimuli; subjects have to attentively evaluate coherence of visual and auditory motion. Data analysis will be carried out jointly with partner UOS and will involve time-frequency-analysis, coherence analysis, autoregressive modelling and information theoretic analysis. The physiological results will be employed to develop (i) behavioural constraints for the robot experiments (task 3.4), (ii) heuristics for computational strategies and algorithms used in the models (tasks 1.4, 2.3-2.5).

Task 1.6. Identification of brain regions involved in multisensory attentional control using fMRI. [UOFM 4PM, **UKE** 6PM]

We will carry out MR measurements to map location and degree of activation of low-level unimodal cortical areas as well as poly-modal association areas. In addition, we will test which sites are undergoing attentional modulation. This will be achieved by replicating experiments 1,2 and 4 as described in task 1.5 using fMRI measurements that are applied to the same subjects that EEG or MEG has been recorded from. This approach will clarify where saliency maps are located in the brain that subserve attentional control, and whether such candidate sites are involved in attentional control irrespective of sensory modality. Moreover, these experiments will be used to guide source localization for the EEG/MEG data obtained in task 1.5. The results will be employed to develop guide the development of computational strategies and algorithms used in the models (tasks 2.3-2.5).

Task 1.7. Computational model of visual and auditory attention [INRIA 10PM, UOFM 3PM, UKE 6PM, FCTUC 6PM, **USFD** 10PM]

Attention serves two roles in perception. First, it is required to direct processing capacity towards new events. Second, it helps to track an ongoing acoustic or visual source such as the utterances of a speaker in a background of competing sources of both visual and auditory nature. This task will develop a computational model capable of both responding to new visual/acoustic stimuli and following existing ones in a mixture. The bottom-up component of the model will utilize an "old+new" principle in determining whether incoming information should be considered as part of an ongoing hypothesis or as the start of a new event. For the top-down component, an existing probabilistic decoder capable of tracking single visual and audio sources in a mixture will be extended to handle attentional focus via the dynamic loading of detailed prior models for the attended source together with a more generic model for the background.

List of milestones.

M12: Neurophysiological models for audio-visual mechanisms of attention based on quantitative modelling of psychophysical experiments.

Associated computational models based on neurophysiological models and psychophysical findigs.

M24: Further validation of the neurophysiological models.

The coupling between recorded psychophysical data and an audiovisual robotic head.

Computational models of attention fully completed.

M30: Final experimental validation and quantitative modelling of audiovisual attention: neurophysiological and computational models are available.

List and content of deliverables.

D1.1: [REPORT] Scientific reports (interim at m12 and final at m24) on bottom-up attention in the visual and auditory domains using single features.

D1.2: [REPORT] Scientific reports (interim at m12 and final at m24) on the interactions between several features during attention.

D1.3: [REPORT] Scientific reports (interim at m12 and final at m24) on top-down attention, or attention on purpose.

D1.4: [REPORT] Scientific reports (interim at m12 and final at m24) on the correlation between audio-visual interactions and overt attention.

D1.5: [REPORT] Scientific reports (interim at m12 and final at m24) on the measurements of neural interactions during overt attention with EEC/MEG.

D1.6: [REPORT] Scientific reports (interim at m12 and final at m24) on the brain representation of multisensory attention.

D1.7: [REPORT] Scientific and technological reports (interim at m12 and m24, and final at m30) on computational models of attention derived from tasks 1.1 to 1.6.

WP2: Integration of visual and auditory cues

WP2	INRIA	UOFM	UKE	FCTUC	USFD	Total
Person-months	30	20	18	6	20	94
Start/End	6/30					
Leader	USFD (partner 5)					

Description of work. Extraction and integration of visual and auditory cues are fundamental building blocks for understanding perception. Cues must be properly extracted from the raw data and their individual merit and reliability must be characterized and quantified. Eventually they must be combined together within a probabilistic framework based on their geometric and temporal properties. We will concentrate on both visual and auditory cues and, moreover, on the interaction of spatial and temporal stimulus features. First, we will study stimuli individually, second we will propose theoretical models for spatial fusion and for temporal fusion, third we will devise a statistical method for sensory integration, and fourth we will study the biological mechanisms of sensory integration. Regarding possible biological mechanisms, this WP aims at testing the hypothesis that coherence of neural signals can change during the formation of cross-modal conjunctions or during cross-modal interference in human subjects. We assume that changes of neural synchrony between areas processing different stimulus modalities will occur in instances of cross-modal feature binding and, moreover, that temporal patterning will be influenced by cross-modal shifts of attention.

Task 2.1. Extraction of visual cues. [**INRIA** 10PM, FCTUC 6PM]

Partner 1 and Partner 4 will make available algorithms and software for extracting visual cues. In particular Partner 1 has recently developed a stereo method that computes a dense depth (disparity) map under the form of a 3D set of surface patches. Partner 4 has recently developed an optical flow method that computes a 2D velocity field in real time. These approaches are embedded into a probabilistic framework such that a reliability measure of the depth and motion maps will be provided as well.

The computing power available today allows, in principle, fast implementations (at 15 frames per second) of these algorithms. Therefore, particular emphasis will be but on achieving and delivering efficient software libraries.

The novel work also to be carried out within this task will be to combine a 3D depth map with a 2D velocity field in order to hypothesize a 3D velocity field. This *instantaneous* depth-velocity temporal/geometric representation, with its associated measure of confidence, will be used to build a primary stimulus-based visual saliency map which in turn will be used as input by tasks 2.3 and 2.4.

Task 2.2. Extraction of auditory cues. [**USFD** 8PM]

Starting from an existing model of peripheral auditory processing, the goal of this task is to form a description of the incoming binaural signals in terms of dominant spectro-temporal elements tagged with their spatial position. Interaural time and level differences will be used to localise the acoustic energy at a given time and frequency. However, these cues are unreliable in natural reverberant conditions, and robust estimates can only be achieved by integrating across time and frequency. Cues such as onset time and harmonicity will be used to group energy into larger time-frequency regions which can be robustly localized. Prior knowledge of the location of sources, gained by tracking the sources in the audio-visual domain, will be fed back into the primitive processing stages to reduce

ambiguity and improve the clarity of acoustic event descriptions. Attention processing and sensory-motor coordination (described in WPs 1 and 3) will ensure that for much of the time the sound source of interest will be at a known location with respect to the head (typically at an azimuth of 0 degrees). Active audiovisual strategies to improve the acoustic signal-to-noise ratio for the sound source of interest will feed into this task.

Task 2.3. Geometric and probabilistic fusion of spatial visual and auditory cues. [INRIA 4PM, **USFD** 4PM]

The goal of this task will be to represent both visual and auditory cues in a common sensor-centered reference frame and to hypothesize associations between visual events (such as a moving object) and auditory events (such as the presence of a sound source). This will consist in an optimal estimation (using the maximum likelihood estimator) of the geometric transformation between the 3D visual space and the 3D auditory space, given the initial estimates provided by Tasks 2.1 and 2.2. This will allow to associate sound sources with a visual depth map and to predict spatial locations where both motion cues and auditory cues are likely to be simultaneously present.

Task 2.4. Synchronization and fusion of temporal visual and auditory cues. [INRIA 4PM, USFD 2PM, **UKE** 6PM]

This task will take input from tasks 2.1, 2.2, and 2.3 and will take from granted the fact that visual and auditory information has been spatially integrated. The goal of this task is to investigate methods and algorithms allowing to correlate temporal auditory events with visual motion detection. If several sources of sound are present in the scene, then the task of grouping auditory events emitted by the same source into a coherent temporal auditory signal (such as a speech signal) is a tremendously difficult problem. On one side, there is evidence from neurobiological data and from psychophysical experiments that *auditory attention* can be positively biased by visual motion detection. On the other side, the visual information is evenly and continuously distributed in space and it is difficult to decide whether a visual object is more relevant than another visual object; this is therefore a case where localization and identification of auditory sources may help vision.

It is therefore crucial to develop a formalism allowing the synchronization between time-varying motion events and auditory events. The output of this task will be an audiovisual "tracker" – able to provide the 3D trajectory of an audiovisual object.

Task 2.5. Development of statistical methods for cross-modal integration. [**INRIA** 12PM, UKE 6PM, UOFM 8PM, USFD 6PM]

An appealing approach for cross-modal integration could be to rely on statistical methods only. This has been a successful approach in speech recognition where 1D hidden Markov models (HMM) have extensively been used to associate sound signals to identities of phonemes. 2D HMMs have also been applied, with some success, to visual data in order to associate "labels" and visual cues, where labels may refer to object parts, human gestures, facial expressions, etc. Researchers in computational neuroscience believe that statistics underlie many mental processes, such as perception, thinking, and acting.

In more detail, cross-modal integration can be achieved by modelling the correlated dynamics of the acoustic and visual features of an event. An approach based on factorial HMMs, developed for

modelling acoustic mixtures, will be extended. Here, an independent HMM is employed to model the energy contribution of each sound source. By employing primitive processes to segment the time-frequency plane into a small number of source fragments (task 2.2), the decoding search space can be controlled. These techniques have been demonstrated in simultaneous-speaker speech recognition tasks and work well when the speakers are acoustically distinct. However, performance may be greatly enhanced if the acoustic models are supplemented with features from the visual modality. The associations between modalities that are learnt by training on isolated speakers mean that the acoustic fragments are bound to the source with the matching visual model. Although techniques for the visual parameterization of speech are well developed, it is less clear how other sounds can be described visually. One of the challenges in this task is to generalize to non-speech events.

These developments should lead to such paradigms as *seeing a sound* which form the bases of the interactions that exist between visual attention and auditory scene analysis.

Task 2.6. Biological mechanisms for multi-sensory integration. [**UOFM** 12PM, UKE 6PM]

Objective of this task is to identify biological mechanisms underlying multisensory integration and to derive relevant algorithmic principles than will be used in the technical implementation. Cross-modal integration will be investigated at the cellular level using multi-electrode recordings in anesthetized animals. Ferrets will serve as a model system. The experiments will address putative saliency map sites, corresponding to the superior colliculus (SC) and the anterior ectosylvian cortex (AEV), a multi-modal cortical area. Multi-electrode recordings with visual-auditory stimuli will be performed to characterize neuronal receptive fields, response properties and dynamic neural interactions in these putative saliency maps. The stimulation paradigms will include uni-modal stimulus presentations as well as cross-modal settings where the spatial and temporal relation between the visual and auditory stimulus are varied. Properties of the multi-modal integration process will be investigated with information-theoretic methods. As a baseline, we will use correlation analysis and autoregressive models that are linear methods. As a next step we use non-linear methods including entropy and mutual information. These will be applied to the physiological data and the signals of both modalities. Importantly, we will partition the information content with respect to rates, signal correlations and noise correlations. These measures allow a characterization of the integration of information while keeping the modality specific information accessible. The results will be employed to guide the development of computational strategies and algorithms used in the models (Tasks 2.3–2.5).

List of milestones.

M18: Elaboration of methods and algorithms for cross-modal integration (audio and video) based on stimulus (sound and light) processing.

M30: Elaboration of an abstract model for cross-modal integration based on statistical methods.

List and content of deliverables.

D2.1a: [PROTOTYPE] Algortihms and software for extracting visual cues (interim and final prototypes).

D2.1b: [REPORT] Description of models and methods used in D2.1.a

D2.2a: [PROTOTYPE] Algortihms and software for extracting auditory cues (interim and final prototypes).

D2.2b: [REPORT] Description of models and methods used in D2.2.a

D2.3a: [PROTOTYPE] Algortihms and software for spatial fusion of auditory and visual cues (interim and final prototypes).

D2.3b: [REPORT] Description of models and methods used in D2.3.a

D2.4a: [PROTOTYPE] Algortihms and software for temporal fusion of auditory and visual cues (interim and final prototypes).

D2.4b: [REPORT] Description of models and methods used in D2.4.a

D2.5a: [PROTOTYPE] Algortihms and software for cross-modal integration using statistical methods (interim and final prototypes).

D2.5b: [REPORT] Description of models and methods used in D2.5.a

D2.6: [REPORT] Biological models for multi-sensory integration and their computational modelling.

WP3: Sensory-motor coordination

WP3	INRIA	UOFM	UKE	FCTUC	USFD	Total
Person-months	-	25	3	24	20	72
Start/End	6/30					
Leader	UOFM (partner 2)					

Description of work. The goal of this work-package is the design and development of principles, methods and algorithms to allow the coordination of the motor activities with the sensor measurements. The rationale is to be able to direct attention towards scene areas with activity relevant to the task at hand. In this project we deal with two sensor modalities, namely vision and hearing. The control of the positions and orientations/poses of the robotic device will be performed using the results of work-packages 1 and 2. The information used to locate the focus of attention will be extracted from both sounds and videos. Since the motion of the cameras and microphones is known, it will be used to predict the positions and orientations of the relevant scene elements in the images. Such predictions are an essential element in processes such as smooth pursuit. Prediction in this type of systems can be regarded as performing a kind of information feed-forward. Active listening will be used to improve the robustness of the motor coordination. Gaze and oculomotor control will be performed using saccades, smooth pursuit and vergence. As a result of the work developed in this work-package the robotic device will be able to respond to both the visual and the auditory stimuli. Coordination is obtained by synchronizing the image and sound measurements with the motor control.

Task 3.1. Sensor, geometric, and dynamic calibration. [**FCTUC** 6PM]

To perform the control of the robotic device the sensors have to be calibrated. That means the estimation of the intrinsic parameters of the cameras and their relative positions and orientations. The control of the active system requires the knowledge of the inverse kinematics. The estimation of the inverse kinematics implies the geometric calibration of all the system. The overall system has also to be dynamically calibrated so that the estimates for the overall bandwidth can be computed. The outputs of this task are: –Software/algorithms for geometric calibration of the cameras (including intrinsic parameters and relative pose) and –The inverse kinematics of the robotic device;

Task 3.2. Oculomotor control. [**UOFM** 13PM, FCTUC 6PM]

The oculomotor control will be performed using outputs of work-package 2. Feedback and predictive feedforward control will be used. The motions of the audiovisual head will be made up of three visual behaviours: –Saccadic motions; these are fast responses to the detection of an event and are performed in open loop, i.e., visual information will not be used to control motion. At low-level the motor bandwidths will be used to define maximum speeds and accelerations; –Smooth pursuit; these are motions that use visual measurements to define the positions and velocities errors on a feedback loop. On the other hand to compensate for the delays due to the visual processing predictive control will be used to update the positions and velocities of the motors; –Vergence; the goal of this behaviour is to have the event or element that is the source of attention located in the center of both images in a binocular active system. Position (stereo) and velocity (motion) disparities will be used to guarantee vergence on the same visual element. Therefore control is performed using both measured positions and estimated velocities.

The outputs of this task will be the implementations of the visual behaviours described above.

Task 3.3. Audio-motor control. [UKE 2PM, FCTUC 6PM, **USFD** 20PM]

Active listening has enormous potential to improve the robustness of both auditory and visual scene analysis. Within the constraints imposed by the active audiovisual head, this task will investigate, both behaviourally and computationally, a range of hypothesized active listening strategies. Possible strategies include: moving to a location which maximizes the signal-to-noise ratio (SNR) (e.g. moving towards a source of interest); orienting the head to a position which maximizes SNR (e.g. using head shadow to attenuate the strongest competing source, or moving the head to face the source in order to reduce ambiguities in cues to location); moving to get an un-occluded view of the target source (e.g. for speech reading); and moving away from walls and hard surfaces to reduce the effect of reverberation. Since relatively little is known about the strategies used by humans to improve the identifiability of a target in noise, the behaviour of participants in adverse environments such as crowded meetings will be observed and annotated in an attempt to better understand the types of active strategies used and the situations in which they are employed. Findings from this investigation will feed into algorithm development for active listening which in turn will be implemented and demonstrated in the active audiovisual head.

The outputs of this task will be the development and implementations of: –active listening strategies; –a set of auditory behaviours for directing the head towards the source of interest;

Task 3.4. Cross-modal motor control. [**UOFM** 12PM, UKE 1PM, FCTUC 6PM]

In this task the results of work-packages 1 and 2 as well as the results of tasks 3.1, 3.2, and 3.3 will be used to develop methods for motor control using information from both vision and audition. In this task the control algorithms will be designed taking into account the specific nature of the cross-modal measurements. Problems to be dealt with in this task are:

- Identification of the behaviours for actively directing the attention of the robotic device using both vision and audition. These behaviours can be based on the visual behaviours adaptively changed to use audition. Attention can be directed to a visually occluded target as well as to a silent target. When measurements from both modalities occur and are associated to the events feedback control will use the measurements estimated from both modalities;
- Development of the feedback control whose task is to direct the attention towards the positions and velocities of selected world events.
- Feedforward strategies to account for processing delays. Since these delays will be highly variable adaptive predictive techniques will be used;

List of milestones.

M18: Elaboration of methods and algorithms for occulo-motor control and for audio-motor control.

M30: Elaboration of methods for cross-modal motor control.

List and content of deliverables.

D3.1: [PROTOTYPE] Software for audiovisual head calibration.

D3.2.a: [PROTOTYPE] Algorithms and software for occulo-motor control (interim and final prototypes).

D3.2.b: [REPORT] Description of models and methods used in D3.2.a.

D3.3.a: [PROTOTYPE] Algorithms and software for audio-motor control (interim and final prototypes).

D3.3.b: [REPORT] Description of models and methods used in D3.3.a.

D3.4.a: [PROTOTYPE] Algorithms and software for cross-modal motor control (interim and final prototypes).

D3.4.a: [REPORT] Description of models and methods used in D3.4.a.

WP4: Development of methodological and experimental platforms

WP4	INRIA	UOFM	UKE	FCTUC	USFD	Total
Person-months	22	5	3	36	20	86
Start/End	1/36					
Leader	FCTUC (partner 4)					

Description of work. This work-package will integrate the theoretical findings and methodological results of tasks in work-packages 1, 2, and 3 and will build a proof-of-concept demonstrator within a well-specified application scenario. There will be an audiovisual robot-head prototype fully designed, developed, and built by the partners. This prototype will integrate both neurophysiological findings (tested on special-purpose platforms – tasks 1.1-1.6, 2.6, and 3.4) and computational developments (tasks 1.7, 2.1-2.5, and 3.1-3.3). We follow a twofold approach integrating both engineering (methods and algorithms) and neurophysiological and psychopysical models and experiments. On one hand we attempt to quantify and formalize (from a computational point of view) the experimental results (tasks 1.5, 2.6, and 3.4), so that they are suitable to be implemented in an artificial system. Next we subject the existing (or the appropriate successor at that time) control system of the artificial sensory system to the same type of experiment as the human subjects, and compare performance. We will analyse to what extent feedforward predictive and feedback control contribute to system performance. An important breakthrough will be to demonstrate a dynamic system rather than a static one. A mid-term (month 24) and a final demonstrator (month 36) will be developed.

Task 4.1. Mid-term and final demonstrators. [**INRIA** 1PM, FCTUC 1PM, USFD 1PM]

Under this task we will specify in detail the interim and final project demonstrators based on WP1, WP2, and WP3 and which will be implemented onto the hardware and software platforms described below. First we will conduct a survey in order to target the set of applications that are likely to take advantage of the POP's results. The core of the demonstrator will show the behaviour of the audio-visual head in a situation like a crowd of people where the POP system should be able to associated (bottom up) auditory and visual cues and locate one or several speakers. Next, the head should be able to concentrate its attention (top down) to one speaker and follow his body and head motions. An ambitious goal would be to enhance automatic speech recognition in these difficult conditions.

Task 4.2. Development of an audio-visual robot platform. [INRIA 21PM, UOFM 1PM, **FCTUC** 35PM, USFD 19PM]

The platform will be composed of hardware – an audio-visual "head" (cameras, microphones, and actuators) and of software – developed within WP2 and WP3 and combined with existing software libraries and packages from the POP partners. The hardware (audio-visual head) will be made up of two pan&tilt units in a stereo configuration. Each pan&tilt unit will have two rotational degrees of freedom. In each unit the degrees of freedom will move a camera and a microphone. The actuators of the pan&tilt units are DC motors. Position and velocity feedback will be obtained by means of encoders. A general purpose control board will be used on a PC to control the motion of the audio-visual head. Low level control software will be developed allowing the control of all the degrees of freedom. The audiovisual head will be built by FCTUC (partner 4) and the technical specifications will be provided to partner 1 such that it can easily duplicate the prototype. USFD (partner 5) will specify the auditory equipment and will provide the software packages necessary to operate this

equipment. INRIA (partner 1) will specify the visual equipment and will provide the corresponding software packages. UOFM (partner 2) and UKE (partner 3) will provide specifications for the type of data sets that such a platform should provide for their experiments (see task 4.5 below).

The performance of the platform will be characterized by means of a set of demonstrations. In one of the cases a single source of attention (with both visual and auditory content) will be used to test the bandwidth of the robot unit by using the encoders measurements. To test the response to a complex set of stimuli the audio-visual head will be tested in a situation where mutiple people are present, along scenarios defined in Task 4.1.

Task 4.3. Methodological platform for investigation of biological mechanisms. [**UOFM** 2PM, UKE 1PM]

In this task the techniques for simultaneous eye tracking and EEG recording are developed. We will modify the mechanical setup of the head-mounted eye tracking system and apply electrical shielding to make it compatible with multichannel EEG recordings. As eye movements induce large signals into the EEG recordings, in a second equally important step techniques have to be developed to separate these from the EEG signals proper. These will be based on on-line feedback of the tracked eye-position to the electrophysiological recordings, independent component analysis and time-delayed auto regressive models. This task is a prerequisite for the experiments carried out in task 1.5.

Task 4.4. Development of a stimulus database for multisensory studies in electrophysiology, psychophysics and robots. [**UOFM** 1PM, UKE 1PM]

The stimulus materials used in WPs 1 and 2 comprise photographs, videos and sounds of isolated objects and naturalistic scenes. Important technical aspects are faithful color calibration, inclusion of disparity information by stereoscopic images and validation by laser range scans. For the later experiments are ratings of the emotional valence, association with different possible actions and subjective similarity for each record in the database, and the option to use congruent and incongruent multi-modal stimuli. "Normalized" stimuli are matched in 2nd order statistics, but retain differences in higher order structure. Development of the stimuli is a prerequisite for the experiments carried out in tasks 1.1-1.6, 2.1-2.2, and 3.3-3.4.

Task 4.5. Development of a platform for testing biologically-based algorithms. [UKE 1PM, **UOFM** 1PM]

The models and algorithms hypothesized as a result of the psychophysics and neurophysiology will be tested with the audiovisual head (task 4.2). For that purpose the robot platform will be able to respond to sequences of images and sounds recorded through psychophysical experiments. These experiments using the robot platform will be conducted jointly by partners 1, 2, 4, and 5. This includes, as described in task 3.1, sensor geometric and dynamic calibration. Furthermore, the recorded data will be input to the system and the system response (the motor encoders) will be recorded. As a result, the effects of the algorithms acting via the robotic platform can be evaluated and compared to the psychophysical data. A similar problem arises in the joint aquisition of stereoscopic images and laser scanning data. Here we will strive for a joint software platform for interoperability of robot head, stereo cameras and laser scanner.

List of milestones.

M12: First prototype (hardware) of the audiovisual head.

M24: Second prototype (hardware and software) of the audiovisual head.

First POP demonstrator.

M36: Final prototype of the audiovisual head.

Final POP demonstrators.

List and content of deliverables.

D4.1a: [DEMONSTRATOR] The first POP demonstrator.

D4.1b: [DEMONSTRATOR] The second POP demonstrator.

D4.2a: [PROTOTYPE] The hardware components of the audiovisual POP head.

D4.2b: [PROTOTYPE] The software components of the audiovisual POP head (first and final demonstrators).

D4.3: [PROTOTYPE] Platform for simultaneous eye-tracking and EEG recording.

D4.4: [DATABASE] A stimulus database for studying cross-modal integration with humans and robots.

D4.5: [PROTOTYPE] Experimental platform for testing the robot head using psychophysical data.

WP5: Exploitation, training, and dissemination of results

WP5	INRIA	UOFM	UKE	FCTUC	USFD	Total
Person-months	3	3	3	3	5	17
Start/End	1/36					
Leader	USFD (partner 5)					

Description of work. This work-package implements the exploitation, training, and dissemination activities through the following tasks.

Task 5.1. Creation and maintenance of a website.

Task 5.2. Consortium publication activities.

The exploitation-dissemination manager will coordinate the publication efforts of the consortium by making sure that the scientific results go out for publication and that papers are submitted to appropriate journal and conferences. The work-package 5 manager will also have the responsibility to verify that the material to be published is consistent with the IPR policy as agreed by the consortium.

Task 5.3. Industrial liaison.

This task will coordinate the transfer of knowledge and of technology from the partners towards companies that manifest their interest in the projects' outcomes. The work-package manager will make sure that these partner/company relationships are in line with the management of knowledge and intellectual property rights.

An industrial workshop (at month 32) will be organized in order to show the demonstrators to potential users.

Task 5.4. Links with other projects and consortia.

The dissemination and training activities will not be carried out in isolation but in collaboration and coordination with other EC and national projects, such as:

- The coordinator (partner 1, INRIA) is currently involved in a Marie Curie Early Stage Training action (VISITOR, MEST-CT-2004-008270). VISITOR is a single-partner action that started in December 2004 for 4 years and that is coordinated by the INRIA Rhone-Alpes computer vision group involved in POP. Within EST VISITOR there will be 8 full-time PhD students and 8-10 PhD visiting students (for 3-6 month periods of time). VISITOR's objectives and topics of research (perception, learning, and autonomy) are strongly related to the POP project.
- The Research Training Network action VISIONTRAIN (MRTN-CT-2004-005439), coordinated by INRIA (Radu Horaud) and that involves 11 partners from 11 countries, started in May 2005 for 4 years. RTN VISIONTRAIN's objectives and topics of research are to establish cognitive and computational models for vision systems, and therefore VISIONTRAIN is fully relevant to the POP project.

- The University Hospital Hamburg-Eppendorf (partner 3) is involved in coordinating the EU Network of Excellence Neuro-IT (IST-2001-35498), which provides a European platform for coordination and dissemination of research in the field of neuroengineering, biorobotics and computational neuroscience. Andreas Engel is acting as a steering committee member in the board of Neuro-IT.

- The University of Sheffield (partner 5) participates to the AMI integrated project (Augmented Multi-Party Interaction) that started in January 2004 for a duration of 4 years. The University of Sheffield's speech and hearing group is the AMI training coordinator. AMI is composed of 9 academic partners and 5 industrial partners from 7 European countries.

- The POP partners commit to participate to the activities likely to be organized by the euCognition coordinating action.

The work-package 5 manager will contact these projects and consortiae, will organize meetings whenever appropriate, and will establish bi-lateral and multi-lateral formal activities such as workshops, tutorials, seminars, and thematic schools.

Task 5.5. Exploitation.

The POP researchers and engineers will develop software, hardware, an integrated robot platform, as well as various laboratory devices for carrying out psychophysical experiments. Proper exploitation and dissemination of these outcomes needs protection of intellectual property rights. The latters will be effective through licensing and patenting and these activities will be coordinated by the exploitation-dissemination manager in coordination with the steering committee and with legal offices of each one of the partners.

Task 5.6. Training.

The POP double-cultural topics of research will be attractive to talented postdoctoral researchers. However, very few will possess the required interdisciplinary skills from the outset. Consequently, all project researchers will be given the opportunity for cross-training in disciplines in which project partners have expertise. They will be encouraged to develop the technical and collaborative skills needed to solve complex problems. A training programme with the following elements will be implemented:

- Postdoctoral researchers will be expected to spend a significant proportion of the first 18 months outside their nominal home institution in order to transfer their expertise and resources to other partners, and to complement and broaden their existing skills set.

- Doctoral students will have the opportunity to spend part of their time in a partner institution and to attend the thematic schools organized by the POP partners.

- Senior scientists involved in POP will be encouraged to spend periods ranging from short visits to longer sabbaticals in other partner institutions.

- A series of tutorials will be organised and delivered at the internal and dissemination workshops that are planned. Tutorials will cover foundational elements of the discipline (e.g. auditory scene analysis, visual attention, active stereo) and will offer training in relevant hardware (e.g. active vision head) and software.

In particular there will be 2 thematic schools (at months 12 and 24) that will be organized by the POP coordinator in conjunction with the thematic schools planned within the RTN action VISION-TRAIN.

Task 5.7. Workshops

The work-package 5 manager will coordinate the organization of a number of workshops: 3 internal workshops (at months 6, 18, and 24) and 3 workshops in collaboration with other European projects (at months 12, 24, and 36).

WP6: Management

WP6	INRIA	UOFM	UKE	FCTUC	USFD	Total
Person-months	3	3	3	3	3	15
Start/End	1/36					
Leader	INRIA (partner 1)					

Task 6.1 Communication with the EC.

Collect, monitor and integrate all the technical, administrative, and finacial data from the partners and prepare appropriate documents for the European Commission: management reports, progress reports, final report, cost and financial statements, deliverables, etc.

Task 6.2. Organisation of EC reviews and audits.

The project coordinator will be responsible for organizing the audits concerning all the aspects of the project: technical audits including the annual project reviews, financial audits, as well as any other audits that the EC wishes to organize.

Task 6.3. Organisation and preparation of internal meetings.

The project coordinator, in coordination with the site managers and with the work-package managers will prepare and organize the Steering Committee meetings and the technical meetings.

Task 6.4. Legal, financial, and administrative management.

The project coordinator will be responsible of this task: receive payments from the EC, transfer payement to the partners, prepare the consortium agreement, obtain the audit certificates when required, etc.

Task 6.5. Technical management.

The work-package managers, supervised by the project coordinator will be in charge of the monitoring, the coordinating, and the controlling of the scientific and technical progress of the project. They will be responsible for preparing the technical deliverables of the project as well as of the annual technical reports and final report of the project.

8 Project resources and budget overview

8.1 Efforts for the project

Work-package	INRIA (1)	UOFM (2)	UKE (3)	FCTUC (4)	USFD (5)	Total
Research & innov.						
WP1	25	25	30	12	20	112
WP2	30	20	18	6	20	94
WP3	-	25	3	24	20	72
WP4	22	5	3	36	20	86
WP5	3	3	3	3	5	17
Sub-total 1	**80**	**78**	**57**	**81**	**85**	**381**
Management						
WP6	3	3	3	3	3	15
Sub-total 2	**3**	**3**	**3**	**3**	**3**	**15**
Total per part.	**83**	**81**	**60**	**84**	**88**	
Overall total						**396**

Table 6: *The break-down of efforts (in person months) for each partner and for each work-package.*

8.2 Overall budget for the project

Forms A3.1 and A3.2 from the CPF are included at the end of the document.

8.3 Management level description of resources and budget

The table below indicates the own resources for the AC contractors expressed in person-months.

AC contractors	UOFM (2)	UKE (3)	FCTUC (4)	USFD (5)	Total
Person-months	12	18	27	18	75

Table 7: *Own resources for the AC contractors.*

The major costs of the project are the followings:

- **Personnel** costs will represent 72% of the requested EC funding. POP will hire 2 PhD researchers and 6 post-doc researchers for the whole duration of the project. In addition, there will be 17 person-months directly allocated to the exploitation and training activities.
- **Consumables** will represent 6% of the requested EC funding. These resources will be used for the development of the robotic platform, for interfacing the audio and video sensors with the

computers and with the actuators, and for fMRI sessions (subject costs, maintainance, micro-electrodes).

- **Durable equipment** will represent 11% of the requested EC funding. These resources will be used for the development of an audio-visual head, the acquisition of a mobile robot, the acquisition of an eye tracker for use with an existing EEG and for psychophysics experiments, and computers.

- **Travel** will represent 5.3% of the requested budget. These resources will be used for cross-visits between the partners, project meetings and audits, participation to thematic schools and to training activities organized by other EC projects, and for attending workshops and international conferences.

9 Ethical issues

In-vivo data on neuronal dynamics will be used to develop algorithms for interface programming and robot control. Studying the interactions between neurons processing visual and auditory information will provide critical insights into mechanisms underlying multisensory integration. WP 2 (Integration of visual and auditory cues) will conduct in-vivo experiments with anesthetized ferrets. These experiments will be conducted by partner 3 (UKE). All experimental procedures will be conducted in accordance with the latest revised version (12 April 2001) of the German Animal Protection Law. Experiments will comply fully with European Community guidelines (EUVD 86/609/EEC) regulating the care and use of laboratory animals. Every in-vivo experiment will be submitted for approval to the local authority (Regierungspräsidium Hamburg) who will assess (i) whether the anticipated benefits justify the use of animals; (ii) the number of animals used in each experiment; (iii) the procedures adopted to ensure that animal suffering is minimised. Most experiments will be performed using procedures that are similar to those already in use at University Hospital Hamburg-Eppendorf, and that have already been approved by the local authorities in Hamburg. We certify that we will inform the EC of local authority approval before the start of the in-vivo experiments.

As part of WP1 and WP4, measurements are performed on human subjects using EEG, fMRI and psychophysical methods. All techniques are non-invasive and do not involve any risk for participants. All participants are volunteers. They are informed about the goals and procedures of the study and give their written consent. Participants always have the possibility to terminate the experiment without given reasons. All studies are performed in accordance with the ethical standards laid down in the 1964 declaration of Helsinki (most recent pass: Edinburgh, Scotland, October 2000).

A Consortium description

A.1 Participants and consortium

Partner 1: INRIA

Contribution to the POP project. There will be two research groups involved in the POP project – the computer vision group and the statistics group. The computer vision group will provide the scientific and technological expertise needed to integrate various vision modules into a working system, i.e., everything from camera acquisition and camera calibration software, to methods for recovering depth using stereo, motion segmentation methods, and object tracking methods. The statistics group will provide theoretical insights for the development of robust statistical methods. These methods will be used for the 3-D reconstruction and tracking algorithms that must operate in the presence of noisy, bad, or partially missing data.

The expertise on real-time 3-D reconstruction of humans from image contours will be used within the context of the active stereo sensor to be developed by the POP partners. The immense expertise on camera calibration using generic camera models will be used for the calibration of the ative stereo head.

Both the statistics and the vision groups will contribute to the development of a theoretical and methodological model of perceptual attention, in particular they will investigate together the most robust and efficient way to model interactions between high-level knowledge and low-level sensorial information. The statistics group is particularly competent in the domain of Bayesian decision theory and therefore will have a major contribution to model the integration of auditory and visual cues.

Our experimental and development platform/laboratory will be made available to POP researchers. One in-house development engineer (Hervé Mathieu, member of the vision group) will supervise the software development activities of POP in order to integrate and build the POP's experimental platform.

Personnel involved in POP.

Florence Forbes: 8PM. http://mistis.inrialpes.fr/index.html

Radu Horaud. 7PM. http://perception.inrialpes.fr/

Frédéric Devernay. 8PM.

Emmanuel Prados. 8PM.

Partner 2: University of Osnabrück

Contribution to the POP project. The partner at the Institute of Cognitive Science at the University of Osnabrück will contribute psychophysical experiments on human overt attention under natural conditions; advanced tools for data analysis; develop models of sensory processing and sensori-motor coupling in the mammalian brain and interface with the development of behaving artefacts. Firstly, we will study the contribution of low-level features and high-level tasks to the guide of human overt attention under natural conditions. Controlled modifications of stimuli allow dissecting purely correlative effects from causal influences that can be described by quantitative mechanistic models. This is complemented by an investigation of effects of context and task.

The investigation of the interaction of such top-down effects with the bottom-up signals forms the second contribution. It will be based on advanced mathematical tools like auto-regressive techniques and an information theoretic analysis. The results will be matched to an implementation suitable to guide a behaving artefact.

Thirdly, the adaptation of receptive field properties to the properties of natural stimuli complements the former studies. Using techniques of unsupervised learning over the space of natural visual and auditory stimuli we study the emergence of cross-modal integration.

Personnel involved in POP.

Peter König. 6PM. http://www.cogsci.uni-osnabrueck.de/ NBP/peterhome.html

Selim Onat. 6PM.

Partner 3: The University Hospital Hamburg-Eppendorf

Contributions to the POP project. Within POP, the UKE research team will contribute in the area of neurophysiology and carry out experimental work on visual-auditory interaction and the relation of such interactions to attention and scene segmentation. Experiments will be carried out using EEG and fMRI in humans. Complementing the human experiments, microelectrode recordings will be performed in animals, using the ferret as a model system to study cellular mechanisms of visual-auditory interactions. In addition, the UKE team will contribute theoretical knowledge from the field of neurophysiology and neurobiology to the modelling and robotics work of the other POP partners.

Personnel involved in POP.

Andreas K. Engel. 3PM. http://www.40hz.net/index.html

Andrej Kral. 5PM.

Stefan Debener. 5PM.

Gerhard Engler. 5PM.

Partner 4: FCTUC (University of Coimbra)

Contributions to the POP project. FCTUC will contribute to several of the scientific topics of the project and, in particular, to the aspects related to sensory-motor coordination and design of the robotic devices. The Computer Vision Lab at FCTUC has developed several active vision systems, two of which are currently being used to study perception-action coordination in Robotics. The Computer Vision Lab has made several relevant contributions in active vision namely be developing visual control methods based on optical flow and by integrating them with mobile robots. The systems that will be developed within the framework of this project will be active and will be used to develop and test models of focus of attention. That will imply that the mechanical devices will be oriented towards the features deemed to be sources of attention. The scientific problems to be solved involve the integration and coordination of the visual and auditory measurements and the control of the mechanical devices. Since the mechanical devices have multiple degrees of freedom they have to be synchronized and coordinated so that they generate the expected behaviour. The FCTUC team will contribute to the design and development of the robotic devices and also to the algorithms to coordinate the motions and activities of the mechanical systems.

In addition FCTUC will contribute to the design and development of algorithms for extracting the visual cues, and to their evaluation as well as to aspects related to the integration and fusion of the visual and auditory cues. The processes and algorithms for visual and auditory information extraction will be tightly coupled to the control algorithms.

FCTUC will also contribute to the development of a platform enabling the test of the biological models and algorithms on the robotic devices. This platform will allow the use, on the robotic devices, of the images and sequences of images tested in the psychophysical experiments. The mechanical response of the mechanism will be recorded and as a result the performance of the biological models and algorithms can be evaluated.

Personnel involved in POP.

Helder Araújo. 6PM. http://www.usr.uc.pt/ helder

Joao Barreto. 8PM. http://www.usr.uc.pt/ jpbar

Paulo Peixoto. 10PM. http://www.usr.uc.pt/ peixoto

Jorge Batista. 9PM. http://www.usr.uc.pt/ batista

Partner 5: University of Sheffield

Contributions to the POP project. The University of Sheffield will contribute expertise in speech, hearing and audiovisual speech processing, both at the level of computational modelling and behavioural experiments. Drawing on an existing collection of software tools, audio and audiovisual corpora, Sheffield will lead the development of algorithms for acoustic cue extraction in noise and robust speech decoding in the presence of multiple talkers, and models of auditory attention. Sheffield will contribute to work on multimodal integration of auditory and visual processes at all levels of the project, including the design of behavioural experiments, development of the theory, algorithmic implementation and integration with hardware and software for an active audiovisual head.

Sheffield will lead a detailed comparison of auditory and visual attention and produce recommendations for modelling work on both auditory attention and multimodal attentional integration. Sheffield will also extend existing algorithms for cue extraction and integrate them with visual cues in a principled fashion. A multi-source decoder, which allows bottom-up and top-down information to be combined within a single probabilistic framework, will be augmented to deal with visual inputs. Sheffield will work on active listening strategies for improving the robustness of machine cognition, and contribute to algorithms which combine auditory and visual information in a synergistic manner.

Personnel involved in POP.

Martin Cooke. 9PM. http://www.dcs.shef.ac.uk/ martin/

Jon Barker. 9PM. http://www.dcs.shef.ac.uk/ jon/

References

[1] Benucci A, Verschure PFMJ, and König P. On the existence of high-order correlations in cortical activity. *Phys Rev E Stat Nonlin Soft Matter Phys.*, 68:041905, 2003.

[2] Tamar Avraham and Michael Lindenbaum. Inherent limitations of visual search and the role of inner-scene similarity. In *Workshop on Attention and Performance in Computational Vision*, pages 16–28, 2004.

[3] J.P. Barker, M.P. Cooke, and Ellis. D.P.W. Decoding speech in the presence of other sources. *Speech Communiation*, (45):5–25, 2005.

[4] J. Barreto and H. Araújo. A general framework for the selection of world coordinate systems in perspective and catadioptric imaging applications. *Int. Journal of Computer Vision*, 57(1):23–47, 2004.

[5] J. Barreto, J. Batista, and H. Araújo. Model predictive control to improve visual control of motion: Application in active tracking of moving targets. In *ICPR'2000*-15^{th} *Int. Conf. on Pattern Recognition*, Barcelona-Spain, September 2000.

[6] J. Barreto, J. Batista, H. Araújo, and A. Almeida. Control issues to improve visual control of motion: Applications in active tracking of moving objects. In *Proc. of AMC'2000*-6^{th} *Int. Workshop on Advanced Motion Control*, pages 13–18, Nagoya-Japan, March-April 2000.

[7] J. Barreto, P. Peixoto, J. Batista, and H. Araújo. Control performance issues in a binocular active vision system. In *Proc. of IROS'98–IEEE/RSJ Int. Conf. on Intelligent Robot and Systems*, pages 886–891. IEEE Press, 1998.

[8] J. Barreto, P. Peixoto, J. Batista, and H. Araújo. Improving 3d active visual tracking. In *ICVS99–First Int. Conf. on Computer Vision Systems*, pages 412–431, 1999.

[9] J. Barreto, P. Peixoto, J. Batista, and H. Araújo. Tracking multiple objects in 3d. In *IROS'99-IEEE/RSJ International Conference on Intelligent Robots and Systems*, Kyongju, Korea, October 17–21 1999.

[10] J. Barreto, P. Peixoto, J. Batista, and H. Araújo. Evaluation of the robustness of visual behaviors through performance characterization. In M. Vincze and G. Hager, editors, *Robust Vision for Vision-Based Control of Motion*, chapter 12, pages 145–161. IEEE Press, 2000.

[11] J. Batista, P. P. Peixoto, and H. Araújo. Real-time vergence and binocular gaze control. In *IROS97–IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Grenoble, France, September 1997.

[12] J. Batista, P. Peixoto, and H. Araújo. Real-time visual behaviors with a binocular active vision system. In *ICRA97–IEEE Int. Conf. on Robotics and Automation*, New Mexico, USA, April 1997.

[13] J. Batista, P. Peixoto, and H. Araújo. Visual behaviors for real-time control of a binocular active vision system. *IFAC Journal on Control Engineering Practice*, 5(10):1451–1461, 1997.

[14] J. Batista, P. Peixoto, and H. Araújo. Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking. In *Proc. of the IEEE Workshop on Visual Surveillance*, pages 18–25, 1998.

[15] P.W. Battaglia, R.A. Jacobs, and R.N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, 20(7):1391–7, Jul 2003.

[16] P. Bertelson and M. Radeau. Ventriloquism, sensory interaction, and response bias: Remarks on the paper by choe, welch, gilford, and juola. *Perception and Psychophysics*, 19(6):531–535, 1976.

[17] Körding KP Betsch BY, Einhäuser W and König P. The world from a cat's perspective - statistics of natural videos. *Biol Cybern*, 90:41–50, 2004.

[18] L. Boucher, A. Lee, Y.E. Cohen, and H.C. Hughes. Ocular tracking as a measure of auditory motion perception. *Journal of Physiology*, 98:235–248, 2004.

[19] M. Brecht, R. Goebel, W. Singer, and A. K. Engel. Synchronization of visual responses in the superior colliculus of awake cats. *Neuroreport*, 12(1):43–7, 2001.

[20] M. Brecht, W. Singer, and A. K. Engel. Correlation analysis of corticotectal interactions in the cat visual system. *J Neurophysiol*, 79(5):2394–407, 1998.

[21] M. Brecht, W. Singer, and A. K. Engel. Patterns of synchronization in the superior colliculus of anesthetized cats. *J Neurosci*, 19(9):3567–79, 1999.

[22] N. A. Busch, S. Debener, C. Kranczioch, A. K. Engel, and C. S. Herrmann. Size matters: effects of stimulus size, duration and eccentricity on the visual gamma-band response. *Clin Neurophysiol*, 115(8):1810–20, 2004.

[23] Kayser C, Körding KP, and König P. Learning the nonlinearity of neurons from natural visual stimuli. *Neural Computation*, 8:1751–1760, 2003.

[24] Kayser C and König P. Population coding of orientation in the visual cortex of alert cats - an information theoretic analysis. *NeuroReport*, (22):2761–4, 2004.

[25] Kayser C, Einhäuser W, and König P. Temporal correlations of orientations in natural scenes. *Neurocomputing*, 52-54:117–123, 2003.

[26] G. A. Calvert and T. Thesen. Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal of Physiology*, 98(1-3):191–205, 2004.

[27] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A component-wise em algorithm for mixtures. *Journal of Computational and Graphical Statistics*, 10:699–712, 2001.

[28] G Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

[29] S. Christy and R. Horaud. Euclidean shape and motion from multiple perspective views by affine iterations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(11):1098–1104, November 1996.

[30] S. Chroust, M. Vincze, J. Barreto, and H. Araújo. Solutions for visual control of motion: Active tracking applications. In *Proc. of ICAR'2001-*10^{th} *Int. Conf. on Advanced Robotics*, Budapest-Hungary, August 2001.

[31] M.P. Cooke. *Modelling Auditory Processing and Organisation*. Cambridge University Press, 1993.

[32] M.P. Cooke. Glimpsing speech. *Journal of Phonetics*, 31:579–584, 2003.

[33] M.P. Cooke and D.P.W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, (35):141–177, 2001.

[34] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and uncertain acoustic data. *Speech Communication*, 34:267–285, 2001.

[35] A. Coy and J.P. Barker. Recognising speech in the presence of a competing speaker using a 'speech fragment decoder'. In *Proc. ICASP 2005*.

[36] S. Debener, C. S. Herrmann, C. Kranczioch, D. Gembris, and A. K. Engel. Top-down attentional processing enhances auditory evoked gamma band activity. *Neuroreport*, 14(5):683–6, 2003.

[37] S. Debener, S. Makeig, A. Delorme, and A. K. Engel. What is novel in the novelty oddball paradigm? functional significance of the novelty p3 event-related potential as revealed by independent component analysis. *Brain Res Cogn Brain Res*, 22(3):309–21, 2005.

[38] Gustavo Deco. The computational neuroscience ofvisual cognition: Attention, memory and reward. In *Workshop on Attention and Performance in Computational Vision*, pages 100–117, 2004.

[39] D. Demirdjian and R. Horaud. Motion-egomotion discrimination and motion segmentation from image-pair streams. *Computer Vision and Image Understanding*, 78(1):53–68, April 2000.

[40] S. Deneve and A. Pouget. A. bayesian multisensory and cross-modal spatial links. *Journal of Neurophysiology*, 98(1-3):249–258, Jan-Mar 2004.

[41] J. Dias, C. Paredes, I. Fonseca, H. Araújo, J. Batista, and A. Almeida. Simulating pursuit with machines: Experiments with robots and artificial vision. *IEEE Trans. on Robot. and Automat.*, 14(1):1–18, 1998.

[42] J. Driver. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381:66–68, 1996.

[43] J. Driver and C. Spence. Multisensory perception: beyond modularity and convergence. *Curr Biol*, 10(20), 2000.

[44] Y. Dufournaud, C. Schmid, and R. Horaud. Image matching with scale adjustment. *Computer Vision and Image Understanding*, 93(2):175–194, February 2004.

[45] M. Eimer, J. van Velzen, and J. Driver. Erp evidence for cross-modal audiovisual effects of endogenous spatial attention within hemifields. *J Cogn Neurosci*, 16(2), 2004.

[46] A. K. Engel, P. Fries, and W. Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nat Rev Neurosci*, 2(10):704–16, 2001.

[47] A. K. Engel, P. Konig, A. K. Kreiter, and W. Singer. Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science*, 252(5010):1177–9, 1991.

[48] A. K. Engel and W. Singer. Temporal binding and the neural correlates of sensory awareness. *Trends Cogn Sci*, 5(1):16–25, 2001.

[49] F. Forbes and N. Peyrard. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(8):1089–1101, 2003.

[50] F. Forbes and A. E. Raftery. Bayesian morphology: Fast unsupervised bayesian image analysis. *Journal of the American Statistical Association*, 94(446):555–568, June 1999.

[51] D.A. Forsyth and J. Ponce. *Computer Vision – A Modern Approach*. Prentice Hall, New Jersey, 2003.

[52] P. Fries, S. Neuenschwander, A. K. Engel, R. Goebel, and W. Singer. Rapid feature selective neuronal synchronization through correlated latency shifting. *Nat Neurosci*, 4(2):194–200, 2001.

[53] P. Fries, J. H. Reynolds, A. E. Rorie, and R. Desimone. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291(5508), 2001.

[54] P. Fries, P. R. Roelfsema, A. K. Engel, P. Konig, and W. Singer. Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proc Natl Acad Sci U S A*, 94(23):12699–704, 1997.

[55] P. Fries, J. H. Schroder, P. R. Roelfsema, W. Singer, and A. K. Engel. Oscillatory neuronal synchronization in primary visual cortex as a correlate of stimulus selection. *J Neurosci*, 22(9):3739–54, 2002.

[56] Gerald Fritz, Christin Seifert, Lucas Paletta, and Horst Bischof. Attentive object detection using an information theoretic saliency measure. In *Workshop on Attention and Performance in Computational Vision*, pages 29–41, 2004.

[57] Daniel Gatica-Perez, Guillaume Lathoud, Iain McCowan, and Jean-Marc Odobez. A mixed-state i-particle filter for multi-camera speaker tracking. In *IEEE Int. Conf. on Computer Vision Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, 2003.

[58] Daniel Gatica-Perez, Guillaume Lathoud, Iain McCowan, Jean-Marc Odobez, and Darren Moore. Audio-visual speaker tracking with importance particle filters. In *IEEE International Conference on Image Processing (ICIP)*, 2003.

[59] C.M. Gray, P. Konig, A.K. Engel, and W. Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338:334–337, 1989.

[60] Fred Henrik Hamker. Modeling attention: From computational neuroscience to computer vision. In *Workshop on Attention and Performance in Computational Vision*, pages 118–132, 2004.

[61] S. Harding, J.P. Barker, and G. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. Submitted to IEEE Trans. on Speech and Audio Processing.

[62] L. Hérault and R. Horaud. Feature Grouping and Figure-Ground Discrimination: A Recursive Neural Network Approach. In *Proc. of the IEEE International Joint Conference on Neural Networks*, pages 2606–2611, Singapore, November 1991.

[63] L. Hérault and R. Horaud. Figure-ground discrimination: a combinatorial optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):899–914, September 1993.

[64] R. Horaud and M. Brady. On the geometric interpretation of image contours. *Artificial Intelligence*, 37(1–3):333–353, December 1988.

[65] R. Horaud, F. Dornaika, and B. Espiau. Visually guided object grasping. *IEEE Transactions on Robotics and Automation*, 14(4):525–532, August 1998.

[66] R. Horaud, D. Knossow, and M. Michaelis. Camera cooperation for achieving visual attention. Technical Report RR-5216, INRIA, INRIA Rhône-Alpes, Montbonnot, June 2004. To appear in Machine Vision and Applications.

[67] R. Horaud and Th. Skordas. Stereo matching through feature grouping and maximal cliques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(11):1168–1180, November 1989.

[68] R. Horaud and H. Sossa. Polyhedral object recognition by indexing. *Pattern Recognition*, 28(12):1855–1870, 1995.

[69] Körding KP Kayser C and König P. Processing of complex stimuli and natural scenes in the visual cortex. *Curr. Opin. Neurobiol.*, 14(4):468–73, 2004.

[70] D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, 55:271–304, 2004.

[71] R. Klinke, A. Kral, S. Heid, J. Tillein, and R. Hartmann. Recruitment of the auditory cortex in congenitally deaf cats by long-term cochlear electrostimulation. *Science*, 285(5434):1729–33, 1999.

[72] A. Kral, R. Hartmann, J. Tillein, S. Heid, and R. Klinke. Congenital auditory deprivation reduces synaptic activity within the auditory cortex in a layer-specific manner. *Cereb Cortex*, 10(7):714–26, 2000.

[73] A. Kral, R. Hartmann, J. Tillein, S. Heid, and R. Klinke. Delayed maturation and sensitive periods in the auditory cortex. *Audiol Neurootol*, 6(6):346–62, 2001.

[74] A. Kral, R. Hartmann, J. Tillein, S. Heid, and R. Klinke. Hearing after congenital deafness: central auditory plasticity and sensory deprivation. *Cereb Cortex*, 12(8):797–807, 2002.

[75] A. Kral, J. H. Schroder, R. Klinke, and A. K. Engel. Absence of cross-modal reorganization in the primary auditory cortex of congenitally deaf cats. *Exp Brain Res*, 153(4):605–13, 2003.

[76] C. Kranczioch, S. Debener, J. Schwarzbach, R. Goebel, and A. K. Engel. Neural correlates of conscious perception in the attentional blink. *Neuroimage*, 24(3):704–14, 2005.

[77] B. Lamiroy, B. Espiau, N. Andreff, and R. Horaud. Controlling robots with two cameras: How to do it properly. In *Proc. of IEEE International Conference on Robotics and Automation*, pages 2100–2105, San Francisco, CA, April 2000.

[78] T. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2002.

[79] Tai Sing Lee. Computations in the early visual cortex. *Journal of Physiology*, 2003.

[80] Siegel M and König P. A functional gamma-band defined by stimulus-dependent synchronization in area 18 of awake behaving cats. *J Neurosci*, 23:4251–60, 2003.

[81] D. Marr. *Vision*. W. H. Freeman, San Francisco, 1982.

[82] J. J. McDonald, W. A. Teder-Salejarvi, F. Di Russo, and S. A. Hillyard. Neural substrates of perceptual enhancement by cross-modal spatial attention. *J Cogn Neurosci*, 15(1), 2003.

[83] McGurk and MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[84] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715), 1985.

[85] B. C. Motter. Focal attention produces spatially selective processing in visual cortical areas v1, v2, and v4 in the presence of competing stimuli. *J Neurophysiol*, 70(3), 1993.

[86] M. H. Munk, P. R. Roelfsema, P. Konig, A. K. Engel, and W. Singer. Role of reticular activation in the modulation of intracortical synchronization. *Science*, 272(5259):271–4, 1996.

[87] G. Nase, W. Singer, H. Monyer, and A. K. Engel. Features of neuronal synchrony in mouse visual cortex. *J Neurophysiol*, 90(2):1115–23, 2003.

[88] K.J. Palomaki, G.J. Brown, and J.P. Barker. Techniques for handling convolutional distortion with 'missing data' speech recognition. *Speech Communication*, (43):123–142, 2004.

[89] V. Pavlovic, A. Garg, J.M. Rehg, and T.S. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *CVPR00*, pages II: 34–41, 2000.

[90] P. Peixoto, J. Batista, and H. Araújo. Integration of information from several vision sensors for a common task of surveillance. *Robotics and Autonomous Systems*, 31:99–108, 2000.

[91] P. Peixoto, J. Batista, and H. Araújo. Real-time human activity monitoring exploring multiple vision sensors. *Robotics and Autonomous Systems*, 35:221–228, 2001.

[92] G. Potamianos, C. Neti, G. Gravier, and A. Garg. Automatic recognition of audio-visual speech: Recent progress and challenges. *Proceeding of the IEEE*, 91(9), 2003.

[93] J. H. Reynolds and R. Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24(1), 1999.

[94] Salazar RF, Kayser C, and König P. Effects of training on neuronal activity and interactions in primary and higher visual cortices in the alert cat. *J Neuroscience*, 24:1627–1636, 2004.

[95] A. Riehle, S. Grun, M. Diesmann, and A. Aertsen. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, 278(5345), 1997.

[96] P. R. Roelfsema, A. K. Engel, P. Konig, and W. Singer. Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature*, 385(6612):157–61, 1997.

[97] P. R. Roelfsema, V. A. Lamme, and H. Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395(6700), 1998.

[98] E. T. Rolls and G. Deco. *Computational Neuroscience of Vision*. Oxford University Press, 2002.

[99] A. Ruf and R. Horaud. Projective rotations applied to a pan-tilt stereo head. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–150, Fort Collins, Colorado, June 1999. IEEE Computer Society Press.

[100] A. Ruf, F. Martin, B. Lamiroy, and R. Horaud. Visual control using projective kinematics. In John M. Hollerbach and Daniel E. Koditschek, editors, *Robotics Research, The Ninth International Symposium*, pages 95–104. Springer, 2000.

[101] S. Shimojo, C. Scheier, R. Nijhawan, L. Shams, Y. Kamitani, and K. Watanabe. Beyond perceptual modality: Auditory effects on visual perception. 22(2):61–67, 2001.

[102] W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1), 1999.

[103] D. Sodoyer, J.L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten. Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli. *EURASIO Journal of Applied Signal Processing*, 11:1165–1173, 2002.

[104] Meredith MA Stein BE. *The merging of the senses*. The MIT Press, Cambridge (MA), 1993.

[105] P. N. Steinmetz, A. Roy, P. J. Fitzgerald, S. S. Hsiao, K. O. Johnson, and E. Niebur. Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404(6774), 2000.

[106] J. Vermaak, A. Blake, M. Gangnet, and P. Pérez. Sequential Monte Carlo fusion of sound and vision for speaker tracking. In *Proc. IEEE Int. Conf. Computer Vision, ICCV'01*, volume 1, pages 741–746, Vancouver, Canada, July 2001.

[107] Einhäuser W and König P. Does luminance-contrast contribute to a saliency map of overt visual attention? *Eur J Neurosci*, 17:1089–97, 2003.

[108] H. Wagner. A comparison of neural computations underlying stereo vision and sound localization. *Journal of Physiology*, 98(1-3):135–45, Jan-Jun 2004.

[109] Andrei Zaharescu, Albert L. Rothenstein, and John K. Tsotsos. Towards a biologically plausible active visual search model. In *Workshop on Attention and Performance in Computational Vision*, pages 133–147, 2004.

Forms A3.1 and A3.2

Contract Preparation Forms

EUROPEAN COMMISSION

6th Framework Programme on Research, Technological Development and Demonstration

Specific Targeted Research or Innovation Project

A3.1

Please use as many copies of form A3.1 as necessary for the number of partners

Proposal Number	027268	Proposal Acronym	POP

Financial information - whole duration of the project

Participant nº	Organisation short name	Cost model used	Estimated eligible costs and requested EC contribution (whole duration of the project)		Costs and EC contribution per type of activities: RTD or innovation related activities (1)	Demonstration activities (2)	Consortium Management activities (3)	Total (4)=(1)+(2)+(3)	Total receipts
5	USFD	AC	Eligible costs	*Direct Costs (a)*	371 225,00		23 000,00	394 225,00	
				of which subcontracting				,00	
				Indirect costs (b)	74 245,00		4 600,00	78 845,00	
				Total eligible costs (a)+(b)	445 470,00	,00	27 600,00	473 070,00	
			Requested EC contribution		445 470,00		27 600,00	473 070,00	
4	FCTUC	AC	Eligible costs	*Direct Costs (a)*	222 000,00		8 667,00	230 667,00	
				of which subcontracting				,00	
				Indirect costs (b)	44 400,00		1 733,00	46 133,00	
				Total eligible costs (a)+(b)	266 400,00	,00	10 400,00	276 800,00	
			Requested EC contribution		266 400,00		10 400,00	276 800,00	
3	UKE	AC	Eligible costs	*Direct Costs (a)*	320 625,00		14 708,00	335 333,00	
				of which subcontracting				,00	
				Indirect costs (b)	64 125,00		2 342,00	66 467,00	
				Total eligible costs (a)+(b)	384 750,00	,00	17 050,00	401 800,00	
			Requested EC contribution		384 750,00		17 050,00	401 800,00	
2	UOFM	AC	Eligible costs	*Direct Costs (a)*	276 450,00		12 000,00	288 450,00	
				of which subcontracting				,00	
				Indirect costs (b)	55 290,00		2 400,00	57 690,00	
				Total eligible costs (a)+(b)	331 740,00	,00	14 400,00	346 140,00	
			Requested EC contribution		331 740,00		14 400,00	346 140,00	

Contract Preparation Forms

EUROPEAN COMMISSION

6th Framework Programme on Research, Technological Development and Demonstration

Specific Targeted Research or Innovation Project

A3.1

Please use as many copies of form A3.1 as necessary for the number of partners

Proposal Number	027268	Proposal Acronym	POP

Financial information - whole duration of the project

Participant nº	Organisation short name	Cost model used	Estimated eligible costs and requested EC contribution (whole duration of the project)		Costs and EC contribution per type of activities: RTD or innovation related activities (1)	Demonstration activities (2)	Consortium Management activities (3)	Total (4)=(1)+(2)+(3)	Total receipts
1	INRIA	FC	Eligible costs	*Direct Costs (a)*	361 500,00		13 020,00	374 520,00	
				of which subcontracting				,00	
				Indirect costs (b)	386 800,00		15 020,00	401 820,00	
				Total eligible costs (a)+(b)	748 300,00	,00	28 040,00	776 340,00	
			Requested EC contribution		374 150,00		28 040,00	402 190,00	
TOTAL			Eligible costs		**2 176 660,00**	**,00**	**97 490,00**	**2 274 150,00**	**,00**
			Requested EC contribution		**1 802 510,00**	**,00**	**97 490,00**	**1 900 000,00**	

Contract Preparation Forms

EUROPEAN COMMISSION

6th Framework Programme on Research, Technological Development and Demonstration

Specific Targeted Research or Innovation Project

A3.2

Proposal Number	027268	Proposal Acronym	POP

Estimated breakdown of the EC contribution per reporting period

Reporting Periods	Start month	End month	Estimated Grant to the Budget	
			Total	In which first six months
Reporting Period 1	1	12	597 633,00	,00
Reporting Period 2	13	24	803 639,00	401 820,00
Reporting Period 3	25	36	498 728,00	249 365,00
Reporting Period 4			,00	,00
Reporting Period 5			,00	,00
Reporting Period 6			,00	,00
Reporting Period 7			,00	,00