Seventh Framework Programme
Theme ICT-2009.2.1
Cognitive Systems and Robotics

Grant agreement for: Small or medium scale focused research project

Annex 1 – "Description of Work"

Project acronym:	**HUMAVIPS**
Project full title:	**Humanoids with auditory and visual abilities in populated spaces**
Grant agreement no.:	247525

List of beneficiaries:

Number	Name	Short name	Country	Enter date	Exit date
1 (Coord.)	Institut National de Recherche en Informatique et Automatique	INRIA	France	month 1	month 36
2	The Czech Technical University	CTU	Czech Republic	month 1	month 36
3	Aldebaran Robotics	ALD	France	month 1	month 36
4	IDIAP Research Institute	IDIAP	Switzerland	month 1	month 36
5	Bielefeld University	BIU	Germany	month 1	month 36

Contents

A.1 Overall budget breakdown for the project

			Estimated budget (whole duration of the project			Requested EU contribution
No.	Name	Country	RTD	Management	Total	
1	INRIA	FR	640412	148365	788777	628636
2	CTU	CZ	606080	21360	627440	475920
3	ALD	FR	572800	30300	603100	459900
4	IDIAP	CH	730624	26536	757160	574504
5	BIU	DE	634560	21120	655680	497040
		Total:	3184476	247681	3432157	2636000

A.2 Project summary

Humanoids expected to collaborate with people should be able to interact with them in the most natural way. This involves significant perceptual, communication, and motor processes, operating in a coordinated fashion. Consider a social gathering scenario where a humanoid is expected to possess certain social skills. It should be able to explore a populated space, to localize people and to determine their status, to decide to join one or two persons, to synthetize appropriate behavior, and to engage in dialog with them. Humans appear to solve these tasks routinely by integrating the often complementary information provided by multi sensory data processing, from low-level 3D object positioning to high-level gesture recognition and dialog handling. Understanding the world from unrestricted sensorial data, recognizing people's intentions and behaving like them are extremely challenging problems. The objective of HUMAVIPS is to endow humanoid robots with audiovisual (AV) abilities: exploration, recognition, and interaction, such that they exhibit adequate behavior when dealing with a group of people. Proposed research and technological developments will emphasize the role played by multimodal perception within principled models of human-robot interaction and of humanoid behavior. An adequate architecture will implement auditory and visual skills onto a fully programmable humanoid robot. An open-source software platform will be developed to foster dissemination and to ensure exploitation beyond the lifetime of the project.

A.3 List of beneficiaries

Number	Name	Short name	Country	Enter date	Exit date
1 (Coord.)	Institut National de Recherche en Informatique et Automatique	INRIA	France	month 1	month 36
2	The Czech Technical University	CTU	Czech Republic	month 1	month 36
3	Aldebaran Robotics	ALD	France	month 1	month 36
4	IDIAP Research Institute	IDIAP	Switzerland	month 1	month 36
5	Bielefeld University	BIU	Germany	month 1	month 36

B.1 Concept and objectives, progress beyond state of the art, S/T methodology and work plan

B.1.1 Concept and project objectives

In recent years, robots have gradually started to move from structured factory floors to populated spaces (public, professional, or private) which are highly unstructured environments. On one side there are robots that share their workspace with people; these robots should be able to carry on their tasks safely and efficiently. On the other side, there is an increasing need for robots that interact, communicate, and cooperate with people [50, 11]. Today's scientific and technological progress allows robots equipped with multiple types of sensors to operate in spaces which were not specifically designed to host robots, i.e., to navigate in complex environments and to solve specific tasks. Comparatively, robots endowed with cognitive capabilities such that they can interact with people in the most natural way, are less developed. In particular it is important to go beyond the single-user and constrained-conditions human-robot interaction paradigm that prevails today.

Imagine an **informal social gathering scenario** where a robot is expected to possess a few simple *social skills*. It should be able to make its way among a group of people without hurting them, to recognize people's activities such as verbal and prosodic communications, body postures, head or hand gestures, to understand people's status (whether two or more persons are engaged into a conversation), to decide to join a small group of people, and to synthetize appropriate behavior such as signaling its presence using hand gestures, heading towards the currently speaking person, requesting conversation turns, etc.

The **robot-with-group-of-people** interaction example just cited belongs to a broader class of complex scenarios raising a wealth of key questions that have not yet been completely answered. In this project we plan to put strong emphasis on studying and developing **robots with auditory and visual skills**. In particular we want to provide methodological solutions to the following problems: How should a robot learn and recognize objects that are both seen and heard? Which auditory or visual patterns allow a robot to distinguish between several people or between people and artifacts that also emit sounds? How many persons are in a room and where are they located? Which is the optimal exploration strategy for sensing a complex audiovisual (AV) scene? How can a robot communicate with people both ways (analysis and synthesis) in non-controlled settings, i.e., in the real physical world? How can we build a robot-centered architecture that combines multisensor exploration with adequate behavior?

To summarize:

> *The objective of HUMAVIPS is to endow robots with audiovisual (AV) abilities, i.e., AV exploration, AV recognition, and AV interaction, such that they exhibit adequate behavior when dealing with a group of people.*

In order to fulfill this ambitious goal, to implement the project's scientific and technological innovations, and to demonstrate the project's outcomes, an adequate autonomous robotic platform must be carefully selected. In particular we believe that the following *anthropomorhic* features are important:

- **Synthesis of complex motions.** The robot should be able to perform human-like actions when involved in exploration and interaction tasks, such as walk, gesticulate, turn the body, shake, nod, or tilt the head, gaze to someone, point an object, etc. These features are of special interest for multimodal dialog modeling where different levels of understanding, e.g., non-verbal, acoustic, visual, semantic, need to be signaled [17].

- **Autonomy.** A robot-centred sensor architecture is particularly well suited for modeling the group-of-people-to-robot-interaction scenarios that we plan to develop. The robot should be able to place itself with respect to a person in order to optimize the audiovisual interaction, then rapidly turn its AV attention to another person, etc. This rules out the use of remote sensors such as camera or microphone networks mounted onto the walls or onto the ceilings, or of wearable devices such as head-mounted microphones.

- **Reactive behavior.** The analysis of human AV cues followed by the synthesis of robot behaviors require complex, yet *efficient sensorimotor loops*. This implies that transmission and processing of sensory data, decision-making and control loops must be optimally implemented. To this end, the computation burden could be distributed between onboard processors and remote ones, using modern communications techniques.

- **User acceptability.** It seems that humans activate different brain regions when interacting with a humanoid as opposed to a more functional looking robot [75]. This suggests that users speculate more about the goals and intentions of a robot if it is more human-looking. Within the HUMAVIPS paradigm, this is important as the reactions of the robot need to be interpreted by the users in a similar way as they would interpret human reactions.

To conclude, the HUMAVIPS proposers believe that a humanoid robot best achieves a balance between the criteria above, and hence it is a well suited robotic platform to develop, implement and demonstrate a wealth of innovative methodologies:

- **A humanoid that explores an unstructured environment.** HUMAVIPS will investigate methods to dynamically build a 3D description of its surrounding objects through unsupervised extraction and fusion of relevant sensor information gathered with a few microphones and cameras, i.e., spatial hearing and stereoscopic vision. The emphasis will be on the detection and localization of humans and on the characterization of their motion patterns and status (silent, emitting sounds, speaking, etc.). In particular, HUMAVIPS will develop methodologies allowing a humanoid to robustly deal with very general situations such as a varying number of people that wander around, gesticulate, emit speech and non-speech sounds, all in the presence of reverberations or other auditory sources, other objects, etc.

- **A humanoid that recognizes, understands and interacts with people.** HUMAVIPS will explore the roles of *multimodality* and *active sensing* to design a robust speech-, prosody- and gesture-based **humanoid-human interface**. Emphasis will be put on informal settings where several people are present. The humanoid will have to: (i) select a person who is available (easy to reach, not committed in a private interaction

with another person, etc.), (ii) optimally place itself in front of the selected person in order to robustify the humanoid-to-human AV interactions, and (iii) communicate with that person using a combination of verbal modalities (e.g., speech and prosody recognition and synthesis) and non-verbal modalities (e.g., gesture recognition and synthesis).

- **A humanoid with a memory-centered cognitive architecture.** HUMAVIPS will develop an architecture needed for the challenge of a humanoid being engaged in interaction with several people in parallel. Indeed, a continuous balancing between active multisensor exploration, on one side, and adequate synthesis of behavior, on the other side, demands for a systemic architectural approach. HUMAVIPS will hence investigate the potential of memory-centered architecture comprising short- and long-term memories with a special focus on fusing *social* and *perceptual* abilities in cognitive models. This will result in a novel humanoid-focused robot architecture informed by cognitive foundations such as associative retrieval of knowledge, visuo-spatial sketch pads, active perception loops, as well as arbitration on the basis of action primitives such as particular body movements, gestures, and bipedal locomotion.

To summarize, HUMAVIPS will be a highly innovative project, for the following reasons:

(i) *It will emphasize the crucial role played by multimodal perception within a principled computational model of humanoid behavior in the presence of a group of people*; **(ii)** *auditory and visual fusion, recognition and communication will be integrated from the ground up*; **(iii)** *an adequate architecture will implement audiovisual skills onto a fully programmable commercial humanoid*; **(iv)** *the impact of HUMAVIPS in terms exploitation and dissemination on the consumer's market will be encouraged by a well thought open-source software platform with an adequate licensing strategy.*

HUMAVIPS will deliver the followings:

1. Methods and algorithms allowing a humanoid robot to fuse audio-visual observations, to extract meaningful information in order to characterize several people composing an unstructured environment, and to interact with one or several persons (WP3, page 25, WP4, page 28 and WP5, page 31).

2. A cognitive architecture for representing the humanoid's short- and long-term perceptive history as well as its low- and high-level knowledge that are needed for robust robot-to-several-people interactions based on audio and visual cues (WP2, page 23).

3. Proof-of-concept demonstrators of increasing complexity which are relevant to humanoid applications (WP1, page 20 and WP7, page 35).

4. The HUMAVIPS scenarios (a humanoid among a group of people) will be trained with a carefully annotated set of corpora gathered in a realistic setting. These data sets will be made publicly available to be used for benchmarking purposes (WP1, page 20).

5. The demonstrators will be implemented on an European-built commercial humanoid robot equipped with an improved audio-visual head (WP7, page 35).

6. An open-source software (OSS) collaborative platform which will host HUMAVIPS's software packages (WP6, page 34). These software packages will be compatible with NAO's distributed architecture concept as well as with any other programmable robotic platform equipped with the necessay sensors, actuators, and processing units.

7. There will be clear dissemination and exploitation strategies and intellectual property right (IPR) policies associated with this OSS platform (WP8, page 38). This software platform will allow strong intra- and inter-project collaborations and synergies during and beyond the lifetime of HUMAVIPS.

The primary goal of HUMAVIPS is to integrate auditory and visual abilities at multiple levels within a robot-centred cognitive architecture, from low-level exploration of an unstructured populated space to non-verbal and verbal interactions through analysis of human audio-visual activities and synthesis of appropriate behavior. HUMAVIPS thus addresses the following issues:

- HUMAVIPS will develop robots operating in real world environments, in particular it will investigate complex scenarios involving people and groups of people whose behavior varies over time in an unpredictable way.

- The AV skills to be developed will allow both understanding of a wide spectrum of human non-verbal and verbal communicative cues and synthesis of human-like behavior. There will be no need to wear special devices (such as head-mounted microphones) in order to communicate with the robot. Likewise, there will be no need for specially equipped rooms.

- HUMAVIPS will develop methods and algorithms based on *stereoscopic vision* and *auditory scene analysis* to learn and recognize audio-visual objects, i.e. objects that are both seen and heard. These objects (humans or artefacts) will be located, counted, recognized, and described as they appear in the 3D space.

- The HUMAVIPS's OSS collaborative platform, will allow a principled way of joint software development by several teams, both from academia and from industry, and of integrating these software packages with fully programmable robots. Indeed, the use of an open-source platform, developed along the stragegy proposed in HUMAVIPS, will be compatible with the commercial robot NAO developed by ALD, as well as with any programmable robot developed elsewhere. The potential designers, manufacturers, developers, and users will have to adhere to IPR exploitation, policies and licensing rules commonly practiced in open-source communities.

- The application scenarios will guide the acquisition of a set of annotated corpora. These data sets will be gathered by a humanoid with its auditory and visual sensors. The data will be annotated with the ground-truth parameters as well as with scripts describing in detail the social situations that the humanoid is supposed to understand. This will be made public to serve to a wider community of robotic developers and manufacturers.

B.1.2 Progress beyond the state of the art

Four major themes in HUMAVIPS are **audio-visual exploration**, **audio-visual recognition**, **audio-visual interaction and communication**, and **interaction-enabling architectures**. The structure of the project is designed around these four areas. This section discusses previous work in these areas that is relevant to the work proposed in HUMAVIPS.

A. Audio-visual exploration

A common characteristic of the vast majority of robots and systems that handle multimodal data is that *hearing and vision are treated relatively independently using modality-specific subsystems* whose results are combined afterwards at a higher level. The performance of such procedures in realistic situations is limited. Difficulties arise from various sources including background acoustic and visual noise, acoustic reverberation, and visual occlusions. The work in HUMAVIPS will focus on the following scientific problems.

Audio-visual data fusion. The first question to be addressed is: *where should the fusion of the AV data take place?* Various strategies are possible. In contrast to the fusion of independently processed modalities [58], the integration could occur at the feature level. In this case audio and video features are directly combined to a larger feature vector which is then used to perform the task of interest. However due to the very different physical nature of audio and visual stimuli, this *concatenation* is not straightforward.

There is no obvious way to associate *dense and spatial visual maps* with *sparse and temporal sound sources*. Two major directions can be identified, depending on the type of *synchrony* being used.

The first one focuses on *spatial synchrony* and implies combining those signals that were observed at a given time, or through a short time window, and correspond to the same AV object (e.g. a speaker). Generative probabilistic models [12], [76] for single speaker tracking achieve this by introducing dependencies of both auditory and visual observations on locations in the image plane. Although [12] suggested an enhancement of the model that would tackle the multi-speaker case, it has not been shown to work for conditions where people's behavior is natural. The explicit dependency on the source location that is used in generative models can be generalised by replacing the parametrised distribution with a set of particles. Particle filters were used for the task of single speaker tracking [128], [114], [91], [26] and multiple speaker tracking [25], [47], [26]. In the latter case the parameter space grows exponentially as the number of speakers increases, so efficient sampling procedures were suggested [47], [26] to keep the problem tractable.

The second direction focuses on *temporal synchrony*, which generalises the previous approach by making no *a priori* assumptions about AV object location. Signals from different modalities are grouped if their evolution is correlated through time. In [36] it is shown how principles of information theory can be used to select those features from different modalities that correspond to the same object. Although the setup consists of only a single camera and a single microphone and no special signal processing is used, the model is capable of selecting the speaker among several persons that are visible. Another illustrative approach within this strategy is offered in [7]. Matching is performed based on audio and video onsets (times at

which sound/motion begins). This model is successfully tested even on the case with multiple sound sources. Most of these approaches are, however, non-parametric and highly dependent on the choice of appropriate features. Moreover they usually require learning or ad-hoc tuning of quantities such as window size and temporal resolution. They appear relatively sensitive to artifacts and would likely benefit from more robust implementations.

Active listening and looking. Head and body movements have great potential in improving SNR in settings where other sound sources and reverberation are present. Head rotations could, in principle, be used to improve SNR at the ear closest to the source of interest, or to attenuate interfering sources, or to locate the target source in the frontal (high-resolution) part of the azimuth plane. Listener motion could be used to improve SNR by moving closer to the source of interest or away from interfering sources, or highly reverberant surfaces, or to obtain line-of-sight for visual cues. However, apart from demonstrations that head rotation can disambiguate auditory front-back confusions [116], and that motion can improve auditory distance estimation [104], there have been no studies of the kind of movements listeners make in realistic settings. Models of these effects would help anticipate listener behavior, and could potentially be used to inform the robot's own motor control strategies.

Scene representation. Another important issue is that of constructing a spatial representation of the environment, based on multi-sensory data, such that a robot is able to localize itself with respected with the sensed objects. This is an instance of the *simultaneous localization and mapping* (SLAM) problem. Previous research in SLAM has lead to the development of statistical methods based on processing various sensory inputs, most often laser, sonar, GPS or camera-based [111], [32]. However, little attention has been given to the use of natural *audio* cues despite their obvious advantages: listening for cues to the physical layout of the environment is non-obtrusive, omni-directional and *complementary to the use of visual cues*. We note that the *audio-visual SLAM* concept, that we plan to investigate and develop in HUMAVIPS, is a new paradigm that has not been studied in the past.

With respect to **audio-visual exploration**, HUMAVIPS will investigate novel approaches for fusion of auditory and visual data for localization, active listening and looking, and scene representation, i.e, WP3, page 25.

B. Audio-visual recognition

Recognition from AV data and the characterisation of AV objects, and more particularly people, is a very active research field. This is especially the case for the multi-party conversational domain, in which many sensors and computational resources are often available, and for human-computer interfaces, in which audio and video signals recorded from people are usually of good quality. In these contexts, some of the technologies employed in HUMAVIPS are already established (e.g. frontal face detectors, single speaker localisers, and prosody feature extractors), and indeed some of these components have been imported as off-the-shelf modules for previous robotics research. Many other tasks and cue extraction problems, however, still represent fundamental challenges in vision and audio processing (e.g. accurate estimation of head pose,

focus-of-attention, multi-party speech segmentation), especially in cases of limited processing and sensing capabilities (cluttered audio recorded with distant microphones, low resolution images), as often encountered with autonomous robots in complex scenarios. In particular, the case of multiple people concurrently interacting with the robot, which will be studied in HUMAVIPS , represents a challenging situation for such tasks and has not been solved so far. The state-of-the-art in the subjects addressed in HUMAVIPS can be summarized as follows.

Audio-visual object categorisation. In order to be socially situated, a robot needs to distinguish its main interlocutors from other social agents and objects in its environment. Hence, it must perform *audio-visual object categorisation*. In many robotics applications (e.g. affective robots), this is not an issue as only one person is interacting with the robot. Except in very specific robot learning and imitation scenarios, little research has been done in this domain, where proximity or the presence of a face or of an audio sound have often been considered as main cues for the task [37]. The interplay of these cues, both spatially and temporally, along with the appropriate discrimination of audio-visual sources need to be investigated when considering human-robot interactions in more cluttered environments, such as those investigated in HUMAVIPS .

Spatio-temporal saliency analysis. Research has been done on finding strategies to share processing between different tasks, and especially to control active systems like stereo gazing systems towards conspicuous regions of the scene. These models often follow rules based on the extraction of visual saliency modeling relying on color, static interest points or dynamical features [65]. Application of such visual attention models to robot interaction [16] or localisation in static scenes [97] has been investigated. Similar saliency approaches have been considered in audio signals, but there has been little research on the construction and exploitation of audio-visual saliency maps for environment exploration and resource allocation. These aspects will be studied in HUMAVIPS .

Extraction of low-level behavioral cues. AV processing is particularly useful for the extraction of low-level behavioral cues. Regarding audio, the extraction of audio features (turn-taking patterns, prosodic features such as pitch, loudness, rhythm) from close-talk or lapel microphones is mature [8]. However, in cluttered multi-party settings recorded from distant microphones, reliably extracting speaker-turns is still challenging, and relatively little work has been done on the extraction of prosodic features. Methods can be divided in model- and appearance-based ones. Model-based methods often use an articulated representation for the human body [74], [117], a model for hands [106], [107], a general model of the human face [61], or motion-primitives patterns [112]. Appearance-based methods rely on color or intensity-based features detected locally as skin blobs, head-and-shoulder contours pattern [127], hands [18], [85], human heads [51], [108], head poses [4], and entire humans [93]. Overall, while model-based approaches can in essence be very accurate, they usually require higher resolution images/videos and have difficulty coping with automatic (re)initialisation. Appearance-based approaches, on the other hand, offer increased robustness and higher speed for the lower resolution head images while still achieving good accuracy.

Extraction of non-verbal social cues. The main non-verbal social cues that have been studied include specific hand [18, 86] and head gestures (nods, shakes) [84] in single person-robot interaction cases, but other cues, including spatial person-robot placement relations, have been considered as well [64]. Less research has been performed on the extraction of the visual focus of attention defined by eye gaze, despite the important role it plays in the understanding of communicative behaviors. As standard gaze tracking technologies can not often be applied, head orientation has been studied as a surrogate for gaze [108, 89, 4]. In most cases, the head pose-gaze direction relationship is learned from training data, or the gaze is assimilated to the head direction [89], although psychosocial findings show that this relation is more complex in practice [87]. Finally, the recognition and interpretation of the above cues should not be considered separately, but jointly, as some of the behaviors provide context to some others. This includes head pose for hand gesture recognition [86], dialog acts [84], or speaking patterns of multiple participants [108, 89], and the use of artifacts such as slides [4], for joint visual focus of attention recognition in multi-person conversational settings.

With respect to **audio-visual recognition**, HUMAVIPS will investigate new methods for object categorisation, spatio-temporal saliency analysis, and extraction of low-level behaviors and mid-level social cues, i.e., WP4, page 28.

C. Audio-visual interaction and communication

Human-robot interactions are currently far from natural. Existing systems do not understand or reproduce the subtle gestural and vocal cues that allow humans to communicate effectively in multiparty situations. Even in one-to-one exchanges, the robot component struggles to react appropriately when it has misunderstood the intent of the communcation [49]. Multimodal input and output promises to increase both the naturalness and robustness of human-robot interaction. The challenge is how to achieve an adequate modality synergy which takes maximum advantage of all modalities involved.

Multi-party dialog modeling. Dialog modeling in robotics has for a long time been a neglected area of research resulting in dialog modules with hand-crafted dialog scripts explicitly assigned to each robot-internal state. Only recently the concept of grounding has been adopted in human-robot interaction [79]. Grounding describes the process of two interaction partners to incrementally build a common ground of shared beliefs [27]. However, although in theory more than two partners can take part in a conversation, computational models have for a long time only taken dyads into account. An explicit account for multi-party dialog has been developed in an immersive VR scenario [113]. Although this approach draws from the theory of grounding, it focuses on the problem of how to determine when to react to which contribution. Accordingly, different levels of communication management are considered, among others: contact, attention, turn-taking, and initiative management. In [113], *contact* describes if and how other individuals can be accessible for communication (e.g. via visual or vocal information), or for technically mediated interaction (radio). Since [113] is based on military communication, it neglects social aspects of contact, such as whether a potential interaction partner wants to interact or not. In human-robot interaction there exist approaches that estimate a person's interest in interacting

with a robot, based on her spatial movements towards the robot [1]. Similarly, in order for the robot to initiate contact, it has been shown that there are strong preferences by human subjects to be approached by a robot from front right or left, whereas a frontal or rear approach is not appreciated at all [31, 123]. Based on these results human-aware navigation approaches have been developed for service robots [99, 98]. These findings represent a significant step towards socially aware human-robot interaction. However, the engagement of a robot in a continuous interaction with a group of humans has not yet been studied. [52] defined eight dimensions along which spatial relations between humans can be described that are relevant for interaction. How other proxemic factors may be used by a robot in order to be perceived as a social interaction partner remains an open issue, together with the question of how these factors have to be taken into account by a robot when addressing a group of humans.

Turn-taking management. A specific task in dialog modeling is *turn-taking* management. Many features have been reported to play a role in managing turn-taking in human-human interaction. In [35] a set of features based on phonetics, paralanguage and body motion was proposed. In [69] gaze direction has been found to convey important information for the timing of turns and the overall naturalistic modeling of turn-taking. In contrast to these findings, approaches to model natural turn-taking behavior in artificial systems are often based on perceptual-level features. More recent approaches try to incorporate more semantic information such as prosodic cues [118] [59], information about word classes [110] or word-content mappings [45] in order to determine the timing for feedback. However, evaluation studies indicate that such models still yield behavior that is significantly different from natural behavior (e.g. [21]). In summary, turn-taking behavior in multi-party interactions has not yet been modeled in human-robot interaction although computational appraoches for predicting turn change in multi-party conversation exist. In HUMAVIPS, we plan to investigate novel models for multi-party HRI turn-taking, which will rely on a number of interaction patterns extracted automatically.

Detecting addressing patterns. Members of multi-party interactions need to keep track of who is speaking (the addressor) and who is primarily being addressed (the addressee) and which party members are attending to what is being said. However, there are relatively few studies on automatic identification of addressees in conversations from non-verbal cues. [68] presented a study on the identification of addressees between two people and a simulated robot using head pose (as a surrogate for gaze) and audio features, finding that head pose is a strong cue for addressee identification, and that audio-visual fusion can be useful. Overall, automatic addressing modeling from audio-visual sensors in real situations remains an open problem, in particular for robotic platforms. The latter will be the case studied in HUMAVIPS in the context of constrained computational resources for this task.

Monitoring participant engagement. Not all members of a multi-party conversation are equally engaged. Awareness of which members are 'tuned in' supplies a context that helps predict the conversational flow. Further, monitoring engagement can signal which segments of conversations are worth attending to – a useful skill for a robot with limited resources. Modeling

engagement is an emerging problem that has been explored in multi-person conversational settings [124, 70, 48, 90]. However, with a few exceptions which have explored the use of audio-visual cues [48, 90], existing work has only analysed the relation between interest and the speech modality. [124] introduced the concept of hot-spots (i.e. speech utterances where participants are highly involved in a discussion), and related it to the concept of activation in emotion modeling. Emphasis for speech utterances were also defined [70], acknowledging that this concept and emotional involvement might be acoustically and perceptually similar. [48] studied the performance of audio-visual cues in discriminating high versus neutral group interest-level segments in multi-party conversations (simultaneously segmenting a meeting and classifying the segments) and showed that, while the audio modality is dominant, audio-visual fusion improves performance and is thus beneficial. In the robotic context, AV fusion can potentially be more robust given the expected degradation of the audio signal quality in a complex scene sensed with distant microphones. This will be investigated in HUMAVIPS .

Judging dominance and influence. Patterns of interaction in multiparty conversations are determined in part by the relative social dominance of the party members. Dominance is a well-studied phenomenon in social psychology and a solid body of knowledge about its multimodal nature exists [34]. However, the automatic analysis of dominance has been studied only recently [9, 94, 63]. For example, a probabilistic approach for the automatic discovery of pair-wise influences (a phenomenon related to dominance) in a room equipped with cameras and microphones, and using inexpensive audio-visual cues, has been proposed [9]. [94] proposed a supervised learning approach to classify people in conversations into a small number of dominance levels. [63] have recently shown that simple models and computationally efficient non-verbal activity cues can predict the most dominant person in a conversation with reasonable accuracy. A specific challenge for HUMAVIPS is the identification of dominant people from distant sensors, a direction that has started to be addressed in [62], with clear relevance for work in robotics.

With respect to **audio-visual interaction and communication**, HUMAVIPS will investigate novel approaches for short-term and long-term modeling of multiparty communicative behavior and for multiparty dialog modeling and management, i.e., WP5, page 31.

D. Interaction-enabling architectures

HUMAVIPS is original in its aim to address the challenges of multi-modal active perception and social competence for a humanoid robot as a joint effort. This calls for context-aware selection of appropriate actions on the robotic platforms, effecting its sensors and actuators directly or indirectly. In parallel, multi-modal fusion for perception and the close coupling to the actuator needs to be established in order to facilitate active perception loops. This poses challenges that HUMAVIPS will address through an interaction-enabling memory architecture. The proposed robot system will feature *behaviors* operationalized according to different aims, focusing on perception and interaction with groups of humans, and built around an appropriate memory architecture.

Behavior-based robotics [2] has been studied for some time. It is characterized by mostly

concurrent activation of behaviors and appropriate arbitration mechanisms, e.g. applying priorization in the subsumption architecture [19]. Behavior-oriented design is proposed in [20] as a promising decomposition and implementation strategy also for cognitive (and interactive) systems. In HUMAVIPS, the arbitration and decision making are inherently context-driven, consulting models of social situations and perception prediction. [33] defined a context as "any information that can be used to characterize situations". Based on this notion, [29] presented an ontology for context and situation that provides an operational theory of context awareness, which can serve as a foundation to the specialization of general context-awareness towards models of social interaction-awareness [105] that in HUMAVIPS represent the driving force for action selection and arbitration. Multi-modal fusion processes that define social context and effect the robot's behavior are particular subjects to be researched. This research will also be substantiated by architecture issues. The problem of continuously linking percepts to concepts has been investigated allowing to match and bind attributes of perception and model of entities, thereby enabling high-level tracking. The study of these challenges in HUMAVIPS will be informed by basic cognitive principles. The cognitive principles of working memory have been operationalized for artificial systems before. Most prominent examples are cognitive architectures build upon the well-established ACT-R and SOAR paradigms [66], which focus on autonomous reasoning and deliberation, while the HUMAVIPS scenarios also demand an architecture that supports responsive control and adaptation. The cognitive principles of memories and memory processes [6] have also informed more recent works on cognitive systems. The visual active memory perspective [10] and the CAST perspective [56] can be seen as two recent representatives. These memory architectures [55, 125] have proven their general suitability as a promising architectural concept for larger scale research systems, characterized by their demands for continuous integration and evolving requirements. The approaches in [55] and [125] encounter the challenge of low-coupling between software components to facilitate rapid progress in development while allowing flexible learning and close-coupled binding between information domains stored and exchanged via respective memories.

With respect to **interaction-enabling memory architectures**, HUMAVIPS will investigate the implementation and design of interactive behaviors on top of a specific robot control layer by means of a cognitive process model, i.e., WP2, page 23.

Contributions of HUMAVIPS

With respect to the scientific and technological state of the art, HUMAVIPS will have the following original contributions:

- HUMAVIPS will address **fundamental features of audio-visual integration** within natural environments with all the complexity they imply: moving speakers, background and natural noise confusion, complex social multimodal dialogs, all from **unrestricted** auditory and visual inputs. This approach is rather different that what has been investigated investigated so far: analysis of formal meetings, speech recognition using a head-mounted microphone), or acquisition of linguistic skills through learning.
- HUMAVIPS will put a strong emphasis on **audiovisual scene analysis**, e.g., source

localization and separation, discrimination between humans and artifacts. Moreover, **hearing and vision will be put on an equal footing** and will be combined whenever unimodal stimuli are too weak, too noisy, or too ambiguous. This differs from the currently available technology: the auditory stimulus is used only for speech recognition and auditory analysis is not performed in the context of robotics.

- HUMAVIPS will couple **scientific investigations** with the engineering of a **humanoid robot** in a mutually beneficial way. The use of an anthropomorphic robot will rise several challenges that have not been addressed by previous projects. This will allow the implementation of **reactive behavior through sensorimotor loops involving two rich sensorial modalities**. This differs from available technologies such as the use of a wheeled robot, audio limited to speech, the sensors are worn by humans, robots with no auditory capabilities, etc. Surprisingly, robotics research omitted auditory sensing. We also note that there are very few attempts that make use of a fully programmable humanoid that is commercially available at an affordable price.

- HUMAVIPS's main original contribution will be the ability for a robot to communicate with groups of people in the most natural manner. HUMAVIPS's outcome will be a very general **audio-visual robot concept**: audio-visual integration will be addressed in a much broader sense than it has been done in the past: AV active exploration, AV recognition and characterization of human behavior, and AV interaction and communication.

- HUMAVIPS will aim at demonstrating that cognition could (and should) be studied in true physical situations, embodied onto a fully autonomous humanoid robot.

B.1.3 S/T methodology and associated work plan

B.1.3.1 Overall strategy and general description

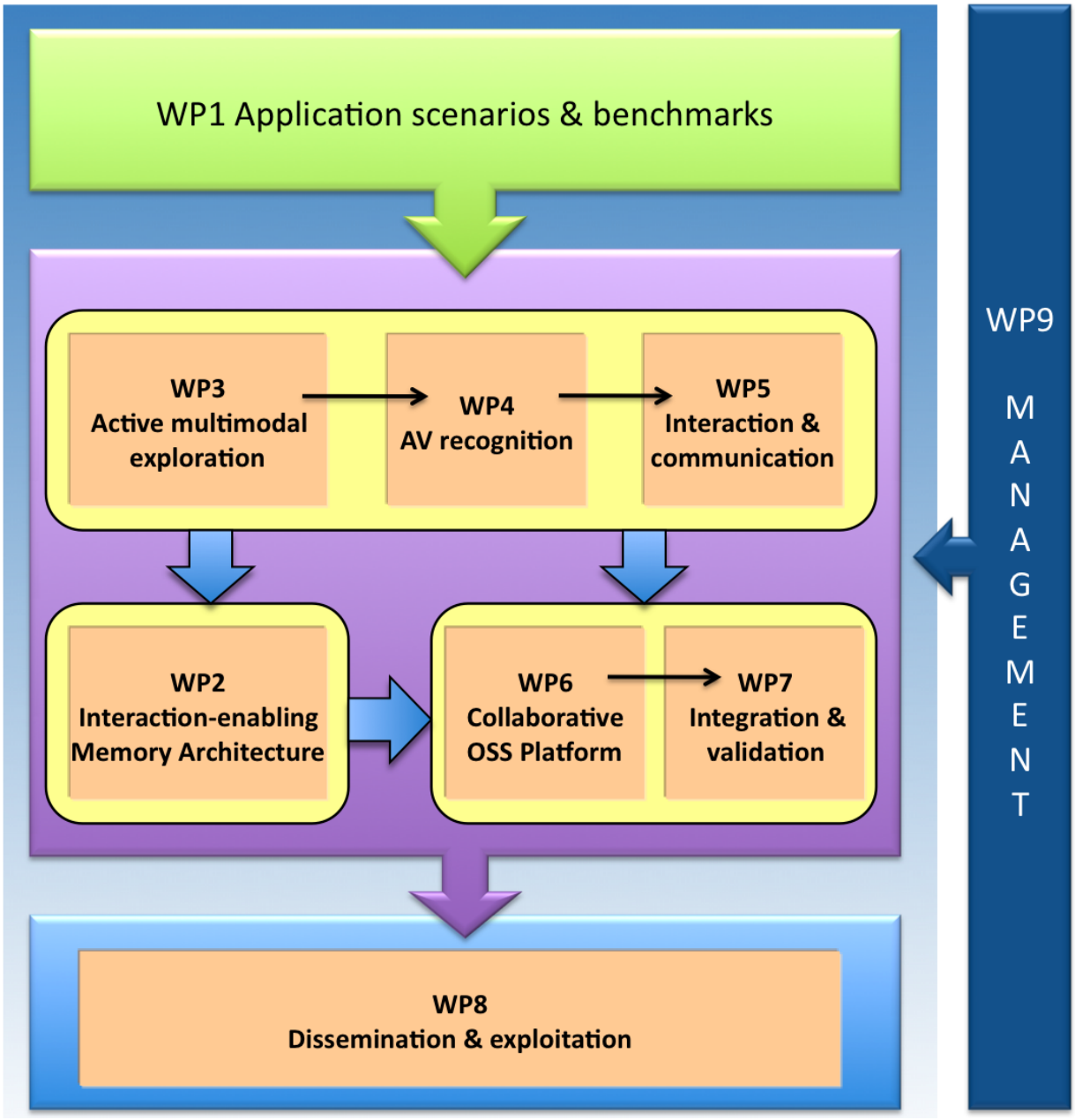
The project objectives will be implemented according to the following work plan. The project's S&T achievements will be guided by WP1 where the application scenarios will be defined in detail, as well as the collection of data sets and their annotation. The scientific and methodological foundations will be developed in four workpackages: WP2, WP3, WP4, and WP5, along the objectives defined by WP1. The outcomes of these workpackages will be twofold: (i) New methods and algorithms for audio-visual humanoid robots and (ii) software packages implementing these algorithms and which will be make publicly available on a open-source software (OSS) platform. The development of the OSS platform itself will be carried out in WP6. The platform will integrate software from all partners. WP6 will use software engineering tools in order to facilitate the integration: maintenance of source code, changes and versions track, percentage of author contribution for each source module, associated documentation, etc. The setup and maintenance of this OSS platform will be key to successful integration between the methods developed in WP2-WP5 and their implementation onto a programmable robot (WP7). Technology (hardware-software) integration, hardware development, and proof-of-concept demonstrators will be achieved in WP7: HUMAVIPS S&T findings will be carefully implemented onto the NAO humanoid robot within the framework of the NaoQi architecture.

NaoQi provides an interface between the developed software modules and the robot hardware. Hence the integration will be done gradually and tested accordingly.

In parallel, WP8 will ensure proper dissemination and exploitation of the project's outcomes and results. WP9 will implement the project management. The milestones MS3, MS4, MS5, and MS6 will assess the progress accross these workpackages. The timing of the workpackages' tasks is shown on page 16. The interconnections between the project's main components are shown on page 17.

B.1.3.2 Timing of workpackages and their components

	TASKS	YEAR 1				YEAR 2				YEAR 3			
WP1	**Application scenarios and benchmarks**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T1.1	Scenarios of project demonstrators												
T1.2	Corpus of annotated data sets												
WP2	**Interaction-enabling Memory Architecture**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T2.1	Memory Spaces & Processes												
T2.2	Interaction-aware Arbitration & Coord.												
T2.3	Modelling, Learning, Activation of Beh.												
WP3	**Active multimodal exploration**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T3.1	Fusion of auditory & visual data												
T3.2	Recognition of audio-visual contexts												
T3.3	Detection of events using AV percepts												
T3.4	Active listening & looking												
T3.5	Audio-visual scene representation												
WP4	**Audio-visual recognition**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T4.1	Spatio-temporal saliency maps												
T4.2	Audio-visual ob ject categorisation												
T4.3	Extraction of low-level non-verbal cues												
T4.4	Extraction of non-verbal social cues												
WP5	**Interaction and communication**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T5.1	Short-term mod. of multiparty comm. beh.												
T5.2	Long-term mod. of multiparty comm. beh.												
T5.3	Modeling group as an interaction partner												
T5.4	Multiparty dialog modeling & management												
WP6	**Collaborative OSS platform**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T6.1	Surveys and recommendations												
T6.2	Collaborative env. def., implemen. & setup												
T6.3	OSS community initiation & animation												
WP7	**Integration and validation**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T7.1	Hardware adaptation of the robotic head												
T7.2	Development of robotic behaviors												
T7.3	Integration of inno. HUMAVIPS functions												
T7.4	Validation of scenarios												
WP8	**Dissemination and exploitation**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T8.1	Dissemination act. (academia, industry, ...)												
T8.2	Exploitation plans for the humanoid platf.												
T8.2	Exploitation plans for software components												
T8.3	Management of IPR & OSS licensing												
WP9	**Management**	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
T9.1	Project control												
T9.2	Administrative & financial management												
T9.3	Quality & documentation management												


WP1 Application scenarios & benchmarks
WP3
Active multimodal exploration
WP4
AV recognition
WP5
Interaction & communication
WP2
Interaction-enabling Memory Architecture
WP6
Collaborative OSS Platform
WP7
Integration & validation
WP8
Dissemination & exploitation
WP9
MANAGEMENT

B.1.3.3 Work package list/overview

WP no	Work package title	Type of activity	Lead part. no.	Lead short name	Person months	Start month	End month
WP1	Application scenarios and benchmarks	RTD	3	ALD	**22**	M1	M27
WP2	Interaction-enabling memory architecture	RTD	5	BIU	**48**	M1	M36
WP3	Active multimodal exploration	RTD	1	INRIA	**53**	M1	M24
WP4	Audio-visual recognition	RTD	2	CTU	**92**	M1	M36
WP5	Interaction and communication	RTD	4	IDIAP	**64**	M6	M36
WP6	Collaborative OSS platform	RTD	1	INRIA	**15**	M1	M36
WP7	Integration and validation	RTD	3	ALD	**71**	M1	M36
WP8	Dissemination and exploitation	OTH	1	INRIA	**12**	M3	M36
WP9	Management	MGT	1	INRIA	**16**	M1	M36
	TOTAL				**393**		

B.1.3.4 Deliverable list

PU*: These deliverables will be publicly available immediately after acceptance of the corresponding scientific publications.

No.	Deliverable name	WP no.	Lead	P-M	Nat.	Diss.	Date
D1.1	Scenario and detailed specification of M12 demonstrator	WP1	ALD	1	R	CO	M3
D9.1	Website	WP9	INRIA	2	R	PU	M3
D2.1	Tutorial on event-driven memory architectures in robotics	WP2	BIU	2	R	PU	M6
D3.1	Methods for independent extraction of auditory and visual features	WP3	CTU	7	R	PU*	M6
D6.1	Strategy for open-source licensing and community development	WP6	INRIA	1	R	CO	M6
D6.2	Collaborative development environment	WP6	INRIA	6	P	CO	M9
D7.1	AV head compatible with the project requirements	WP7	ALD	6	P	CO	M9
D1.2	First corpus with annotated data sets	WP1	INRIA	10	O	PU	M12
D3.2	Methods for audio-visual fusion, context building, and event detection	WP3	INRIA	23	R	PU*	M12
D4.1	Computer vision methods for scene understanding, core implementations	WP4	CTU	20	R	PU*	M12
D5.1	Robot-to-group interactions and dialog interface modeling	WP5	BIU	10	R	PU*	M12
D6.3	First release of OSS packages	WP6	INRIA	3	P	PU	M12
D7.2	First project demonstrator	WP7	INRIA	20	P	PU	M12
D1.3	Scenario and detailed specification of M24 demonstrator	WP1	ALD	0.5	R	CO	M15
D1.4	Second corpus with annotated data sets	WP1	ALD	10	O	PU	M24
D2.2	Arbitration and fusion architecture for social behaviors	WP2	IDIAP	23	R	PU*	M24
D3.3	Methods for audio-visual scene representation with an active observer	WP3	INRIA	23	R	PU*	M24
D4.2	Spatio-temporal saliency maps, methods, implementations	WP4	CTU	40	R	PU*	M24
D5.2	Short-term and long-term multiparty communicative behavior recognition models	WP5	IDIAP	20	R	PU*	M24
D5.3	Multimodal interactive models for robot-to-group dialog management	WP5	BIU	20	R	PU*	M24
D6.4	Second release of OSS packages publicly available	WP6	INRIA	2	P	PU	M24
D7.3	Second project demonstrator	WP7	ALD	20	P	PU	M24

No.	Deliverable name	WP no.	Lead	P-M	Nat.	Diss.	Date
D1.5	Scenario and detailed specification of M36 demonstrator	WP1	ALD	0.5	R	CO	M27
D4.3	Audio-visual functionalities, methods, implementations	WP4	IDIAP	32	R	PU*	M33
D2.3	A memory architecture for a situation-aware humanoid	WP2	BIU	23	R	PU*	M36
D5.4	Communication and interaction models	WP5	IDIAP	14	R	PU*	M36
D6.5	Final release of OSS packages publicly available	WP6	INRIA	3	P	PU	M36
D7.4	Third project demonstrator	WP7	ALD	25	P	PU	M36

B.1.3.5 Work package descriptions

WP1: Application scenarios and benchmarks

WP no.	WP1	Start date or event:			M1
WP name	Application scenario and benchmarks				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	INRIA	CTU	**ALD**	IDIAP	BIU
P-M/part	3	2	9	3	5

Objectives

- Specify three scenarios of gradually increasing complexity. The scenarios will serve as a basis for the project's demonstrators at M12, M24, and M36. Detailed demonstrator specifications based on these scenarios will be available at M3, M15, and M27.
- Design and create annotated audio-visual benchmark data sets allowing both qualitative and quantitative evaluation of the proposed methods. These data-sets will enable the consortium to test their methods and disseminate the results at an early stage of the project. All the annotated data sets will be publicly available.

Approach. Define a set of ambitious scenarios for the project's proof-of-concept demonstrators; these scenarios will be carefully selected to address promising applications issued from the developments of HUMAVIPS. Gather and annotate datasets for benchmarking the scenarios. The datasets will be publicly released.

Description of work

T1.1 Scenarios of project demonstrators (M1-M27). (INRIA, CTU, ALD, IDIAP, BIU). We will describe in detail three scenarios corresponding to three demonstrators of increasing complexity to be delivered by the project (WP7):

1. The *Approaching a person* scenario will demonstrate the ability of a humanoid robot to detect and recognize audio-visual events and patterns, orient itself with respect the environment, and use its on-board sensorimotor capabilities to select a person, to locate him/her, to move towards this person, and to get his/her attention. The robot will operate in a room where several people (their number is not known by the robot) are present. The scenario will be mainly based on the methods developed in WP3 and WP4. This will also demonstrate, at a relatively early stage of the project, the integration between (i) methods and associated software already available at the start of the project, i.e., partners' *background knowledge*, (ii) methods developed during the first year of the project and available on the OSS platform, i.e., WP6, as well as (iii) compatibility with the humanoid platform made available by partner ALD.

2. The *Continuous audio-visual task* scenario will demonstrate the ability of the robot to deal with a continuous **and** varying audio-visual flow of information. For example, the humanoid communicates with a person through both auditory and visual data (gestures, speech and non-speech) while other people wander around. The scenario will deal with a "noisy" environment, e.g., a TV is on, a window is open, etc. The scenario will be mainly based on the methods developed in WP2 and WP5, as well as on the methods available from WP3 and WP4.

3. The *Humanoid as a social companion* scenario will illustrate the overall achievements of the project in an integrated demonstrator. In particular the scenario will demonstrate the new cognitive capabilities developed in HUMAVIPS and based on audio-visual interactions between a humanoid and several people. Imagine, for example, an informal gathering: six to ten people stand up, wander around, talk to each other, while gesticulating with their heads and hands. The humanoid will *join the party* and will attempt to understand the current social situation, and to behave appropriately. The scenario will implement the methods developed in WP2 which will strongly depend on the basic audio-visual exploration, recognition, and communication capabilities developed in WP3, WP4, and WP5.

T1.2 Corpus of annotated data sets (M3-M24). (INRIA, CTU, ALD, IDIAP, BIU) There is a need for annotated audio-visual data sets gathered with camera and microphones on board of a humanoid robot. To date, such robot-centered audio-visual data are not publicly available. In particular there is a crucial need for 3D data because they provide a richer description of a physical world populated with humans, than existing data sets. HUMAVIPS will gather such data sets and put them into a coherent corpus. The data will be appropriately annotated (exact 3D positions and trajectories of auditory sources and of

the visual objects of interest, scripts to describe human actions – such as waving a hand, text corresponding to speech, etc.) The data sets will be used by the project's partners during the development stages (benchmarking and off-line demonstrators). They will be made public for dissemination and benchmarking purposes. The first corpus (D1.2) will be gathered with the POPEYE robot head available at INRIA and developed under the European project POP. This robot has the facilities to record synchronized audio and visual data with two cameras and two microphones. The second corpus (D1.4) will be gathered using NAO equipped with its new head. The humanoid robot will perform a specified trajectory and behavior among a group of people. Ethical issues associated with these data sets are addressed in Section B.4, on page 62.

Risk	Description	Chance	Impact	Contingency plan
R1	Data gathering and data annotation will take more time than predicted.	medium	high	It is crucial for the project that the annotated data are gathered and made available as initially scheduled (M12 and M24). In practice this task (T1.2) will be achieved using INRIA's and BIU's laboratory facilities with the assistance of the associated technical staff. If needed, partners INRIA and BIU can allocate additional resources during the critical periods (M3-M10 and M15-M21) to ensure that the data are gathered, annotated, and delivered as planned.

Deliverables

D1.1 Scenario and detailed specification of M12 demonstrator (M3).

D1.2 First corpus with annotated data sets (M12)

D1.3 Scenario and detailed specification of M24 demonstrator (M15)

D1.4 Second corpus with annotated data sets (M24)

D1.5 Scenario and detailed specification of M36 demonstrator (M27)

WP2: Interaction-enabling memory architecture

WP no.	WP2	Start date or event:			M1
WP name	Interaction-enabling Memory Architecture				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	INRIA	CTU	ALD	IDIAP	**BIU**
P-M/part	-	3	7	2	36

Objectives

- Facilitate the implementation and design of interactive behaviors featuring situation-aware intiative and flexible dialog strategies on the high-level of abstraction on-top of specific robot control layer by means of computational model for cognitive process.
- Provide humanoid-centered arbitration mechanisms on the basis of working memory dynamics that balance socially appropriate behavior and active perception loops.
- Fuse multi-modal perception for socially-aware action selection and learning of memory contents.
- Promote and prove the idea of memories also as a powerful decomposition strategy for building interactive systems and investigate social memories not only as a cognitive foundation but also as a principle to inform the software engineering process in integrated projects w.r.t. decomposition strategies, component decoupling, and distributed and asychnronous processing.

Approach. WP2 will be carried out in close collaboration with WP7 and will define a decomposition strategy into asynchronous memory processes with well-defined memory semantics. Exploiting space-based collaboration techniques in these memories will allow to have access to properties and information regarding all processes, while abstracting from hardware specific lower architectural layers. Active perception loops are working in high-performance short-term memories, while the multi-modal perception is consolidated and integrated in a kind of visuo-spatial sketch pad as a part of the working memory. Inter-memory processes (between semantically different memory types) will be developed that facilitate situation-aware *activiation* of knowledge from long-term memories in order to select appropriate actions.

Description of work

T2.1 Memory spaces & processes (M1-M30). (CTU, BIU) This task will extend prior software frameworks for memory-centered architectures towards social memories further exploiting space-based collaboration and promote a decomposition strategy along a 2.5

layers architecture, featuring long-term memory, short-term (working) memory, and immediate reactive loops. Informed by cognitive foundations, implementations of a visuo-spatial sketch pad towards a "social" sketch pad in the working memory providing temporal fusion and consolidation of perception and recognition from WP3 and WP4 will be developed, exploiting generic memory processes (e.g. following concepts of anchoring [67]). Furthermore, an architectural framework that allows associative reference, "fuzzy" information retrieval, and perceptual consolidation will be investigated. This task will provide the scaffoling for integration and adapt to the interfaces of WP7. In order to ensure early integration and commitment to the developed concepts, WP2 will organize an internal workshop on memory architectures in robotics in the first year of the project, i.e., D2.1 below.

T2.2: Interaction-aware arbitration and coordination (M6-M30). (ALD, BIU) In order to address the challenge of arbitration between interactive behaviors and active perception this task will integrate a *central executive* as a part of the working memory with a focus on humanoid robot control strategies. Situation-aware arbitration and conflict resolution of the sometimes contradictory demands will be in the focus of investigation.

T2.3: Modelling, learning, and activation of behaviors (M6-M36). (IDIAP, BIU) In close collaboration with WP5 this task will investigate and implement appropriate models and representations of socially adequate behaviors (such as entering a gathering). It will exploit the "situational gists" to associate this behaviors with respective social situations. The models of knowledge and behavior activation will allow mixed-initiative driven by either situational gists or commanded by dialogic strategies on the basis of memory processes actively retrieving knowledge from long-term memories via associative lookup and instantiation into working memory representations. Consequently, machine learning techniques will be exploited to enable learning of situation-related behaviors and to provide powerful associative and generic activation.

Risk	Description	Chance	Impact	Contingency plan
R2	The memory driven architecture is difficult to implement in practice.	high	high	Partners BIU and ALD have strong competences in this type of architectures. If needed, partner BIU will bring in additional competence from passed implementations of similar architectures on other humanoid robots.

Deliverables

D2.1 Tutorial on event-driven memory architectures in robotics (M6).

D2.2 Arbitration and fusion architecture for social behaviors (M24).

D2.3 A memory architecture for a sitation-aware humanoid (M36).

WP3: Active multimodal exploration

WP no.	WP3	Start date or event:			M1
WP name	Active multimodal exploration				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	**INRIA**	CTU	ALD	IDIAP	BIU
P-M/part	40	10	3	-	-

Objectives

The humanoid will be able to:

- Fuse audio and visual information in a head-based spatial representation.
- Recognize the context of an everyday scenario from AV cues.
- Detect events characterized by AV percepts.
- Build a spatial map of the environment, including the humanoid's current location.

Approach. Build spatio-temporal audio and visual representations using signal and image processing coupled with the geometry of the sensors (both cameras and microphones). Use unsupervised statistical methods to associate audio and visual observations with higher-level AV events in the spatial domain. Investigate sensorimotor strategies to build a comprehensive model of an AV scene.

Description of work

T3.1 Fusion of auditory and visual data (M1-M12). (INRIA, ALD) Auditory and visual inputs will be *fused*, in order to make use of the spatial and temporal correlation between the two streams. For example, speech sounds will usually be simultaneous with visual motion (such as head movements). Furthermore, if a pair of microphones is used, then the azimuth of an auditory source can be estimated from the inter-aural time-difference (ITD). This azimuth will typically be aligned with the visual direction of the AV source. Finally, if a pair of cameras is used, then the *distance* of the source can be estimated from binocular disparity information. It will be convenient to fuse the AV sources in a *head-based* 3D coordinate system. This representation, as described above, reveals the spatial and temporal correlations in the data. We will refer to temporal and spatial synchrony, two strongly correlated processes. We plan to implement a mixture model with *dynamic model selection* where the number of components, i.e., the number of audio-visual sources

vary over time. This may result in an ambiguous AV scene representation (on probabilistic grounds) and in this case, active listening and looking come into play, where the robot ignores the ambiguous stimuli and attempts to concentrate on the most prominent ones (T3.4). T3.1 will focus on algorithms for robust fusion of the AV data, e.g. in the presence of reverberations. This will require a geometric characterization of the sensors (binocular and binaural), as well as statistical models of the incoming signals and of the fusion process.

T3.2 Recognition of audio-visual contexts (M6-M24). (INRIA, CTU) This task will enable the humanoid to recognize the AV context from a history of stored examples. Furthermore, we will investigate the minimal AV representation that is sufficient for context recognition. In particular, we will consider transformations of the data that reveal certain types of information, while discarding others. The auditory spectrogram, which represents the time/frequency energy (but not phase) is one such transformation. The visual analogue can be found in the *spatial envelope* representation [88]. Another, more flexible, representation of this kind is the *bag of features* [78]. This approach preserves the local structure of the image, but not the global structure. We will investigate the extension of the bag-of-features approach to AV data streams for the description and recognition of contexts. This will pave the road towards the concept of *audio-visual scene gist*.

T3.3 Detection of events using AV percepts (M6-M24). (INRIA, CTU) How can the humanoid localise and identify scene *events* that are both seen and heard? The task will adopt a probabilistic inference principle. We will estimate the joint probability of visual features, auditory features, and audio-visual events. The humanoid will operate in egocentric mode meaning that the observations will be related to the observer centred coordinate system. This tasks will provide basic event detection capabilities needed by WP4. The objects of interest are mainly humans, audio sources and various other AV objects. Detected observations will be represented by features on lower level or labels on higher level of abstraction. These features/labels will be used in WP4, WP5, and WP2.

T3.4 Active listening and looking (M6-M24). (INRIA, ALD) The AV input received by the humanoid is a function of both the external environment and of the humanoid's action. Current research has tended to focus on fixed perceivers. However, psychophysical evidence suggests that humans use small head and body movements, in order to optimize the location of their ears with respect to the source. Similarly, by walking or turning, the humanoid may be able to improve the incoming visual data. For example, in binocular perception, it is desirable to reduce the viewing distance to an object of interest. This allows the 3D structure of the object to be analyzed at a higher depth-resolution. Conflicts could arise between the active listening and looking strategies. For example, the pose of the head that is optimal for listening may compromise the humanoids field of view. This task will implement robust AV source localisation algorithms, which are capable of resolving such conflicts.

T3.5 Audio-visual scene representation (M12-M24). (INRIA, CTU, ALD) In this task the humanoid will construct a spatial representation of the environment, and will be able to localize itself in this map, i.e., AV SLAM. Note that this *scene based* representation is complementary to the *head based* representation developed in T3.1 and will be carried out in parallel with T3.4. At each time instance, i.e., a time interval of approximately 0.2 seconds, the robot will detect AV sources (T3.1). In addition, the robot will have to locate AV objects in 3D and this will be done within a probabilistic approach. Each AV object is a component in a Gaussian mixture model (GMM) whose parameters are the 3D locations (the Gaussian means), the uncertainty in location (the Gaussian covariances) and the AV status (emitting sound or not). The AV SLAM data will include changes in spatial sound source cues as the humanoid moves through an AV scene. The humanoid will learn to localise itself in relation to the AV map that it generates of its environment. The exploratory behavior will be controlled by the action generation model developed in T3.4. Initial work will investigate environments where sound sources are predominately in fixed positions (e.g. noises from artifacts that do not move). The task will progress to tackling more challenging scenes containing mobile sound sources (e.g. humans) where it becomes necessary to distinguish the acoustic foreground (content) from the stationary background (context).

Risk	Description	Chance	Impact	Contingency plan
R3	Extraction and characterization of sensory information is difficult in real conditions, i.e., acoustic reverberations, bad lighting conditions, etc.	high	high	Partners INRIA, CTU and IDIAP have strong competences in robust statistical techniques such as robust mixture models, kernel methods, robust regression, M-estimators, random sampling, etc. Based on this know how, several methodological solutions will be investigated and tested. The experimental conditions will be controlled such that real physical environments of increasing complexity are made available. Hence, the technology can be gradually updated without any impact on the final project outcomes.

Deliverables

D3.1 Methods for independent extraction of auditory and visual features (M6).

D3.2 Methods for audio-visual fusion, context building, and event detection (M12).
D3.3 Methods for audio-visual scene representation with an active observer (M24).

WP4: Audio-visual recognition

WP no.	WP4	Start date or event:			M1
WP name	Audio-visual recognition				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	INRIA	**CTU**	ALD	IDIAP	BIU
P-M/part	10	47	3	32	-

Objectives

- Extraction of low- and mid-level behavioral descriptors for scene objects, with emphasis in the extraction of non-verbal behavioral cues for detected people.
- Categorisation of audio-visual sources in a scene.
- Generation of spatio-temporal saliency maps for behavior analysis.
- Generation of a model to acquire and store knowledge about the co-occurrence of actions and changes in audio-visual sensory input.

Approach. WP4 will develop principled models and efficient algorithms to characterise audio-visual scene objects. Objects include people and artifacts people interact with. Through short-term and long-term observation of the environment, and integrating some of the results from WP2 and WP3, WP4 will develop strategies, ranging from biologically inspired mechanisms to advanced audio-visual processing and probabilistic inference techniques, aimed at describing scene objects by their category (e.g. human vs. non-human) and their behavior. Emphasis will be placed on people, and extracted descriptors will include location, trajectories, and non-verbal cues related to speaking activity and to body language including attentional cues, and basic gestures and actions.

Description of work

T4.1 Spatio-temporal saliency maps for driving behavior analysis (M12-M24) (INRIA) This task will develop computationally efficient saliency maps designed to direct the attention of the computationally more expensive object categorisation and behavior analysis mechanisms in subsequent tasks. The task will use memory maps available from WP2. In particular, the short-term saliency-map concept will be extended to the temporal domain in order to allow detection of salient behavior. Recent work in this area has related

saliency in video to the predictability of the motion tracks of 2D visual features. Extending these ideas to work on a robot platform, where changes in the sensory input may be due to object motion or egomotion, presents a chalenging problem that will be investigated. Moreover, the task will generalise this vision-based approach to the audio modality where it will be used to detect salient features in the development of the spectro-temporal representation of the acoustic scene (e.g., responding to a tonal component with a non-predictable frequency trajectory or an amplitude modulation with an irregular temporal envelope).

T4.2 Audio-visual object categorisation (M1-M33) (CTU, INRIA) Methods to categorise the audio-visual features and events detected and located by the methods developed in WP3 into a small number of meaningful categories will be investigated. We aim at binary classification, human or non-human, as well as multi-class cases where non-human object classes will be further refined. Unlike many existing approaches that build specific models for each object category, we plan to build on top of the concepts and methods investigated in WP3, namely the audio-visual bag of words/features representation, object category models from generic descriptors. The introduction of spatio-temporal context in the bag of words/features model (e.g. while both a television and a person might generate human sounds, the former is not likely to change location over time). We will extend the use of probabilistic topic models for audio-visual object categorisation, and investigate their application in both supervised (classification) and unsupervised (discovery) category learning [92].

T4.3 Extraction of low-level non-verbal behavioral cues (M6-M36) (IDIAP, INRIA, CTU) A challenge is the distant setting of the sensors with respect to the audio sources of interest. While non-verbal audio cues can be reasonably extracted from close-talk microphones, the robotic setting requires the investigation of location-based speaker segmentation and speech enhancement techniques. For audio-visual sources categorised as people, we will study robust techniques to extract a small ensemble of cues known to be effective in characterising speaking activity (e.g. energy, pitch, and speaking rate). Audio information will also help to extract head pose. Regarding visual cues, we will extend methods to extract head and body pose for people in proximity to the robot. Head pose, as a surrogate for gaze, plays a crucial role in behavior modeling. The challenge will be the low-resolution images. We will extend our previous work on appearance-based representations and probabilistic graphical models [46], that jointly estimate location and pose.

T4.4 Extraction of non-verbal social cues (M9-M36) (IDIAP, INRIA, CTU) Two types of social cues are considered: visual focus of attention (VFOA), and basic conversational head and body gestures, and will build on the models and results produced in T4.3. For the former, we will investigate novel generative probabilistic models of gaze, eye/head/body motion, and visual attention, and their application to basic yet rich conversational situations with a moderate number of relevant target visual foci (e.g., looking at the robot vs. looking at another person). We plan to design a cognition-inspired computational

model of VFOA, relying on findings of behavioral research [103, 44, 57] (which describe the relative contribution of head and eye when performing gaze shifts), and on attention models that integrate some of the aspects of audio-visual contextual saliency investigated in WP2 and WP4. In the latter research line, detection and recognition of conversational head and body gestures will be studied using the appearance-based models developed in T4.3. The emphasis will be on natural and simple gestures useful for robot-human interaction (e.g., head nodding to express approval, hand gestures to display an inviting attitude, etc).

Risk	**Description**	**Chance**	**Impact**	**Contingency plan**
R4	Recognition and categorisation methods are too complex to be used in practice by the robot.	medium	high	The implementation of such recognition and categorization methods may require a lot of computer resources both in terms of memory and of processsing units. Currently, the algorithms developed by partners CTU, INRIA and IDIAP run in real-time on a PC cluster. If needed, the robot will be linked to such a PC cluster for the purpose of the project demonstrators and for proving the concept of using these methods within the planned communicative behaviors. In practice, NAO is already equiped with such a facility (physical link to a PC cluster).

Deliverables

D4.1 Computer vision methods for scene understanding, core implementations (M12)

D4.2 Spatio-temporal saliency maps, methods, implementations (M24)

D4.3 Audio-visual functionalities, methods, implementations (M33)

WP5: Interaction and communication

WP no.	WP5	Start date or event:			M6
WP name	Interaction and communication				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	INRIA	CTU	ALD	**IDIAP**	BIU
P-M/part	-	-	3	29	32

Objectives

- Define models for both verbal and non-verbal robot behaviors for active interaction with a group
- Derive appropriate representations of group social situations
- Build predictive models of multiparty communicative behavior, use them to interpret the non-verbal cues extracted in WP4 and maintain the robot social situation representation of the group
- Investigate novel grounding dialog models for robot-to-group interaction
- Deliver a multiparty dialogue manager composing adaptive robot behaviors through mixed verbal and non-verbal interactions

Approach. This WP will deliver a robot-human interface able to handle communication with a group of several people through both non-verbal and verbal behaviors. We will extend the usual robot-to-human communication paradigm to the more challenging robot-to-group of humans case by defining appropriate prototypal multimodal dialog moves for the robot, as well as description of valid social situations (models of admissible sequence of moves) (Task 5.3). The composition and activition of these moves will be handled by the multimodal dialogue manager (T5.4), in relationship with the arbitration and coordination architecture developed in WP2. The robots communicative abilities in the task T5.4 will be informed by a model-based understanding of human-human communicative behavior relying on the probabilistic behavioral models developed in tasks T5.1 and T5.2 which filter the primitive human-behavioral information delivered by WP4 (location, head motion, gaze direction, speech activity, etc.).

Description of work

T5.1 Short-term modelling of multiparty communicative behavior (M12-M30) (IDIAP, BIU, ALD). This task aims to model the short-term communication patterns in multiparty conversations with the goal of continously tracking the conversational role of participants (including the robot) and determining who is talking, who is being addressed and who are the listening people at the present time. These questions cannot be answered

reliably by analysing individual time-slices but require a model of the short-term context. Innovative probabilistic approaches will be developed which employ the multimodal social and behavioral cues extracted by tasks T4.3 and T4.4. In particular, we will investigate dynamic Bayesian networks for multi-person estimation of their location, visual focus, and conversation modes. The models will use a joint multi-person-robot state-space formulation integrating audio-visual cues and the robot's internal dialog state. The model will allow for the explicit definition of interaction models to account for the fact that people in conversations are not fully independent, but react to others' behaviors, and will let the robot infer whether itself or a person is the current addressee for further processing.

T5.2 Long-term modelling of multiparty communicative behavior (M18-M36). (IDIAP) This task will model the longer term patterns of social interaction. Key goals in this task will be (i) assessing the levels of engagement and attention of conversation participants (for instance, to recognise a talkative person or a "hot-spot" moment in a discussion) and (ii) establishing a judgement of the relative social dominance of the interacting people. The discovery and recognition of these patterns often involve the aggregation over time of non-verbal cues exchanged among people in a conversation. We will investigate both static and dynamic statistical models for the estimation of these basic social behavior patterns, which will integrate non-verbal cues extracted in WP4 (speaking turns along with speaking rate and pitch, head and body gestures) along with cue integration models documented in social psychology, which emphasise the relational nature of conversations (e.g. people are in general both speakers and listeners, and the active interplay of such roles in a discussion contribute to the emergence of social saliency patterns).

T5.3 Modeling group as an interaction partner (M6-M24) (BIU). This task will define new models for robot-to-group interaction modeling. This will be achieved by defining a multi-layer set of communication prototypical moves for the robot involving contact (e.g. establishing or breaking contact with the group), attention (e.g. showing attention to a speaker through backchannel signals), conversation turns (requesting, taking or releasing turns) and initiatives using both verbal and non-verbal signals. From this set, studies will be conducted to define and learn appropriate representations of social situations which should be maintained by the robot. This task will address specific issues such as: the development of models for dialog grounding in robot-to-group interactions, based on perceptual level cues rather than understanding; how to detect inconsistent reactions from the group members to a robot move, based on changes of the conversational gists and the modeling of the turn-taking statistical models; how to switch between robot-to-group and robot-to-individual dialog models (e.g. to decide who the robot should address when taking an initiative).

T5.4 Multiparty dialog modeling and management (M6-M36) (BIU,ALD). This task deals with the design and implementation of the necessary components of the dialog system. It will integrate the different perceptual models developed in the task 5.1. to 5.3 with the memory architecture defined in WP2. One important challenge of the project

(addressed from the architecture view point in task T2.3), is the development of flexible compositions of the behavior prototypes defined in the task T5.3 which are linking perception and action, and in particular coordinate active perception loops (e.g. so as to keep an update representation of its environment) with the appropriate social behavior defined in this WP. To address this issue, we will investigate in this task computational models to identify the specific moments to trigger a dialog move and assess the reaction of the group participants using the communicative and social behaviors models developed in tasks 5.1 and 5.2. The goal there is to answer such questions as: *when* is the right moment to take a turn or take an initiative (like asking the group to follow the robot, proposing to take a picture or tell a joke); *how* the robot's behavior (using non-verbal, verbal, or physical actions) can show willingness to interact (e.g. by nodding or moving towards the group) or emphasise its wish to take a turn? An important challenge in this task will be to learn how a dialog strategy (e.g. on taking an initiative or making a physical move) can be adapted, based on the group behavior feedback to the robot's dialog behavior.

Risk	Description	Chance	Impact	Contingency plan
R5	PM allocated to WP5 are insufficient	medium	medium	If needed by the project, IDIAP (partner 4) can reallocate 4-6 PM from WP4 to WP5. Indeed these two work packages are closely related. Simirarly, BIU (partner 5) can reallocate 4-6 from WP2 to WP5 without affecting the outcome of the project.

Deliverables

D5.1 Robot-to-group interactions and dialog interface modeling (M12).

D5.2 Short-term and long-term multiparty communicative behavior recognition models (M24).

D5.3 Multimodal interactive models for robot-to-group dialog management (M24).

D5.4 Communication and interaction models (M36).

WP6: Collaborative OSS platform

WP no.	WP6	Start date or event:			M1
WP name	Collaborative OSS platform				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	**INRIA**	CTU	ALD	IDIAP	BIU
P-M/part	7	2	2	2	2

Objectives

- Implement a collaborative development environment suitable to support the project's software developments as well as developments by a wider community beyond the lifetime of the project.
- Provide releases and updates of the project's software packages.
- Initiate a community of developers and users based on the project developments and results, and foster its growth.
- Establish release and repository content management

Approach. We will start with an in-depth analysis and assessment of relevant components in the following *open-source* domains: governance, community building, licenses and collaborative development tools. Recommendations will be issued to drive the implementation of the collaborative development environment as well as building the HUMAVIPS open-source community. Building and maintaining this community will be the next step. This should guarantee the software dissemination activities as well as sustainability of the project community beyond the lifetime of HUMAVIPS.

Description of work

T6.1 Surveys and recommendations (M1-M6). (INRIA) Provide a clear set of licensing recommendations for successful dissemination and exploitation. Licensing recommendations will result from a survey of existing OSS licenses, sources of legal information pertaining to open source licensing practice, and a consideration of the needs of the project partners while bearing in mind the future wider community. Community building recommendations will result from a survey of existing open-source communities in robotics. We will survey existing environments for the collaborative development of open source software (forges) with an emphasis on the project's specific needs for source code development and document management.

T6.2 Collaborative environment definition, implementation and setup (M3-M9). (INRIA, CTU, ALD, IDIAP, BIU) A collaborative development platform and environment

well suited for the project needs will be realized. This will allow early software integration from WP2, WP3, WP4, and WP5. The core features will include content management, source code management, security mechanisms, publication development tools and knowledge sharing facilities. These features will all be available through a single open source portal, the HUMAVIPS Open Portal (HOP), an open-source software collaborative development environment (OSS-CDE). In addition we will develop features to raise awareness of HUMAVIPS in the research community worldwide, such as tools for maintaining a mailing list, news, a discussion forum, a Wiki, community directory listings, social networking and knowledge sharing services.

T6.3 OSS community initiation and animation (M9-M36). (INRIA, CTU, ALD, IDIAP, BIU) We will launch, develop, and maintain a community of humanoid-robotics software developers. This will be facilitated by the compatibility between the project's OSS and the commercial humanoid robots manufactured by ALD. The needs and expectations of the developers and users of ALD's robots will be identified and analyzed.

Risk	Description	Chance	Impact	Contingency plan
R6	Incompatibility between OSS modules and NAO	medium	high	OSS modules will be tested at an early stage of the project. Project partners will report incompatibilities and the software packages will be updated.

Deliverables

D6.1 Strategy for open-source licensing and community development (M6)
D6.2 Collaborative development environment (M9)
D6.3 First release of OSS packages (M12)
D6.4 Second release of OSS packages publicly available (M24)
D6.5 Final release of OSS packages publicly available (M36)

WP7: Integration and validation

WP no.	WP7	Start date or event:			M1
WP name	Integration and validation				
Act. type	RTD				
Part no.	1	2	3	4	5
Part name	INRIA	CTU	**ALD**	IDIAP	BIU
P-M/part	12	15	26	9	9

Objectives. The global objective of this work package is to build complete humanoid applications according to the scenarios defined in WP1 and based on methods and associated software from WP2, WP3, WP4, WP5, and WP6. The innovative and genuine cognitive skills based on AV abilities developed by the project will be integrated onto an autonomous humanoid platform. In particular WP7 will:

- Develop and implement new audio-visual hardware;
- Interface the hardware-specific architectural elements of the humanoid platform with the general memory architectures of WP2 with respect to sensor abstraction and actuator control;
- Perform continuous integration and validation of the project's scientific achievements, and
- Validate the project's scenarios with live demonstrators.

Approach. The innovative and genuine cognitive and social skills developed by the project will be integrated and validated in the context of the autonomous humanoid platform. This workpackage will be closely related to WP2, linking the client-server architecture and internal memory concepts of NaoQi, the humanoid's architecture, with the cognitive memory architecture. Abstraction will be exploited to ensure generality of the developed concepts and at the same time facilitate rapid development for partners working on their own computer systems while assuring easy migration to the robotic platform in short implementation cycles. Furthermore, WP7 will investigate the specific perceptual requirements defined by WP3 and WP4 to design and implement appropriate audio-visual sensors on the platform, e.g. to support stereo vision and advanced spatial audio capabilities. WP7 will ensure full compatibility between the project's software packages (developed in WP2–WP5 and available on the OSS platform developed in WP6) and the software interface of the humanoid used for experimental and validation purposes. Hardware abstractions and simulation environments along with separate hardware components (such as an audio-visual head) are exploited to reduce risks and ensure integration progress at every partner's institute. In WP7, we will furthermore take care of system evaluation regarding the scenarios defined in WP1 to assess recognition, interaction, and communicative skills in an integrated manner in real world settings.

Description of work

T7.1 Hardware adaptation of the robotic head (M1-M9). (ALD) Systemic performance assessment and approach validation will be performed by all partners on the humanoid robotic platform NAO. The current version of NAO's head uses two cameras and four microphones embedded in its head. The cameras cannot be used for stereovision: their outputs are not synchronized and their fields of view do not overlap. ALD is currently

developing a new head for NAO including a stereovision device and advanced auditory sensing. The sensors' outputs will be fed into the robot's software architecture (NaoQi) with distributed processing capabilities. Nevertheless, HUMAVIPS-specific improvements will be necessary to fit the requirements of WP3, WP4, and WP5. NAO's new head with an AV perception-focused layout to be developed in this task will be used in HUMAVIPS for continuous validation during the development stages and for the planned demonstrators. A head prototype will be delivered to each one of the HUMAVIPS's partners.

T7.2 Development of robotic behaviors (M6-M18). (INRIA, ALD, BIU) Several exploration-, recognition-, and interaction-relevant motor behaviors will be developed, such as: Pointing towards objects (people or artefacts), walking towards a target, performing hand or body gestures to attract human attention, or head gestures such as nodding, shaking, tilting, etc. These behaviors will be developed and implemented based on the requirements in terms of exploration, interaction and dialog of WP3 and WP5.

T7.3 Integration of innovative HUMAVIPS functions (M9-M30). (ALD and BIU) As soon as new functions and abilities are implemented in WP2, WP3, WP4, and WP5, they will be integrated onto NAO's embedded sensors and computers. The architectural memory concept of WP2 will be interfaced with ALD's NaoQi architecture. By joining both architecture that already share many similarities achieved in their respective independent development processes such as the coherent event-based and distributed schemes and distributed processing, the outcome of this task will be with an autonomous system ready for evaluation and supporting continuous integration of updated abilities. Hence, architectural principles investigated in WP2 are foreseen to influence the further evolution of the NaoQi architecture realized within this WP, while likewise the generic memory architecture concepts will be fertilized by the particular requirements of the humanoid platform. In particular, constraints of limited CPU and memory need to be taken into account and feedback to research could be necessary to fit these constraints. Each function implemented on NAO will however be tested independently before being integrated in a complete scenario.

T7.4 Validation of scenarios (M9-M36). (INRIA, CTU, ALD, IDIAP, BIU) This task will combine the outcomes of WP2, WP3, WP4, WP5, available as software packages (WP6) with the annotated datasets and the scenarios specified in WP1. While the annotated data sets will be used for training and for benchmarking, there will be *live demonstrators* that implement the scenarios defined in WP1, page 20:

- First demonstrator: Approaching a person;
- Second demonstrator: Continuous audio-visual task, and
- Third demonstrator: Humanoid as a social companion.

A "demonstration" room will be set up at INRIA and made available to the other partners. The setup of this room will be carefully controlled (the identity, location, and number of

fixed objects as well as the lighting and acoustic conditions). An evaluation protocol will be available with each one of the three scenarios such that it will be possible to validate the experiments. The repeatability of the experimental setup will allow quantitative validation (comparison with the ground truth whenever possible) of the robot performances as well as qualitative validation (cognitive skills and behaviors) as the robot proceeds in real-time.

The real-time demonstrators to be implemented by the project partners will be kept as close as possible the planned scenarios (T1.1) such that comparisons with ground truth are possible. The complexity of the demonstration room (see paragraph above) will be quantified as well as the robot performances. User studies will be performed in order to assess the quality of the actual behaviors.

Risk	Description	Chance	Impact	Contingency plan
R7	Expected robotic applications are too ambitious to be achieved in three years	medium	medium	Implementation of scenarios of gradual complexity. Early tests in real environments will validate the project approach.
R8	Computing resources on board of the robot are too limited to execute complex processes developed in WP2, WP3, WP4 & WP5	low	low	The NaoQi middleware provided with NAO offers the possibility to distribute the software in between the on board computers and a remote PC cluster. ALD has strong experience in real-time distributed software. INRIA has experience in real-time distributed 3D reconstruction and associated applications.

Deliverables

D7.1 AV head compatible with the project requirements (M9).

D7.2 First project demonstrator (M12).

D7.3 Second project demonstrator (M24).

D7.4 Third project demonstrator (M36).

WP8: Dissemination and exploitation

WP no.	WP8	Start date or event:			M3
WP name	Dissemination and exploitation				
Act. type	OTH				
Part no.	1	2	3	4	5
Part name	**INRIA**	CTU	ALD	IDIAP	BIU
P-M/part	4	2	2	2	2

Objectives

- Scientific dissemination towards academia.
- Exploitation and dissemination of software packages.
- Exploitation plans for the project's added-value to robotic platforms.
- Management of intellectual property rights and of open-source software licensing.

Approach. Dissemination both in Europe and worldwide. Scientific journals and conferences will be targeted to promote the scientific results of HUMAVIPS. Three workshops will be organized. The project's demonstrations and main S&T results will be presented at special events. All project outcomes (scientific publications, technical reports, open-source software and documentation, events, videos, etc.) will be available on a website.

Description of work

T8.1 Dissemination activities towards academia, industry, and other users (M3-M36). (INRIA, CTU, ALD, IDIAP, BIU). The scientific publications, technical reports, and project documentation will be regularly updated on the HUMAVIPS public website. Project demonstrations will be recorded and shown at different events. Three workshops will be organized, and there will be liaisons establishment with robotic clubs and challenges, and with other similar EU or national projects and initiatives. T8.1 will implement the dissemination plans.

T8.2 Exploitation plans for the humanoid platform (M12-M36). (ALD). Exploitation of the humanoid platform manufactured by ALD and enhanced with the project's outcomes. T8.2 will implement the exploitation strategy.

T8.3 Exploitation plans for software components (M12-M36) . (INRIA, CTU, ALD, IDIAP, BIU). The software components developed in WP2, WP3, WP4, and WP5 and packaged in WP6 will be exploited in their own right because it will be possible to use them on any programmable robotic platform equiped with cameras and microphones. Recommendations for OSS licensing will be available at an early stage of the project, i.e., WP6, task T6.1 and deliverable D6.1. T8.3 will implement the exploitation strategy.

T8.4 Management of IPR and open-source software licensing (M12-M36) (INRIA, CTU, IDIAP, BIU). This task will implement the management of IPR and OSS licensing.

Risk	Description	Chance	Impact	Contingency plan
R9	Incompatibility between OSS modules and humanoid robots others than NAO	low	low	OSS modules will be distributed at an early stage of the project. Third parties will be able to report incompatibilities

WP9: Management

WP no.	WP9	Start date or event:			M1
WP name	Management				
Act. type	MGT				
Part no.	1	2	3	4	5
Part name	**INRIA**	CTU	ALD	IDIAP	BIU
P-M/part	12	1	1	1	1

Objectives

- Ensuring compliance with EC rules and the Consortium Agreement
- Managing overall legal, ethical, administrative and financial matters
- Meeting the objectives of the project within the agreed budget and timeframe
- Coordinating project activities and ensuring effective internal communication
- Carrying out quality control of the work performed and the deliverables
- Providing adequate information to decision-making bodies

Description of work

T9.1 Project control (M1-M36). (INRIA) INRIA as Coordinator will check the project progress against the planned schedule inputs. It will ensure that the deliverables are properly produced in due time and in respect to the quality plan and contractual commitments. Risks will be tracked and assessed at the occasion of specific milestones as described on page 44. If the work plan or the project objectives are affected by any event, the Steering Committee will take the appropriate measures in order to re-define the work plan or project objectives accordingly, with the approval of EC representatives, and in conformity with contractual aspects. Meetings will be organised for the Steering Committee every 6 months. The Coordinator will inform the Steering Committee about the project progress and about potential problems or conflicts. Specific meetings may be organised to solve issues that may affect the whole project or specific work packages.

T9.2 Administrative and financial management (M1-M36). (INRIA, CTU, ALD, IDIAP, BIU) INRIA is responsible for the administrative work related to the periodic and

final reporting to the EC in accordance with the Grant Agreement schedule. It will consolidate the different reports and will globally maintain the communications between the Commission and the consortium.

The coordinator will receive EC funds and will distribute them to the different partners according to the rules set up in the Grant and Consortium agreements. Realised expenditure will be checked against budget claims and the planned schedule. Certificates on Financial Statements or on Methodology will be collected from the different partners if required. The coordinator is also responsible for maintaining contractual aspects which may be affected during the project progress (mainly the Grant Agreement and Consortium Agreement), including IPR-related matters.

T9.3 Quality and documentation management (M1-M36). (INRIA) INRIA will establish and maintain a quality plan stating standards and procedures for the conduct of the project activities such as distribution of information, document numbering and archiving, review of deliverables, public reports, presentations and publications. It will establish qualitative and quantitative indicators for monitoring the quality of the project activities. INRIA will maintain a documentary system for the project documents (contracts, reports, publications and other relevant documents). An internal project interactive website will be set up at the beginning of the project in order to facilitate communication and exchange of those documents between partners, as well as to track the project progress (deliverables and milestones). A public website will be released at the same period as part of the dissemination strategy. The qualitative and quantitative indicators allowing to measure the project activities:

- Success of project demonstrators based on the three scenarios and quantitative measure based on the annotated data sets and on benchmarks.
- Effectiveness of the integration: evaluation of the performances of the new AV robot head, new robot functionalities based on the project's scientific results, hardware/software compatibility.
- Software contributions co-authored by at least two partners
- Size of the OSS community (number of developers and users outside the project).
- Number of journal and/or conference publications co-authored by at least two partners.
- Number of peer-reviewed journal publications.
- Number of peer-reviewed conference publications.
- Number of workshops, tutorials, courses, etc. organized by the project partners.
- List of dissemination activities by partner ALD.
- Other dissemination activities.

Risk	Description	Chance	Impact	Contingency plan
R10	Delay in the delivery of the five head prototypes (planned at M9)	medium	low	First demonstrator (M12) can use the POPEYE head (INRIA) or the BIRON robot (BIU) for gathering annotated data sets (WP1) and for assessing the results of WP3 and WP4
R11	The timing of workpackages and their components is inadequate	low	low	Timing can be adjusted. The project coordinator and the WP leaders can reallocate the project resources

Deliverables

D9.1 Website (M3).

B.1.3.6 Efforts for the full duration of the project

Indicative efforts per beneficiary per WP

Part no	Short name	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	Total P-M
1	INRIA	3	-	40	10	-	7	12	4	12	**88**
2	CTU	2	3	10	47	-	2	15	2	1	**82**
3	ALD	9	7	3	3	3	2	26	2	1	**56**
4	IDIAP	3	2	-	32	29	2	9	2	1	**80**
5	BIU	5	36	-	-	32	2	9	2	1	**87**
	Total	**22**	**48**	**53**	**92**	**64**	**15**	**71**	**12**	**16**	**393**

Indicative efforts per activity type per beneficiary

Activity type	INRIA	CTU	ALD	IDIAP	BIU	TOTAL
RTD/Innovation activities						
WP1: Application scenarios and benchmark	3	2	9	3	5	22
WP2: Interaction-enabling memory architecture	-	3	7	2	36	48
WP3: Active multimodal exploration	40	10	3	-	-	53
WP4: Audiovisual recognition	10	47	3	32	-	92
WP5: Interaction and communication	-	-	3	29	32	64
WP6: Collaborative OSS platform	7	2	2	2	2	15
WP7: Integration and validation	12	15	26	9	9	71
Total RTD/Innovation	72	79	53	77	84	365
Consortium management activities						
WP9: Management	12	1	1	1	1	16
Total management	12	1	1	1	1	16
Other activities						
WP8: Dissemination and exploitation	4	2	2	2	2	12
Total other	4	2	2	2	2	12
TOTAL beneficiaries	88	82	56	80	87	393

B.1.3.7 List of milestones

Number	Milestone name	WP(s)	Lead	Date	Verification
MS1	Detailed specification of application scenarios	WP1	INRIA	M3	D1.1
MS2	Delivery of five prototypes of a new NAO audiovisual head	WP7	ALD	M9	D7.1
MS3	Assessment of scientific results with the first corpus of annotated datasets	WP1, WP3, WP4	INRIA	M12	D1.2, D7.2
MS4	Assessment of compatibility between OSS modules and the NAO robot	WP6, WP7	INRIA	M24	D1.4, D6.4, D7.3
MS5	Assessment of full compatibility between HUMAVIPS's architecture concept, the NaoQi architecture, and the OSS platform	WP2, WP7	BIU	M24	D2.2
MS6	Assessment of HUMAVIPS S & T results using NAO and the new audiovisual head	WP2, WP5, WP7	INRIA	M24	D7.3

B.2 Implementation

B.2.1 Management structure and procedures

Operational, decision-making and advisory bodies

The operational, decision-making and advisory bodies, together with their definition, functions and responsibilities are outlined below.

1. **The Steering Committee (SC)** comprises one representative from each beneficiary (consortium member) empowered to make decisions on behalf of his/her organisation regarding its participation to the project. The Steering Committee duties include: definition and management of the overall strategy, monitoring of the project progress, conflict solving and project implementation.

2. **The coordinator** (INRIA) is in charge of all legal and financial matters. The operational duties of coordination are assigned to the project leader, assisted by the steering committee and the management support team.

3. **The project leader** is Radu Horaud and he is the single point of contact between the European Commission and the Consortium.

4. **The management support team** (INRIA) will be in charge of organising internal communication among partners and external communication between the consortium and the European commission.

5. **The work package leaders** will be responsible for stimulating and monitoring the performance of their WP and ensuring the quality and timely delivery of their deliverables.

Decision process

It will be the project leader's duty to ensure that the decision making entity of the project is the steering committee. However day-to-day decisions will be made by other entities of the management structure such as WP leaders, in as much as they do not affect the project objectives, the financial plan, and the dissemination or exploitation strategies. Decisions within the steering committee are reached by consensus. In the event that no consensus is reached within a reasonable delay (based on a timely execution of the work programme), decisions will be made by a majority of 2/3. Any conflict internal to a WP will be resolved by consensus within the WP under the guidance of its WP leader. If the problem could harm normal progress of the project, or have a direct impact on other WPs or if it cannot be solved within the WP, the issue will be put to the steering committee. Decision-making mechanisms will be precisely defined in the consortium agreement (CA).

Management procedures

Risk management: The management process shall identify and monitor risks that could have an impact on the project schedule and results and shall take appropriate measures to suppress or mitigate their effects.

IPR, dissemination and exploitation of results: The rules regarding protection and dissemination of knowledge are set out in the consortium agreement.

Quality management: the coordinator is in charge of establishing and maintaining a quality plan stating standards and procedures for the conduct of the project.

Meetings: Meetings of the steering committee will take place every six months.

Reports: Reports on deliverables and milestones will be produced by WP leaders in accordance with the work plan and made available to the steering committee.

Financial reports: The actual effort of each partner will be monitored by the management support team on a bi-annual basis and compared to the work plan. Any major deviation will be discussed by the steering committee.

Summary of project organization

	Administrative management	**Strategic management**	**Executive management**	**Operational activities**
WHAT?	Project and financial reports Communication tools and dashboard	New orientations, conflict solving, corrective actions, budget allocation	Implementations of project through WP, inputs to administrative management	Research and innovation, dissemination and exploitation
HOW?	Following instructions from the Steering Committee, interacts with WP leaders for monitoring and reporting	Top-down decisions based on WP leaders' input	Top-down decisions and bottom-up reporting to the Steering Committee	Joint activities implemented in WPs, meetings, workshops
WHO?	Management Support Team, Coordinator	Steering Committee, WP Leaders	Steering Committee	WP Leaders

B.2.2 Beneficiaries

Partner 1: Institut National de Recherche en Informatique et Automatique (INRIA)

The computer vision group (PERCEPTION) (`http://perception.inrialpes.fr`) has expertise in the field of computer vision, and more specifically for the development of methods for extracting 3D descriptions from images and from videos: geometric and photometric modeling of cameras, binocular- and multiview-stereo, surface reconstruction and representation, motion segmentation, human-motion capture, and recognition of human gestures and actions. Recent PERCEPTION publications relevant to HUMAVIPS are [60, 126, 53, 73, 30, 81, 28, 119, 120, 121, 122].

The statistics group (MISTIS) (`http://mistis.inrialpes.fr`) is specialised in the modelling and inference of complex and structured stochastic systems. More precisely, the group aims at developing statistical methods for dealing with complex systems, complex models, and complex data. The applications that are addressed consist mainly in image processing, multisensory fusion, and spatial data problems, as well as applications in biology and medicine. The methods under consideration involve parametric methods such as mixture models, Markovian models, and more generally hidden structure models as well as semi- and non-parametric methods. MISTIS publications relevant to HUMAVIPS are [41, 23, 39, 22, 15, 14, 24, 40, 115, 38, 13, 71, 72, 3].

Radu Horaud, 22% (PhD, University of Grenoble, 1981) holds a position of director of research at INRIA and he is the scientific leader of the PERCEPTION group. Radu Horaud pioneered work on object recognition using range data and on stereo matching using graph representations and heuristic-search techniques based on geometric constraints. He has contributions in visual sensor calibration, 3-D reconstruction, image-based robot control, graph- and point-matching, and human motion analysis.

Florence Forbes, 10% (PhD, University of Grenoble, 1996) holds a position of senior research scientist at INRIA. Florence Forbes' topics of interest and research activities include Bayesian decision theory applied to image analysis and multi-sensory fusion, Markov processes, Markov random fields, hidden structure models, as well as the development of methodological and algorithmic statistical methods such as Expectation-Maximisation (EM) techniques.

Edmond Boyer (PhD, Institut National Polytechnique de Lorraine, 1996) is assistant professor in computer science at Université Joseph Fourier and a member of PERCEPTION. His interests and topics of research cover both computer vision and computational geometry with emphasis onto multivew stereo and human-action analysis.

Stéphane Ribas (M.Sc, University of Surrey, 1996) joined OW@INRIA in 2008. His expertise is in OSS platform development, support, and dissemination.

Partner 2: The Czech Technical University (CTU)

The HUMAVIPS team comes from the **Center for Machine Perception** (CMP), Faculty of Electrical Engineering. The CMP topics of interests are: computer vision, pattern recognition and mathematics for treating uncertainty. The expertise relevant to the proposal is in analysis of video, learning from the pattern recognition standpoint in both statistical and structural representations, 3D vision in general, human face detection and recognition, effective view-based detection of objects in many images by local invariant descriptive frames.

Vaclav Hlavac, 14% `http://cmp.felk.cvut.cz/~hlavac`. Professor at the Czech Technical University, Faculty of Electrical Engineering, since 1998. Head of the Center for Machine Perception since 1996. His research interests are in 3D computer vision, human motion modeling, relation between statistical and structural pattern recognition and industrial applications of machine vision. Relevant publications: [96], [102], [42], and [43].

Tomas Pajdla, 14% , assistant professor at CTU since 1995, `http://cmp.felk.cvut.cz/~pajdla`. T. Pajdla has experience in geometry and algebra of computer vision, visual robot control, image matching, eye-hand calibration and coordination, precise digital optical measurements, photogrammetry, robot navigation using vision, image matching and object recognition. His publications relevant to the project are [109], [83], [80], [54].

Partner 3: Aldebaran Robotics (ALD)

Aldebaran Robotics (`http://www.aldebaran-robotics.com`) is a French SME created in July 2005 and based in Paris. ALD designs, develops, manufactures, and commercialises humanoid robots and corresponding control software. ALD is the world leader in autonomous humanoid robotics. The company has 8 patents in the fields of mechatronics, software architecture and robot control, and has the following technical expertise: mechanical and kinematics design, motor control, electronics and sensors, bus and low level communication protocols, embedded software development, signal processing, image analysis, bi-pedal walking, and motion planning.

David Gouailler, 15% (PhD, Paris University, 2001) has expertise in robot locomotion and stability. He is one of the co-developers of NAO and is in charge of the robot's kinematic hardware and software architecture. Previously, he worked with SONY's AIBO for on obstacle detection and with HONDA's HRP2 on stability issues.

Vincent Meserette, 30% has 8 years of experience in acoustics, audio and signal processing. He has worked for industry as well as for public laboratories on hardware and software acoustic applications. He joined Aldebaran in 2009 to work on embedded audio software and hardware.

Pierre Emmanuel Viel, 20% has a 6 years of experience in vision processing for automotive application and video suveillance. He has expertise in mixed architectures (DSP, FPGA, ARM, PPC, x86). Since 2008, he works on perception of the robot NAO.

David Houssin, 30% is specialised in real-time design and has 9 years of experience in specification, design and development of real-time embedded systems.

Rodolphe Gelin, 22% worked for 20 years at CEA (Commissariat Energie Atomique) as a researcher in robotics and had mainly been involved in robotic solutions for assistance of disabled people. He served as the head of robotics, virtual reality and cognitive program at CEA. He joined Aldebaran Robotics in December 2008.

Jerome MONCEAU, 22% is specialized in software architecture. He has 5 patents and made several international publications on virtual acoustics.

Partner 4: IDIAP Research Institute (IDIAP)

IDIAP conducts R&D in the areas of speech processing, computer vision, machine learning, multimodal interaction, information retrieval, and biometric authentication, with a total staff of 80 people. In particular, IDIAP has expertise in multiple people localisation and tracking, using audio [77], video [101], and audio-visual [46] sensor data, including the estimation of people visual focus [100] from head pose [5]. IDIAP is also recognised for its research in the automatic analysis of social interaction from multimedia data, modeling of social influence [95] and interest [48] in meetings, recognition of group activities [82], and modeling of human visual attention [4].

Jean-Marc Odobez, 15% (PhD 1994, University of Rennes, France) is a senior researcher at IDIAP. He has worked for several years on the development of statistical models applied to tracking, human activity recognition and multimedia content analysis.

Daniel Gatica-Perez, 10% (PhD 2001, U. of Washington, USA) is a senior researcher at IDIAP. He has led research on audio-visual signal processing and social interaction analysis in meetings from sensor data for several years.

Partner 5: Bielefeld University (BIU)

The Research Institute for Cognition and Robotics (CoR-Lab) at Bielefeld University was founded in July 2007 to bundle the research efforts in cognitive robotics. Research at CoR-Lab considers artificial cognition, computer vision, neural networks, robot learning, and software architectures for cognitive systems. Its research profile is shaped by the conviction that substantial success in Intelligent Robotics requires cross-disciplinary integration of experience in engineering and informatics, brain science, and cognition including the humanities and social science.

Gerhard Sagerer received the diploma and the Ph.D. (Dr.-Ing.) degree in computer science from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1980 and 1985 respectively. Since 1990 he is a professor of computer science at the University of Bielefeld, Germany, and head of the research group for Applied Computer Science. His fields of research are cognitive

and perceptive systems, computer vision, speech understanding, intelligent systems, and pattern recognition in natural science domains.

Britta Wrede, 20% received her Masters degree in Computational Linguistics and her PhD in Computer Science from Bielefeld University in 1999 and 2002, respectively. She received a DAAD PostDoc fellowship at the speech group of the International Computer Science Institute (ICSI) at Berkeley. She is currently working in the field of social and developmental robotics with the goal to enable verbal and non-verbal based communication with a robot.

Sebastian Wrede, 20% received the diploma in computer science from Bielefeld University, Germany, in August 2002. Since 2009 he is heading the research group on Cognitive Systems Engineering at the CoR-Lab. His main research interests are software architectures for the development of cognitive computer (vision) systems and robotic systems. Further research interests are XML and database technologies as well as software and process modeling.

B.2.3 Consortium as a whole

	INRIA	**CTU**	**ALD**	**IDIAP**	**BIU**
Hearing & speech	spatial sound localization			audio processing, microphone arrays	speech recognition, dialog handling
Multimodal interfaces	audiovisual object detection			audiovisual fusion, nonverbal communication	
Statistics & machine learning	mixture models, parametric and nonparametric clustering, MRF	kernel methods, large-scale learning, statistical pattern recognition		dynamical graphical models, MCMC	
Robotics	sensorimotor control, sensor calibration	visual SLAM, vision-robot integration	humanoid design, legged robots, motion synthesis		human-robot interaction
Computer vision	3D reconstruction, human motion analysis, action recognition	object recognition, face detection image-based indexing		human activity recognition	
Integration	OSS platforms		real-time systems, embedded software		cognitive architectures

Partner	Role in the project
1 – INRIA	Coordinator. Leader of WP3, WP6, WP8 and WP9. INRIA brings in strong scientific expertise in computer vision, statistics, audio-visual fusion, and sensor-based robot control (WP3 and WP4). It also brings strong competences in the development of OSS platforms (WP6) and of proof-of-concept demonstrators (WP7). INRIA's main contributions will be in the extraction of visual and auditory cues from the real physical world and the bottom-up building of sensorial representations and of saliency maps using unsupervised statistical methods (WP3 and WP4). INRIA will manage the project (WP9) and will lead the dissemination activities (WP8).
2 – CTU	Leader of WP4. CTU brings in strong expertise in visual object recognition, image-based navigation indexing using statistical pattern recognition methods, and human face detection (WP4). It also has an expertise in visual SLAM to be extended to audiovisual SLAM (WP3). CTU has solid technological expertise in sensor calibration and sensor-robotics integration (WP7).
3 – ALD	Leader of WP1 and WP7. ALD has strong technological expertise in humanoid robotics (control, legged locomotion, action/behavior synthesis) as well as hardware/software integration and embedded software. ALD will specify the scenarios in detail, will supevise the collection and benchmarking of the data sets (WP1) and will lead the project's integration effort and proof-of-concept demonstrators (WP7).
4 – IDIAP	Leader of WP5. IDIAP brings in strong expertise in auditory scene analysis, audio-visual processing and modeling of human communication and behaviour. IDIAP's main scientific contributions will be in the extraction and interpretation of nonverbal cues (WP4 and WP5) and the modeling and automatic analysis of multi-party interaction behaviours for robot dialog situation understanding (WP5).
5 – BIU	Leader of WP2. BIU brings in strong expertise in designing cognitive architecture for humanoid robots (WP2) and in dialog modeling approaches to human-robot interaction (WP5). Given its expertise, BIU will play the role of tying the research in WP2, WP3, WP4, and WP5 with the technological implementation in WP7.

B.2.4 Resources to be committed

Workload per beneficiary per personnel category

Beneficiary	Load	PhD	Postdoc	Sen. res.	Admin. staff	TOTAL
INRIA	in PM	72		12	4	88
CTU	in PM	36	36	10		82
ALD	in PM			56		56
IDIAP	in PM	36	36	8		80
BIU	in PM	36	36	15		87
TOTAL						393

Direct and indirect personnel costs

Beneficiary	Costs	PhD	Postdoc	Sen. res.	Admin. staff	TOTAL
INRIA	direct	217237		44479	12069	273785
	indirect	413175		68863	22954	504992
	Total:	630412		113342	35023	**778777**
CTU	direct	129600	162000	52800		344400
	indirect	77760	97200	31680		206640
	Total:	207360	259200	84480		**551040**
ALD	direct			336000		336000
	indirect			201600		201600
	Total:			537600		**537600**
IDIAP	direct	144000	216000	65600		425600
	indirect	86400	129600	39360		255360
	Total:	230400	345600	104960		**680960**
BIU	direct	151200	156600	75000		382800
	indirect	90720	93960	45000		229680
	Total:	241920	250560	12000		**612480**
	TOTAL:					**3160857**

Workload per beneficiary and per senior researcher in %

Beneficiary	Senior researcher	% of workload
INRIA	Radu Horaud	23%
	Florence Forbes	10%
CTU	Vasek Hlavac	14%
	Tomas Pajdla	14%
ALD	David Gouailler	15%
	Vincent Meserette	30%
	Pierre Emmanuel Viel	20%
	David Houssin	30%
	Rodolphe Gelin	22%
	Jerome Monceau	22%
IDIAP	Jean-Marc Odobez	15%
	Daniel Gatica-Perez	10%
BIU	Britta Wrede	20%
	Sebastian Wrede	20%

Travel expenses [1]

Partner	Project meetings (10)	Workshops	Conferences (10)	Indirect	TOTAL
INRIA	-	-	-	-	-
CTU	10000	-	20000	18000	48000
ALD	10000	-		6000	16000
IDIAP	10000	-	17000	16200	43200
BIU	10000	-	17000	16200	43200
TOTAL					**150400**

Equipment and consumables

Partner	Equipment	Consumables	Indirect	TOTAL
INRIA	10000	-	-	10000
CTU	10000	7000	10200	27200
ALD	-	30000	18000	48000
IDIAP	10000	10000	12000	32000
BIU	-	-	-	-
TOTAL				**117200**

Equipment. ALD will provide to the other project partners the latest version available of NAO with no commercial margin (10000 euros).

INRIA allocated 10000 euros to purchase a NAO robot.

CTU allocated 10000 euros to purchase a NAO robot.

IDIAP allocated 10000 euros to purchase a NAO robot.

Consumables:

ALD allocated 30000 euros for purchasing consumables needed for the development of 5 identical prototypes of a new head to be plugged onto the NAO robot. Four out of these five heads will be delivered to the other partners, as detailed in WP7, task T7.1 (D7.1), page 35.

IDIAP allocated 10000 euros: computers to be dedicated to the project, hardware components for wireless communication with NAO, stereo head for early experiments.

CTU allocated 7000 euors: computers to be dedicated to the project, hardware components for wireless communication with NAO.

[1] INRIA's travel expenses are included in the personnel indirect costs

Subcontracting

Partner	Audit certificates	TOTAL
INRIA	-	-
CTU	1200	1200
ALD	1500	1500
IDIAP	1000	1000
BIU	-	-
TOTAL		**3700**

Overall budget

Partner	Personnel	Travel	Equip. & cons.	Subcontracting	TOTAL
INRIA	778777	-	10000	-	**788777**
CTU	551040	48000	27200	1200	**627440**
ALD	537600	16000	48000	1500	**603100**
IDIAP	680960	43200	32000	1000	**757160**
BIU	612480	43200	-	-	**655680**
TOTAL	**3160857**	**150400**	**117200**	**3700**	**3432157**

Laboratory equipment and facilities. The following resources will be made available by the partners during the lifetime of the project, at no extra-cost:

- INRIA has a modern robotic laboratory equipped, among others, with an audiovisual robotic head (POPEYE). It also has a vision laboratory equipped with a network of cameras linked to a PC cluster.
- CTU has a modern computer vision laboratory with a long-term experience in sensor modeling and video-data acquisition and storage.
- ALD has the adequate facilities to design, develop, and manufacture humanoid robots as well as various kinds of robotic modules.
- IDIAP has a modern multimedia meeting room equiped with audio and visual sensors as well as facilities for audiovisual data acquisition and storage.
- BIU's laboratory is equipped with a wheeled robot (BIRON), a NAO robot, and with hardware and software for speech and dialog.

B.3 Impact

HUMAVIPS will have the following broad contributions:

- Advancement of scientific and technological state of the art in computer vision, auditory & speech processing, multimodal dialogs, robotics, machine learning and human-robot interfaces. On one hand, the research results will be integrated into the robot NAO – a humanoid built by partner ALD and which is already widespread in the scientific community. On the other hand, HUMAVIPS results will be released on an open-source software platform, to foster its use on other robotic devices, thus widening exploitation prospects.

- The ambition to innovate in service robots, and industrial production and manufacturing processes. Aldebaran Robotics, a world leader in autonomous humanoid robotics, is directly involved in the consortium – partner ALD. ALD primarily targets three market segments (robots for academics and for technophiles, robots for entertainment, and personal/assistant robots) but the fully programmable capabilities of ALD's product opens up prospects beyond these market segments.

- The emergence of new market opportunities and technologies. Aldebaran Robotics will directly benefit from the project's results, thus opening new opportunities, as well as other companies which will be able to exploit NAO in conjunction with the OSS platform of HUMAVIPS.

- Design and creation of annotated audio-visual benchmark data sets allowing both qualitative and quantitative evaluation of the proposed methods. These data-sets will be made publicly available to foster their use by the scientific community.

B.3.1 Strategic impact

Brief presentation of the market. There are very few real assistant/personal robots today. It is fair to say that, despite the scientific and technological advances in robotics research, fully useful, robust, and affordable service robots are in their infancy.

All future service robots (domestic, industrial, space, medical,...) require significant audio-visual communication capabilities such that they exhibit intelligent or socially-aware behavior. This can be illustrated by two examples, based on *real business cases* studied by partner ALD and initiated by the end-users themselves:

Personal assistants for the elderly or for people with special needs. A worldwide consortium of insurance companies, asked ALD whether robots could assist dependent or elderly people in their homes, watch them and give alerts in cases of emergency. Current solutions use human companions. However, it was observed that, in 80% of the cases, the human companion sleeping next door does not notice that there is a problem. One solution already explored by insurance companies is to equip a room with webcams and a surveillance system. Often, however, an entire home needs to be equipped to monitor people. Although one could imagine

installing a full remote surveillance system for analysis, they are far from being fully reliable, and installing a network of sensors with their associated communication hardware and software involes a significant cost.

Vision centers targeting visually impaired people are exploring similar applications.

Personal-assistant robots will not be able to perform their assigned tasks without a reliable analysis of their environment, including the ability to understand and analyze humans. This is precisely what HUMAVIPS plans to investigate.

Obviously, there is a large discrepancy between what is commonly expected from a robot, and what robot technologies are able to provide today. This discrepancy is due to several technological challenges that are still unsatisfied:

- There is still progress to be achieved in *mechatronics* in order to yield affordable, safe, autonomous, and large enough robots;
- There are strong expectations from *cognitive behaviour* (voice/speech recognition, person/face detection, recognition of objects of various kind) needing mandatory multimodal approaches, and
- There are challenges associated with *cognitive autonomy* such that robots self-understand what is happening around them, are able to alert, move among people, etc.

The last two challenges are directly connected to HUMAVIPS's main topic: **audio-visual, interaction and communication skills are required to cope with unstructured environments or populated spaces in an autonomous manner**.

Robotic applications would largely benefit from HUMAVIPS results following the project's twofold exploitation strategy. First, the software results will be released as open-source software, most probably under a LGPL license, potentially compatible with many programmable robotic platforms. Secondly, the results will be integrated and demonstrated using ALD's NAO. Indeed, ALD forecasts large sales in the next years, reaching tens of thousand of units. This segment is key, because allowing large diffusion of interactive, yet programmable, robots will help the market move towards advanced service robots. The outcomes will not be limited to robotic systems, but to other application fields such as surveillance systems based on networks of cameras and microphones, and improvement of human-machine interfaces.

To conclude, HUMAVIPS will boost research in several disciplines (computer vision, auditory and speech processing, multimodal integration, robotics, machine learning and human-robot interfaces) and will offer capabilities to integrate and consolidate these disciplines together. Through the implementation of an open-source model for software development in robotics, HUMAVIPS will foster the dissemination and exploitation of the project results in the years to come, with the adoption of new directions, new applications, and new standards, while guaranteeing long-life maintenance beyond the project termination. HUMAVIPS will thus offer the necessary tools to improve the competitiveness of European science and technology: Academics and companies will leverage from this project with their own development on robotics, which would be potentially re-invested in the initial HUMAVIPS code base, therefore improving the overall quality of robotic development and functionalities.

B.3.2 Plan for the dissemination of foreground

The dissemination of the project results will be achieved through the following channels:

- *Publications to international journals and conferences.* The HUMAVIPS academic partners will pursue their publication activities. The following journals and conferences are of particular interest:

  IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Audio, Speech, and Language Processing, IEEE Transactions on Robotics, International Journal of Humanoid Robotics, International Journal of Computer Vision, Speech and Communication, Journal of the Acoustic Society of America, International Journal of Robotics Research, Image and Vision Computing.

  IEEE International Conference on Computer Vision and Pattern Recognition, IEEE International Conference on Robotics and Automation, ACM/IEEE International Conference on Human-Robot Interaction, IEEE International Conference on Acoustics, Speech, and Signal Processing, Interspeech.

- *Project documentation.* During the project lifetime and upon its completion a number of technical documents, scientific papers, deliverables, reports, and presentations will be publicly available on the project website.

- *Project demonstrators.* Three demonstrators of increasing complexity will be implemented. These demonstrations are particularly well-suited for conferences: they can be easily presented off-line as multimedia shows or as live demos with a small humanoid robot. They will be used to present the project in prominent events related to robotic systems and solutions. Also, events related to more specific application domains will be pursued.

- *Project workshops.* Three one-day workshops will be organised during the lifetime of HUMAVIPS. They will involve external users, academic and industrial participants, developers involved in the open-source community created around the project results, and representatives from other national and EU related projets.

- *Project web portal.* The HUMAVIPS web portal will be key to ensuring a wide dissemination of the project results and for animating the HUMAVIPS community. First it will provide information on HUMAVIPS, its objectives, innovation, results and exploitation prospects. In particular, it will include multimedia demos of the project outputs. Second, it will give access to the OSS collaborative platform launched by HUMAVIPS.

- *Business-oriented media.* Partner ALD, as a key player for the exploitation of the project's results, is already communicating through TV, magazines, exhibitions and fairs. It will use these opportunities to specifically disseminate HUMAVIPS results and to raise the project's media profile.

- *Robotics clubs and challenges.* HUMAVIPS's communication will take advantage of existing robotics clubs and robotics challenges like Eurobot.

- *Participation in annual and/or topical concertation meetings* organized by the Commission, usually by coordination actions such as EUCognition and robotics coordination actions.

All the dissemination activities will be regularly monitored by INRIA as WP8 leader. Quantified objectives will be set up at the project kick-off (including: number of publications, public presentations, exhibitions and demonstrations, workshops, etc).

A variety of exploitation and use modalities and channels will be adopted to ensure maximum impact from HUMAVIPS results. HUMAVIPS's key exploitable results are:

- *Methods* and *algorithms* allowing a humanoid robot to fuse audio-visual observations, to extract meaningful information from them in order to characterize several people composing an unstructured environment, and to interact with one or several persons. Those capabilities will rely on a *cognitive architecture* for representing the humanoid's short- and long-term perceptive history as well as its low- and high-level knowledge that are needed for robust robot-to-several-people interactions. Methods will be described in various publications. Software modules will be integrated onto the NAO platform (see below) and released onto an OSS platform (see WP6, page 34). They will be modified and upgraded by the HUMAVIPS community of users and developers.

- A *proof-of-concept demonstrator* illustrating relevant applications and integrated onto the NAO humanoid robot. This will require a hardware adaptation of NAO's robotic head to take into account the project audio-visual sensory requirements, the development of behaviours based on WP1 scenarios and the integration of new functions.

- An *annotated set of corpora of audio-visual data* gathered by a humanoid in a realistic setting, to be used for a qualitative and quantitative evaluation of the proposed methods. This set will be made publicly available to contribute to the definition of criteria for benchmarking system properties in the specific R&D context in which HUMAVIPS fits. It will involve both a protocol to obtain consent from all people recorded in the data according to the highest ethical standards (see section 4 on Ethical issues) and the development of an on-line platform through which data will be available for distribution.

Exploitation prospects include:

- The integration of HUMAVIPS results into the next generations of NAO by ALD (at the hardware and software levels).

- The development of R&D, consulting and support services towards academia and industry, for the development of robotics middleware, solutions and applications involving audio-visual interpretation and communication skills, based on the HUMAVIPS know-how;

- The investigation of other application domains (e.g. surveillance systems), even if they are not at the core of the project initial scenarios.

These prospects require appropriate means to be allocated by the HUMAVIPS partners:

- Partner ALD is fully dedicated to upgrading its NAO robot with enhanced skills, as part of its core strategy for the future years.
- All the academic partners are committed to disseminate the HUMAVIPS's research. They expect to generate activities and associated IPR from the provision of external services to industry or other academic institutions in their respective fields of expertise.
- The user community which will be built around the OSS collaborative platform will play a central role in the dissemination and exploitation of the HUMAVIPS results, by providing feedback on usage scenarios and by amending the software modules.
- A Scientific & Usage Advisory Board will be created to involve industrials and additional academic institutions, both to feed and exploit HUMAVIPS results. It will assess HUMAVIPS strategic orientations by covering a larger range of applications than HUMAVIPS itself. Combined with the feedback user community, those tools will allow the definition of viable exploitation plans taking into account user needs and market possibilities.

Integration of HUMAVIPS results onto NAO robots as a direct commercial application for partner ALD. Aldebaran Robotics' (ALD) core business focuses on the development and sales of companion humanoids. NAO was commercialised in 2008 and current sales (March 2009) reach 250 units, the main clients being academics and technophiles. ALD has a strategic plan which gained credibility thanks to Robocup trustees' choice, as well as support from investment funds. The next step is to develop a second generation of humanoids: interactive, robust and safe personal assistant/companion robots at an affordable price to address the consumer market. This new humanoid generation is planned to be released in 2012.

As far as the market prospects are concerned, ALD foresees sales of 10,000 to 20,000 NAO units per year starting in 2009, and 5,000 to 10,000 units per year of its forthcoming new humanoid with interactive capabilities, starting in 2012. Every generation of humanoids requires more advanced skills. In particular, advanced interaction capacities and the ability to move in the complex human environment of a home are mandatory requirements for this market. HUMAVIPS is an ideal opportunity to develop such new humanoids: the project's results will be gradually integrated onto the commercial units.

Promoting open-source softwares and associated services as opportunities for the academic partners. Academic partners will promote the use of HUMAVIPS OSS modules towards third parties such as other research organisations and the industry. They will also be able to offer value-added consultancy services based on their know-how and scientific and technological expertise. HUMAVIPS results will strengthen their position in robotics-related applications. Indeed, service provision opportunities can take several forms:

- Support services based on HUMAVIPS's OSS, its subsystems, libraries and components;
- Value-added consultancy based on the use of HUMAVIPS-derived products or services;
- Software customisation;

- Technology transfers to industry (especially to SMEs) for the development of the next generation of products and services, in different robotic applications;
- Consultancy based on technological expertise in robotics and HUMAVIPS-based technologies.

This exploitation strategy is particularly compatible with the release of HUMAVIPS results under open-source, as detailed in the next paragraph.

Developing the user community and OSS strategy to foster exploitation paths. HUMAVIPS has adopted an open-source software (OSS) strategy for its research results. The consortium strongly believes that creating an OSS community around the project will leverage and highly contribute to the success of the dissemination and exploitation within industries and academiae. This activity will ensure the results and benefit sustainability beyond project termination.

The software modules will be made compatible with NAO's open architecture used for the project demonstrations. This will significantly facilitate the software re-use in a commercially-available, affordable humanoid platform. Being open, those robots can be freely programmed and upgraded (except for low-level functions), **thus allowing the exploration of new usages and concepts.** Nevertheless, the OSS packages will not be hardware-dependent: The HUMAVIPS academic partners will promote their use with any other programmable humanoid equipped with adequate sensors, end-effectors, processing power, and architecture.

HUMAVIPS OSS licence and code base policy. The OSS intellectual property policies will have a significant impact on dissemination and exploitation prospects:

- The right to redistribute the **source code**, as well as modifications and improvements will facilitate the sharing of HUMAVIPS scientific outcomes amongst the wider community and will attract a substantial crowd of developers to work around the HUMAVIPS concepts.
- The right to use the scientific results in conjunction with the open-source code, combined with redistribution rights, will foster the development of a large population of contributors, including SMEs. This will allow them to build up a market for new robotic applications, as well as to adapt the open-source code to other robotic platforms.

Extending exploitation prospects to other application domains. HUMAVIPS exploitation perspectives are limited neither to robots with AV skills nor to personal/assistants robots. Many different research and technological areas may benefit from advanced sensor-guided skills coupled to cognitive abilities. An example of another prominent application is the surveillance sector. Surveillance systems could be improved by HUMAVIPS developments, such as advanced scene analysis combining the use of auditory and visual information.

Management of intellectual property rights

HUMAVIPS partners will take appropriate measures to protect their IPRs while avoiding obstructions to the exploitation of results. Therefore the HUMAVIPS consortium will adopt a clear rationale for IPR as well as management procedures described below. This will provide a basis for steering the technological activities and for strengthening the collaboration during and after the project completion:

IPR and open-source rationale. Defining an IPR rationale implies mapping IPRs onto the main HUMAVIPS deliverables. Key results are indeed new software technologies, including in particular software for AV exploration, interaction and communication and an adapted NAO head taking into account new audio-visual sensory requirements.

The strategy adopted by the consortium will promote the release of the sofware results under an open-source licence to foster the results dissemination and exploitation amongst a large number of potential users. Such an approach will allow users to use and update the software developed during HUMAVIPS for their own applications, using a commercial affordable robot. In addition, abstraction APIs will allow to adapt those open-source solutions to other robotic platforms.

Software results will probably fall under an LGPL license. This will be further investigated in WP8. Special attention will be given to the software that the partners will bring to the project as part of their background IPR, potentially following different licensing principles. The consortium will make sure that there is no conflict between those licences and the research results to be released as open-source, and that all necessary access rights will be provided to the other partners to foster dissemination and exploitation. This will be specified in the Consortium Agreement.

Appropriate protection measures will be taken by ALD to protect the adaptation of NAO's head. Patents applications could be made if relevant. Access rights shall be granted to all the partners who participated in its elaboration, under fair and reasonable conditions. Releasing this head as a standalone research prototype could be considered, if it makes sense for the scientific community.

The forthcoming generations of humanoid robots (by ALD or by others) may be compatible with this head. In any case, NAO middleware technology and hardware will remain the property of ALD, as part of the background and sideground developed by this SME before and during the project execution. ALD will provide software to ease the integration of the different project modules. It will mainly consist of middleware allowing inter-process communication. The middleware itself will be provided as object code. It will be modified internally by ALD if required by the project participants, but the source code will not be released to preserve the company's interests.

B.4 Ethical issues

Audio-visual datasets

As stated in previous sections, HUMAVIPS will collect an audio-visual data set in the context of the HUMAVIPS scenarios (WP1). The collection of such data will be strictly handled, in order to guarantee the highest ethical standards regarding personal data. Procedures involve:

1. All participants in the data set collection process, and all data provided by the participants will be voluntary.

2. All participants will be required to fill in separate official consent forms to grant authorization to record, access, process, and distribute their data for research purposes. Under no circumstance, data will be collected from people without their explicit and active consent.

3. As an additional mechanism, participants will also retain the right to eliminate (on an off-line basis) any content that, although voluntarily provided, might be regarded after the facts as sensitive. For this purpose, a specific system will be implemented for participants to review their recorded data. Participants will retain the right to remove content from the data set at any time during the data collection process.

4. Content will be distributed inside the HUMAVIPS consortium only when all the above issues are strictly satisfied.

5. The data will be publicly disseminated (review meetings, demonstrations, etc.) only for project evaluation and research purposes, and using a strict data release management procedure.

Members of the HUMAVIPS consortium have previous experience in the collection of large multimodal data sets involving people (e.g. IDIAP as part of the AMIDA project, INRIA as part of the POP project), and have successfully implemented procedures to guarantee that all personal data are ethically handled.

References

[1] Hans Jørgen Andersen, Thomas Bak, and Mikael Svenstrup. Adaptive robot to person encounter : by motion patterns. In *Proceedings of the International Conference on Research and Education in Robotics - EUROBOT 2008 : Mission to Mars*, pages 13–23, 2008.

[2] Ronald C. Arkin. *Behavior-based Robotics*. MIT Press, Cambridge, MA, USA, 1998.

[3] E. Arnaud, H. Christensen, Y-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. P. Horaud. The cava corpus: synchronised stereoscopic and binaural datasets with head movements. In *ACM/IEEE International Conference on Multimodal Interfaces (ICMI'08)*, October 2008.

[4] S. Ba and J-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *accepted for publication in IEEE ICASSP*, Las-Vegas, march 2008.

[5] S.O. Ba and J.-M Odobez. Recognizing human visual focus of attention from head pose in meetings. *IEEE Trans. on System, Man and Cybernetics: part B, Man*, 39(1):16–34, February 2009.

[6] Alan Baddeley. *Human Memory: Theory and Practice*. Psychology Press, Hove, East Sussex, 2002.

[7] Z. Barzelay and Y.Y. Schechner. Harmony in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[8] S. Basu. Conversational scene analysis. *PhD thesis, MIT*, 2002.

[9] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *Proc. IEEE CVPR Int. Workshop on Cues in Communication (CVPR-CUES)*, Kauai, Dec. 2001.

[10] C. Bauckhage, S. Wachsmuth, M. Hanheide, S. Wrede, G. Sagerer, G. Heidemann, and H. Ritter. The visual active memory perspective on integrated recognition systems. *Image and Vision Computing*, 26(1):5–14, January 2006. Special Issue on Cognitive Vision.

[11] A. Bauer, D. Wollherr, and M. Buss. Human-robot collaboration: A survey. *International Journal of Humanoid Robotics*, 5(1), March 2008.

[12] M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[13] J. Blanchet and F. Forbes. Triplet Markov fields for the classification of complex structure data. *IEEE Transactions on Pattern Analyis and Machine Intelligence*, 30(6):1055–1067, 2008.

[14] J. Blanchet, F. Forbes, and C. Schmid. Markov random fields for recognizing textures modeled by feature vectors. In *International Conference on Applied Stochastic Models and Data Analysis*, Brest, France, May 2005.

[15] J. Blanchet, F. Forbes, and C. Schmid. Markov random fields for textures recognition with local invariant regions and their geometric relationships. In *British Machine Vision Conference*, Oxford, UK, September 2005.

[16] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision for sociable robots. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 31:443–454, 2001.

[17] Susan E. Brennan and Eric A. Hulteen. Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Systems*, 8:143–151, 1995.

[18] L. Brethes, P. Menezes, F. Lerasle, J. Hayet, C. LAAS, and F. Toulouse. Face tracking and hand gesture recognition for human-robot interaction. In *IEEE International Conference on Robotics and Automation, ICRA*, volume 2, 2004.

[19] Rodney A. Brooks. A robust layered control system for a mobile robot. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1985.

[20] Joanna Bryson. *Agent Technologies, Infrastructures, Tools, and Applications for E-Services*, volume 2592 of *Lecture Notes in Computer Science*, chapter The Behavior-Oriented Design of Modular Agent Intelligence, pages 61–76. Springer, 2003.

[21] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *Proc. European Chapter of the Association for Computational Linguistics (EACL10)*, pages 51—58, 2003.

[22] 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.

[23] 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.

[24] G. Celeux, F. Forbes, C.P. Robert, and M. Titterington. Deviance Information Criteria for missing data models. With discussion. *Bayesian Analysis*, 1(4):651–706, 2006.

[25] N. Checka, K. Wilson, M. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *IEEE Conf. on Acoustics, Speech, and Signal Processing*, pages 881–884. IEEE, 2004.

[26] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *IEEE Proc.*, 92(3):485–494, 2004.

[27] H. H. Clark. *Arenas of Language Use*. University of Chicago Press, 1992.

[28] M. Cooke, Y-C. Lu, Y. Lu, and R. P. Horaud. Active hearing, active speaking. In *International Symposium on Auditory and Audiological Research (ISAAR 2007)*, Helsingor, Denmark, August 2007.

[29] James L. Crowley, Jolle Coutaz, Gaeten Rey, and Patrick Reignier. Perceptual Components for Context Aware Computing. In *UbiComp 2002*, LNCS 2498, pages 117 – 134. Springer, 2002.

[30] F. Cuzzolin, D. Mateus, D. Knossow, E. Boyer, and R. Horaud. Coherent Laplacian 3D protrusion segmentation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, June 2008.

[31] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nehaniv, A. Sisbot, R. Alami, and T. Siméon. How may i serve you?: a robot companion approaching a seated person in a helping context. In *HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 172–179, New York, NY, USA, 2006. ACM.

[32] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[33] Anind K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, 2001.

[34] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):231–257, 2005.

[35] Starkey Jr. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[36] J. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.

[37] T. Fong, I. Nourbakhsh, and K. Dautenhan. A survey of socially interactive robots. *Robotoics and Autonomous Systems*, 42:143–166, 2003.

[38] F. Forbes and G. Fort. Combining Monte Carlo and mean field like methods for inference in hidden Markov Random Fields. *IEEE Transactions on Image Processing*, 16(3):824–837, 2007.

[39] F. Forbes and N. Peyrard. Hidden Markov model selection based on mean field like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 2003.

[40] F. Forbes, N. Peyrard, C. Fraley, D. Georgian-Smith, D. Goldhaber, and A. Raftery. Model-based region-of-interest selection in dynamic breast MRI. *Journal of Computer Assisted Tomography*, 30(4):675–687, July/August 2006.

[41] 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.

[42] V. Franc and V. Hlaváč. An iterative algorithm learning the maximal margin classifier. *Pattern recognition*, 36(9):1985–1996, September 2003.

[43] V. Franc and V. Hlaváč. Simple solvers for large quadratic programming tasks. In Walter G. Kropatch, Robert Sablatnig, and Allan Handbury, editors, *DAGM 2005: Proceedings of the 27th DAGM Symposium*, number 3663 in LNCS, pages 75–84, Berlin, Germany, 8–9 2005. Springer-Verlag.

[44] E. G. Freedman and D. L. Sparks. Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of Neurophysiology*, 77:2328–2348, 1997.

[45] S. Fujie, K. Fukushima, and T. Kobayashi. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In *Proc. ICARA Int. Conference on Autonomous Robots and Agents*, pages 379—384, 2004.

[46] D. Gatica-Perez, G. Lathoud, J-M. Odobez, and I. McCowan. Audio-visual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):601–616, February 2007.

[47] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(2):601–616, 2007.

[48] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, March 2005.

[49] P. Gieselmann and A. Waibel. What makes human-robot dialogues struggle? In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue (DIALOR)*, Nancy, June 2005.

[50] M. A. Goodrich and A. C. Schultz. Human-robot interaction: A survey. *Foundations and Trends*, 3(1):203–275, 2007.

[51] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Pointing 2004, ICPR International Workshop on Visual Observation of Deictic Gestures*, pages 183–191, 2004.

[52] Edward T. Hall. Proxemics. *Current Anthropology*, 9(2/3):83, 1968.

[53] M. Hansard and R. P. Horaud. Cyclopean geometry of binocular vision. *Journal of the Optical Society of America*, 25(9):2357–2369, September 2008.

[54] M. Havlena, T. Pajdla, and K. Cornelis. Structure from omnidirectional stereo rig motion for city modeling. In *VISAPP 2008 - International Conference on Computer Vision Theory and Applications*. INSTICC - Institute for Systems and Technologies of Information, Control and Communication, Setubal, Portugal, January 2008.

[55] Nick Hawes, Aaron Sloman, Jeremy Wyatt, Michael Zillich, Henrik Jacobsson, Geert-Jan Kruijff, Michael Brenner, Gregor Berginc, and Danijel Skočaj. Towards an integrated robot with multiple cognitive functions. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pages 1548–1553. AAAI Press, 2007.

[56] Nick Hawes and Jeremy Wyatt. Developing intelligent robots with cast. In Martin Hülse and Manfred Hild, editors, *IROS Workshop on current software frameworks in cognitive robotics integrating different computational paradigms*, September 2008.

[57] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *TRENDS in Cog. Sciences*, 9(4):188–194, 2005.

[58] M. Heckmann, F. Berthommier, and K. Kroschel. Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 11:1260–1273, 2002.

[59] M. Heldner and J. Edlund. Prosodic cues for interaction control in spoken dialogue systems. In *Working Papers 52: Proceedings of Fonetik 2006*, pages 53–56, 2006.

[60] R. P. Horaud, M. Niskanen, G. Dewaele, and E. Boyer. Human motion tracking by registering an articulated surface to 3-d points and normals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):158–164, January 2009.

[61] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3d head orientation from a monocular image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 1996.

[62] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Mar. 2008.

[63] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, Augsburg, Sep. 2007.

[64] H. Huttenrauch, K.S: Eklundh, A. Green, and E. Topp. Investigating spatial relationships in human-robot interaction. In *Int. Conference on Intelligent Robots ans Systems, Bejing, China*, Oct 2006.

[65] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, Aug 2005.

[66] Todd R. Johnson. Control in act-r and soar. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 343–348. Erlbaum, 1997.

[67] Kai Jüngling and Marc Hanheide. Fusion of perceptual processes for real-time object tracking. In *Proc. Int. Conf. on Information Fusion*, page to appear, Cologne, 2008. IEEE.

[68] M. Katzenmeier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, State College, PA, Oct. 2004.

[69] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychol (Amst)*, 26(1):22–63, 1967.

[70] L. Kennedy and D. Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. ASRU*, Virgin Islands, Dec. 2003.

[71] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. P. Horaud. Audio-visual clustering for multiple speaker localization. In *5th International Workshop on Machine Learning for Multimodal Interaction (MLMI'08)*, LNCS, pages 86–97. Springer, September 2008.

[72] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. P. Horaud. Detection and localization of 3d audio-visual objects using unsupervised clustering. In *ACM/IEEE International Conference on Multimodal Interfaces (ICMI'08)*, October 2008.

[73] D. Knossow, R. Ronfard, and R. P. Horaud. Human motion tracking with a kinematic parameterization of extremal contours. *International Journal of Computer Vision*, 79(2):247–269, September 2008.

[74] F. Korč and V. Hlaváč. *Detection and Tracking of Humans in Single View Sequences Using 2D Articulated Model*, volume 36 of *Computational Imaging and Vision*, chapter 5, pages 105–130. Springer Verlag, Heidelberg, Germany, 1 edition, 2007.

[75] S. Krach, F. Hegel, B. Wrede, G. Sagerer, T. Binkofski, and F. Kircher. Can machines think? Direct interaction and perspective taking with robots investigated via fMRI. *PLoS ONE*, 3, 07 2008.

[76] A. Kushal, M. Rahurkar, L. Fei-Fei, J. Ponce, and T. Huang. Audio-visual speaker localization using graphical models. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 291–294, Washington, DC, USA, 2006. IEEE Computer Society.

[77] G. Lathoud and J.-M. Odobez. Short-term spatio-temporal clustering applied to multiple moving speakers. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(5):1696–1710, July 2007.

[78] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169– 2178, June 2006.

[79] S. Li. *Multi-modal Interaction Management for a Robot Companion.* Phd, Bielefeld University, Bielefeld, 2007.

[80] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, September 2004.

[81] D. Mateus, R. Horaud, D. Knossow, F. Cuzzolin, and E. Boyer. Articulated shape matching using Laplacian eigenfunctions and unsupervised point registration. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, June 2008.

[82] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.

[83] B. Mičušík and T. Pajdla. Structure from motion with wide circular field of view cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1135–1149, July 2006.

[84] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the Int. Conf. on Multimodal Interfaces*, Trento, Italy, October 2005.

[85] K. Nickel, E. Scemann, and R. Stiefelhagen. 3d-tracking of head and hands for pointing gesture recognition in a human-robot interaction scenario. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 565–570, 2004.

[86] K. Nickel, E. Scemann, and R. Stiefelhagen. 3d-tracking of head and hands for pointing gesture recognition in a human-robot interaction scenario. In *Sixth IEEE International Conference on Automatic Face and Gesture RecognitionSixth*, pages 565–570, 2004.

[87] J-M. Odobez and S. Ba. A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME'07)*, Bejing, July 2007.

[88] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[89] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proc. of International Conference on Multimodal Interface (ICMI'05)*, pages 191–198, Trento, Italy, October 2005.

[90] A. Pentland and A. Madan. Perception of social interest. In *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*, Beijing, Oct. 2005.

[91] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *IEEE Proc.*, 92(3):495–513, 2004.

[92] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(9):1575–1590, September 2007.

[93] D. Ramanan, D.A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):65–81, January 2007.

[94] R. Rienks and D. Heylen. Automatic dominance detection in meetings using support vector machines. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[95] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. Detection and application of influence rankings in small-group meetings. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI), Banff*, Nov. 2006.

[96] M. I. Schlesinger and V. Hlaváč. *Ten lectures on statistical and structural pattern recognition*, volume 24 of *Computational Imaging and Vision*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[97] C. Siagian and L. Itti. Biologically-inspired robotics vision monte-carlo localization in the outdoor environment. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2007.

[98] E.A. Sisbot, R. Alami, T. Simon, K. Dautenhahn, M. Walters, S. Woods, K.L. Koay, and C. Nehaniv. Navigation in the presence of humans. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids 05)*, 2005.

[99] E.A. Sisbot, A. Clodic, L. Urias, M. Fontmarty, L. Brthes, and R. Alami. Implementing a human-aware robot system. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 06)*, 2006.

[100] K. Smith, S. Ba, D. Gatica-Perez, and J.-M. Odobez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 30(7):1212–1229, 2008.

[101] K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, June 2005.

[102] M. Šonka, V. Hlaváč, and R. D. Boyle. *Image Processing, Analysis and Machine Vision*. Thomson, Toronto, Canada, 3 edition, April 2007.

[103] D.L. Sparks. The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience*, 3:952–964, December 2002.

[104] J. M. Speigle and J. M. Loomis. Auditory distance perception by translating observers. In *IEEE Symposium on research frontiers in virtual reality*, pages 92–99, Washington DC, 1993.

[105] Thorsten P. Spexard, Shuyin Li, Britta Wrede, Marc Hanheide, Elin A. Topp, and Helge Hüttenrauch. Interaction awareness for joint environment exploration. In *Proceedings of the Special Session on Situation Awareness in Social Robots at the International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 546–551, Jeju Island, Korea, August 2007. IEEE.

[106] B. Stenger, PRS Mendonca, and R. Cipolla. Model based 3d tracking of an articulated hand. In *Proceedings of the Int. Conference on Computer Vision and Pattern Recognition*, volume 2, pages 310–315, 2001.

[107] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(9):1372–1384, September 2006.

[108] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

[109] T. Svoboda and T. Pajdla. Epipolar geometry for central catadioptric cameras. *International Journal of Computer Vision*, 49(1):23–37, August 2002.

[110] M. Takeuchi, N. Kitaoka, and S. Nakagawa. Timing detection for realtime dialog systems using prosodic and linguistic information. In *Proc. of the International Conference Speech Prosody, SP2004*, pages 529—532, 2004.

[111] S. Thrun and M. Montemerlo. The GraphSLAM algorithm with applications to large-scale mapping of urban structures. *International Journal on Robotics Research*, 25(5/6):403–430, 2005.

[112] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR 2008: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 8, Madison, USA, June 2008. IEEE Computer Society, Omnipress.

[113] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 766–773, New York, NY, USA, 2002. ACM.

[114] J. Vermaak, M. Ganget, A. Blake, and P. Pérez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *ICCV '01: Proceedings of the 8th International Conference on Computer Vision*, pages 741–746. IEEE, 2001.

[115] M. Vignes and F. Forbes. Gene clustering via integrated Markov models combining individual and pairwise features. *IEEE Transactions on Computational Biology and Bioinformatics*, 2008. To appear.

[116] Hans Wallach. The role of head movements and vestibular and visual cues in sound localization. *The Journal of Experimental Psychology*, 27(4):339, 1940.

[117] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *European Conference on Computer Vision*, pages 710–724, 2008.

[118] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32:1177–1207, 2000.

[119] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, Anchorage, Alaska, 2008.

[120] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *Proceedings of the International Conference on Computer Vision*, pages 1–7, Rio de Janeiro, Brazil, 2007. IEEE Computer Society Press.

[121] D. Weinland, R. Ronfard, and E. Boyer. Automatic discovery of action taxonomies from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1639–1645, Washington, DC, USA, 2006. IEEE Computer Society.

[122] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, November/December 2006.

[123] S. Woods, M.L. Walters, K. Koay, and K. Dautenhahn. Methodological issues in HRI: A comparison of live and video-based methods in robot to human approach direction trials. In *Proc. 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, pages 51–58, 2006.

[124] B. Wrede and E. Shriberg. Spotting hotspots in meetings: Human judgments and prosodic cues. In *Proc. Eurospeech*, Geneva, Sep. 2003.

[125] Sebastian Wrede. *An Information-Driven Integration Framework for Cognitive Systems*. PhD thesis, Universität Bielefeld, 2008. to appear.

[126] A. Zaharescu and R. P. Horaud. Robust factorization methods using a gaussian/uniform mixture model. *International Journal of Computer Vision*, 81(3):240–258, March 2009.

[127] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 459–466, 2003.

[128] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 11:1154–1164, 2002.