IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2015

Project-Team PERCEPTION

Interpretation and Modeling of Images and Sounds

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team PERCEPTION

Creation of the Team: 2006 September 01, updated into Project-Team: 2008 January 01

Keywords:

Computer Science and Digital Science:

3.4. - Machine learning and statistics
5.1. - Human-Computer Interaction
5.10.2. - Perception
5.10.5. - Robot interaction (with the environment, humans, other robots)
5.3. - Image processing and analysis
5.4. - Computer vision
5.7. - Audio modeling and processing
8.2. - Machine learning
8.5. - Robotics

Other Research Topics and Application Domains:

5.6. - Robotic systems

1. Members

Research Scientists

Radu Horaud [Team leader, Inria, Senior Researcher, HdR]
Siléye Ba [Inria, granted by ERC Advanced Grant VHIA]
Georgios Evangelidis [Inria, granted by ERC Advanced Grant VHIA]

Faculty Member

Laurent Girin [Grenoble INP, Professor, HdR]

Engineers

Soraya Arias [Inria Research Engineer, 40%]
Fabien Badeig [Inria, granted by EU FP7 STREP EARS]
Quentin Pelorson [Inria, granted by ANR MIXCAM]

PhD Students

Yutong Ban [Inria, granted by ERC Advanced Grant VHIA]
Vincent Drouard [Inria, granted by ERC Advanced Grant VHIA]
Israel Dejene Gebru [Inria, granted by Cordi-S]
Dionyssos Kounades-Bastian [Inria, granted by FP7 STREP EARS]
Stéphane Lathuilière [Inria, granted by ERC Advanced Grant VHIA]
Benoît Massé [Inria, granted by ERC Advanced Grant VHIA]

Post-Doctoral Fellow

Xiaofei Li [Inria, granted by FP7 STREP EARS project]

Visiting Scientists

Christine Evers [Imperial College London, UK, granted by FP7 STREP EARS]
Sharon Gannot [Bar-Ilan University, Israel, granted by ERC Advanced Grant VHIA]

Administrative Assistant

Nathalie Gillot [Inria]

Others

Gianni Delannoye-Vignoble [Inria, Intern, from Feb 2015 until Jul 2015]
Pierre Girardeau [Inria, Intern, from Apr 2015 until Sep 2015]
George Sterpu [Inria, Intern, from Feb 2015 until Jul 2015]

2. Overall Objectives

2.1. Overall Objectives

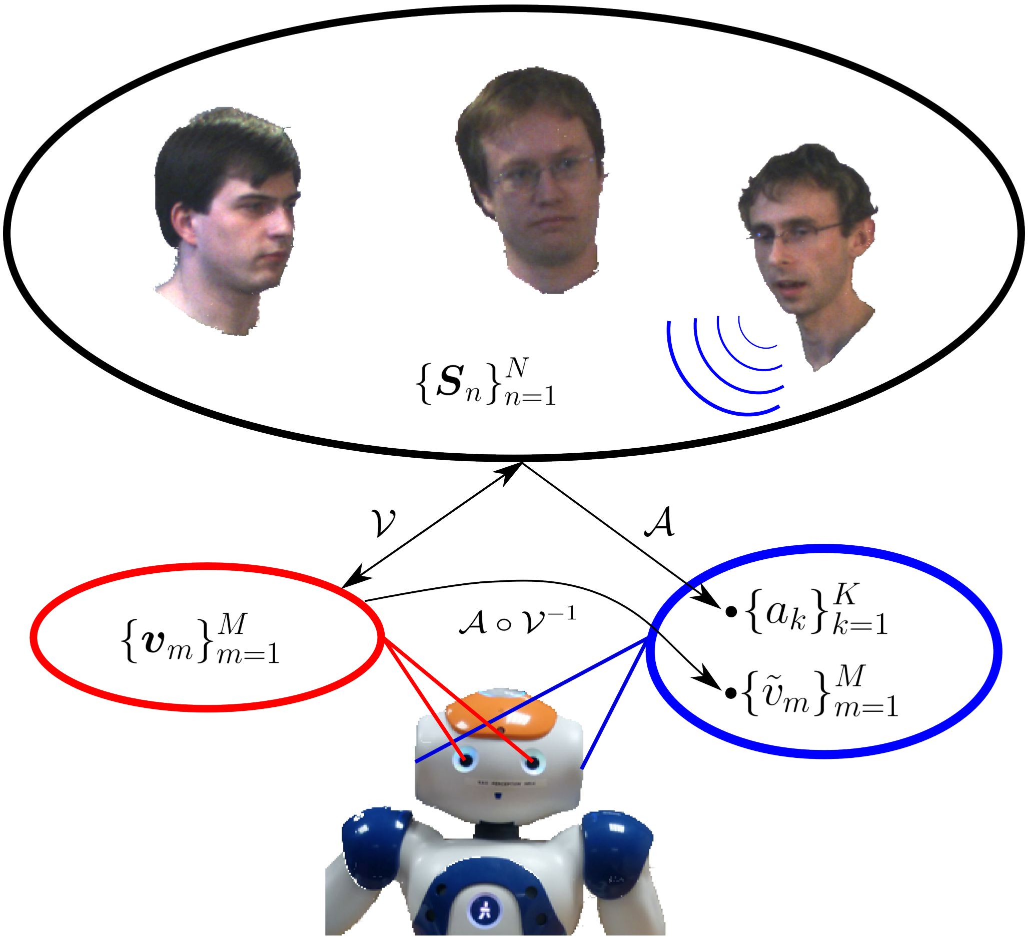
$\{\boldsymbol{S}_n\}_{n=1}^{N}$
$\mathcal{V}$
$\mathcal{A}$
$\{\boldsymbol{v}_m\}_{m=1}^{M}$
$\mathcal{A} \circ \mathcal{V}^{-1}$
$\bullet\{a_k\}_{k=1}^{K}$
$\bullet\{\tilde{v}_m\}_{m=1}^{M}$

Figure 1. This figure illustrates the general principle of the latent-variable mixture models for audio-visual data analysis that the PERCEPTION team have developed [11], [17]. Audiovisual events ($\mathbf{S}$), e.g., speaking faces, are observed with two cameras and two microphones, hence two types of observations are available: 3D binocular features ($\mathbf{v}$) and 1D binaural features (a). By combining the inverse visual mapping with the direct auditory mapping, $\mathcal{A} \circ \mathcal{V}^{-1}$, it is possible to project 3D visual features onto an 1D auditory space, to represent visual and auditory data in the same space, and to properly cluster and classify them.

Auditory and visual perception play a complementary role in human interaction. Perception enables people to communicate based on verbal (speech and language) and non-verbal (facial expressions, visual gaze, head movements, hand and body gesturing) communication. These communication modalities have a large degree of overlap, in particular in social contexts. Moreover, the modalities disambiguate each other whenever one of the modalities is weak, ambiguous, or corrupted by various perturbations. Human-computer interaction (HCI) has attempted to address these issues, e.g., using smart & portable devices. In HCI the user is in the loop for decision taking: images and sounds are recorded purposively in order to optimize their quality with respect to the task at hand.

However, the robustness of HCI based on speech recognition degrades significantly as the microphones are located a few meters away from the user. Similarly, face detection and recognition work well under limited lighting conditions and if the cameras are properly oriented towards a person. Altogether, the HCI paradigm cannot be easily extended to less constrained interaction scenarios which involve several users and whenever is important to consider the *social context*.

The PERCEPTION team investigates the fundamental role played by audio and visual perception in human-robot interaction (HRI). The main difference between HCI and HRI is that, while the former is user-controlled, the latter is robot-controlled, namely *it is implemented with intelligent robots that take decisions and act autonomously*. The mid term objective of PERCEPTION is to develop computational models, methods, and applications for enabling non-verbal and verbal interactions between people, analyze their intentions and their dialogue, extract information and synthesize appropriate behaviors, e.g., the robot waves to a person, turns its head towards the dominant speaker, nods, gesticulates, asks questions, gives advices, waits for instructions, etc. The following topics are thoroughly addressed by the team members: audio-visual sound-source separation and localization in natural environments, for example to detect and track moving speakers, inference of temporal models of verbal and non-verbal activities (diarisation), continuous recognition of particular gestures and words, context recognition, and multimodal dialogue.

3. Research Program

3.1. Audio-Visual Scene Analysis

From 2006 to 2009, R. Horaud was the scientific coordinator of the collaborative European project POP (Perception on Purpose), an interdisciplinary effort to understand visual and auditory perception at the crossroads of several disciplines (computational and biological vision, computational auditory analysis, robotics, and psychophysics). This allowed the PERCEPTION team to launch an interdisciplinary research agenda that has been very active for the last five years. There are very few teams in the world that gather scientific competences spanning computer vision, audio signal processing, machine learning and human-robot interaction. The fusion of several sensorial modalities resides at the heart of the most recent biological theories of perception. Nevertheless, multi-sensor processing is still poorly understood from a computational point of view. In particular and so far, audio-visual fusion has been investigated in the framework of speech processing using close-distance cameras and microphones. The vast majority of these approaches attempt to model the temporal correlation between the auditory signals and the dynamics of lip and facial movements. Our original contribution has been to consider that audio-visual localization and recognition are equally important. We have proposed to take into account the fact that the audio-visual objects of interest live in a three-dimensional physical space and hence we contributed to the emergence of *audio-visual scene analysis* as a scientific topic in its own right. We proposed several novel statistical approaches based on supervised and unsupervised mixture models. The *conjugate mixture model* (CMM) is an unsupervised probabilistic model that allows to cluster observations from different modalities (e.g., vision and audio) living in different mathematical spaces [11], [17]. We thoroughly investigated CMM, provided practical resolution algorithms and studied their convergence properties. We developed several methods for sound localization using two or more microphones [1]. The *Gaussian locally-linear model* (GLLiM) is a partially supervised mixture model that allows to map high-dimensional observations (audio, visual, or concatenations of audio-visual vectors) onto low-dimensional manifolds with a partially known structure [21]. This model is particularly well suited for perception because it encodes both observable and unobservable phenomena. A variant of this model, namely *probabilistic piecewise affine mapping* has also been proposed and successfully applied to the problem of sound-source localization and separation [20]. The European project HUMAVIPS (2010-2013), coordinated by R. Horaud, applied audio-visual scene analysis to human-robot interaction.

3.2. Stereoscopic Vision

Stereoscopy is one of the most studied topics in biological and computer vision. Nevertheless, classical approaches of addressing this problem fail to integrate eye/camera vergence. From a geometric point of view, the integration of vergence is difficult because one has to re-estimate the epipolar geometry at every new eye/camera rotation. From an algorithmic point of view, it is not clear how to combine depth maps obtained with different eyes/cameras relative orientations. Therefore, we addressed the more general problem of binocular vision that combines the low-level eye/camera geometry, sensor rotations, and practical algorithms based on global optimization [5], [13]. We studied the link between mathematical and computational approaches to stereo (global optimization and Markov random fields) and the brain plausibility of some of these approaches: indeed, we proposed an original mathematical model for the complex cells in visual-cortex areas V1 and V2 that is based on steering Gaussian filters and that admits simple solutions [6]. This addresses the fundamental issue of how local image structure is represented in the brain/computer and how this structure is used for estimating a dense disparity field. Therefore, the main originality of our work is to address both computational and biological issues within a unifying model of binocular vision. Another equally important problem that still remains to be solved is how to integrate binocular depth maps over time. Recently, we have addressed this problem and proposed a semi-global optimization framework that starts with sparse yet reliable matches and proceeds with propagating them over both space and time. The concept of seed-match propagation has then been extended to TOF-stereo fusion.

3.3. Audio Signal Processing

Audio-visual fusion algorithms necessitate that the two modalities are represented in the same mathematical space. Binaural audition allows to extract sound-source localization (SSL) information from the acoustic signals recorded with two microphones. We have developed several methods, that perform sound localization in the temporal and the spectral domains. If a direct path is assumed, one can exploit the *time difference of arrival* (TDOA) between two microphones to recover the position of the sound source with respect to the position of the two microphones. The solution is not unique in this case, the sound source lies onto a 2D manifold. However, if one further assumes that the sound source lies in a horizontal plane, it is then possible to extract the azimuth. We used this approach to predict possible sound locations in order to estimate the direction of a speaker [17]. We also developed a geometric formulation and we showed that with four non-coplanar microphones the azimuth and elevation of a single source can be estimated without ambiguity [1]. We also investigated SSL in the spectral domain. This exploits the filtering effects of the head related transfer function (HRTF): there is a different HRTF for the left and right microphones. The interaural spectral features, namely the ILD (interaural level difference) and IPD (interaural phase difference) can be extracted from the short-time Fourier transforms of the two signals. The sound direction is encoded in these interaural features but it is not clear how to make SSL explicit in this case. We proposed a supervised learning formulation that estimates a mapping from interaural spectral features (ILD and IPD) to source directions using two different setups: audio-motor learning [20] and audio-visual learning [22]. Currently we generalize this approach to an arbitrary number of microphones [35], [31]

3.4. Visual Reconstruction With Multiple Color and Depth Cameras

For the last decade, one of the most active topics in computer vision has been the visual reconstruction of objects, people, and complex scenes using a multiple-camera setup. The PERCEPTION team has pioneered this field and by 2006 several team members published seminal papers in the field. Recent work has concentrated onto the robustness of the 3D reconstructed data using probabilistic outlier rejection techniques combined with algebraic geometry principles and linear algebra solvers [16]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [14]. The originality of this work was to represent photometric information as a scalar function over a discrete Riemannian manifold, thus *generalizing image analysis to mesh and graph analysis*. Manifold equivalents of local-structure detectors and descriptors were developed [15]. The outcome of this pioneering work has been twofold: the formulation of a new research topic now addressed by several teams in the world, and allowed us to start a three year collaboration with Samsung Electronics. We developed the novel concept of *mixed camera systems* combining high-resolution color cameras with low-resolution depth cameras [7], [3],[24]. Together with our start-up company 4D Views Solutions and with Samsung, we developed the first practical depth-color multiple-camera multiple-PC system and the first algorithms to reconstruct high-quality 3D content [23].

3.5. Registration, Tracking and Recognition of People and Actions

The analysis of articulated shapes has challenged standard computer vision algorithms for a long time. There are two difficulties associated with this problem, namely how to represent articulated shapes and how to devise robust registration and tracking methods. We addressed both these difficulties and we proposed a novel kinematic representation that integrates concepts from robotics and from the geometry of vision. In 2008 we proposed a method that parameterizes the occluding contours of a shape with its intrinsic kinematic parameters, such that there is a direct mapping between observed image features and joint parameters [12]. This deterministic model has been motivated by the use of 3D data gathered with multiple cameras. However, this method was not robust to various data flaws and could not achieve state-of-the-art results on standard dataset. Subsequently, we addressed the problem using probabilistic generative models. We formulated the problem of articulated-pose estimation as a maximum-likelihood with missing data and we devised several tractable algorithms [10], [9]. We proposed several expectation-maximization procedures applied to various articulated shapes: human bodies, hands, etc. In parallel, we proposed to segment and register articulated shapes represented with graphs by embedding these graphs using the spectral properties of graph Laplacians [18]. This turned out to be a very original approach that has been followed by many other researchers in computer vision and computer graphics.

4. Highlights of the Year

4.1. Highlights of the Year

Robotic Demonstration at ICMI'15. The PERCEPTION team was present at the ACM International Conference on Multimodal Interaction – ICMI'15 (November 2015,Seattle WA, USA) with the demonstration *A Distributed Architecture for Interacting with NAO* [27]. This software package enables robot programming using various languages, e.g. C, C++, Matlab, and Python. This distributed architecture is available under the NAOLab open-source software package. The development of NAOLab is part of PERCEPTION's participation in EU FP7 projects and is funded by STREP project *Embodied Audition for RobotS* (EARS) and ERC Advanced Grant *Vision and Hearing in Action* (VHIA).

The Xerox Foundation University Affairs Committee (UAC) awarded Radu Horaud and Florence Forbes (EPI MISTIS) with a three year grant *Advanced and Scalable Graph Signal Processing Techniques* (2015-2017). Collaboration with Arijit Biswas and Anirban Mondal, research scientists at Xerox Research Center India (XRCI), Bangalore. Information about these awards is available at page 9 of this document available online: http://www.xerox.com/downloads/usa/en/innovation/innovation_xig_brochure.pdf.

MOOC on Binaural Hearing for Robots. In May-June 2015 Radu Horaud taught a five hour MOOC dealing with the fundamental principles of robot hearing, from binaural signal processing to robotic implementations. MOOC content available at https://team.inria.fr/perception/mooc-bhr/ and at https://www.france-universite-numerique-mooc.fr/courses/inria/41004/session01/about.

4.1.1. Awards

- **Vincent Drouard** (PhD student) and his co-authors received the “Best Student Paper Award" (second place) at IEEE ICIP’15 for the paper *Head Pose Estimation via High-Dimensional Regression* . The conference took place in Quebec City, Canada, September 2015. There were five papers awarded, two “Best Paper" and three “Best Student Paper" out of a total of 1033 (oral and poster) papers presented at the conference. IEEE ICIP is the premier international image processing conference series held every year. The work is funded by the ERC Advanced Grant VHIA.
- **Dionyssos Kounades-Bastian** (PhD student) and his co-authors received the “Best Student Paper Award" at IEEE WASPAA’15 for the paper *A Variational EM Algorithm for the Separation of Moving Sound Sources* . The conference took place in New Paltz, NY, USA, October 2015. There were six papers nominated for the award, out of a total of 80 (oral and poster) papers presented at the workshop. The IEEE WASPAA workshop series is among the premier international forums in the field of audio and acoustic signal processing, held every other year. The work is funded by the EU STREP project EARS and the ERC Advanced Grant VHIA.

BEST PAPERS AWARDS:

[28]
V. DROUARD, S. BA, G. EVANGELIDIS, A. DELEFORGE, R. HORAUD. *Head Pose Estimation via Probabilistic High-Dimensional Regression*, in "IEEE International Conference on Image Processing", Quebec City, Canada, Proceedings of the IEEE International Conference on Image Processing, September 2015, https://hal.inria.fr/hal-01163663

[31]
D. KOUNADES-BASTIAN, L. GIRIN, X. ALAMEDA-PINEDA, S. GANNOT, R. HORAUD. *A Variational EM Algorithm for the Separation of Moving Sound Sources*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, NY, United States, IEEE Signal Processing Society, October 2015, https://hal.inria.fr/hal-01169764

5. New Software and Platforms

5.1. Associations of Audio Cues with 3D Locations Library

FUNCTIONAL DESCRIPTION

Library to associate some auditory cues with 3D locations (points). It provides an estimation of the emitting state of each of the input locations. There are two main assumptions:

1. The 3D locations are valid during the acquisition interval related to the audio cues
2. The 3D locations are the only possible locations for the sound sources, no new locations will be created in this module

The software provides also a multimodal fusion library.

- Participants: Xavier Alameda-Pineda, Antoine Deleforge, Jordi Sanchez-Riera and Radu Horaud
- Contact: Radu Horaud

5.2. Supervised Binaural Mapping Software

FUNCTIONAL DESCRIPTION

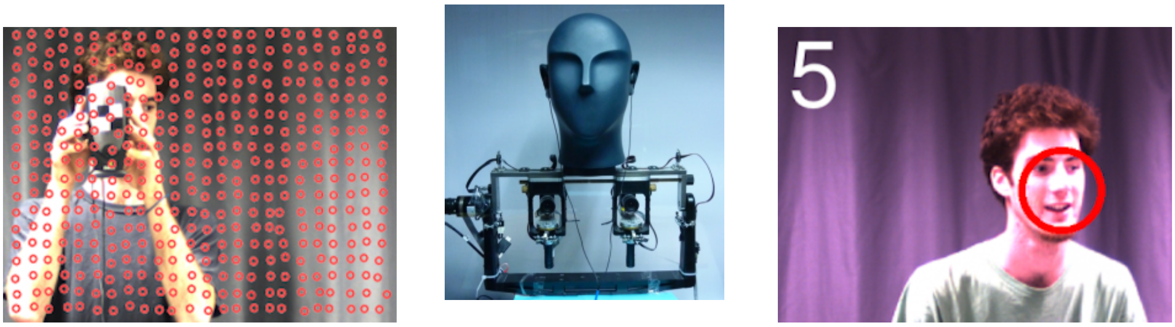

Figure 2. An audio-visual sound source (left) that emits white noise is moved in front of the POPEYE robot (middle). These input-output observation pairs are used to estimate a regression function that is then used to predict the location of a sound (right).

The SBM Matlab toolbox for “Supervised Binaural Mapping", contains a set of functions and scripts for supervised binaural sound source separation and localization. The approach consists in learning the acoustic space of a system using a set of white-noise measurements. Once the acoustic space is learned, it can be used to efficiently localize one or several natural sound sources such as speech, and to separate their signals.

- Participants: Antoine Deleforge, Soraya Arias and Radu Horaud
- Contact: Radu Horaud
- URL: https://team.inria.fr/perception/supervised-binaural-mapping/

5.3. Audiovisual Robotic Heads

FUNCTIONAL DESCRIPTION

The team has developed two audiovisual (AV) robot heads: the POPEYE head and the NAO stereo head. Both are equipped with a binocular vision system and with four microphones. The software modules comprise stereo matching and reconstruction, sound-source localization and audio-visual fusion. POPEYE has been developed within the European project POP in collaboration with the project-team MISTIS and with two other POP partners: the Speech and Hearing group of the University of Sheffield and the Institute for Systems and Robotics of the University of Coimbra. The NAO stereo head was developed under the European project HUMAVIPS in collaboration with Aldebaran Robotics (which manufactures the humanoid robot NAO) and with the University of Bielefeld, the Czech Technical Institute, and IDIAP. The software modules that we develop are compatible with both these robot heads.

- Contact: Radu Horaud
- URL: https://team.inria.fr/perception/popeye/

5.4. MIXCAM Platform

FUNCTIONAL DESCRIPTION

We developed a multiple camera platform composed of both high-definition color cameras and low-resolution depth cameras. This platform combines the advantages of the two camera types. On one side, depth (time-of-flight) cameras provide coarse low-resolution 3D scene information. On the other side, depth and color cameras can be combined such as to provide high-resolution 3D scene reconstruction and high-quality rendering of textured surfaces. The software package developed during the period 2011-2015 contains the calibration of TOF cameras, alignment between TOF and color cameras, TOF-stereo fusion, and image-based rendering.

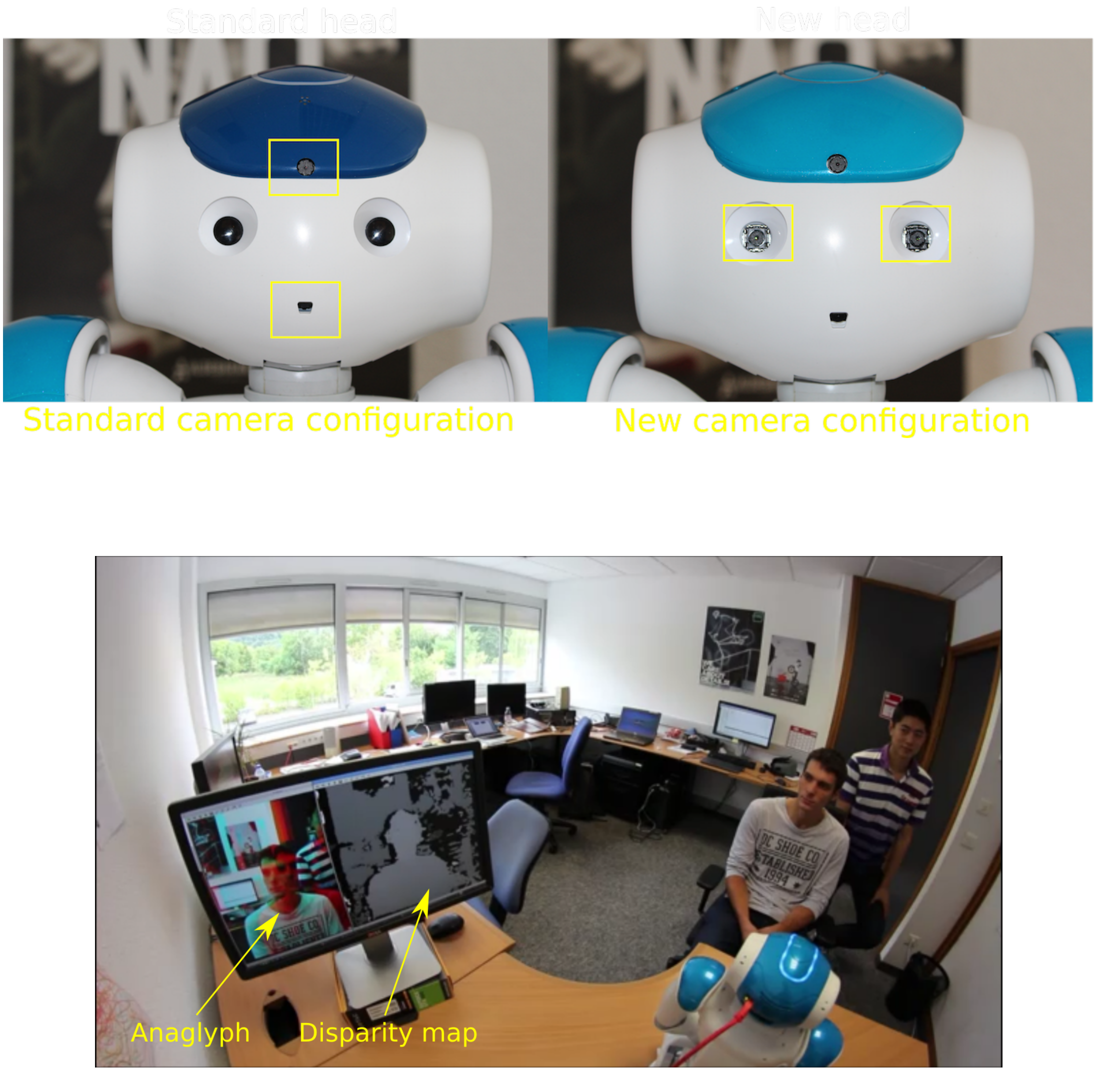

Figure 3. In collaboration with Aldebaran Robotics the team has developed a stereoscopic head for the humanoid robot NAO. Unlike the standard head that has a vertical pair of unsynchronized cameras (top-left), the new head has a horizontal pair of synchronized cameras (top-right). The latest prototype delivers VGA image pairs at 15 FPS. Based on the NAOLab library, we developed a stereo reconstruction method that delivers depth maps at 5 FPS (bottom).

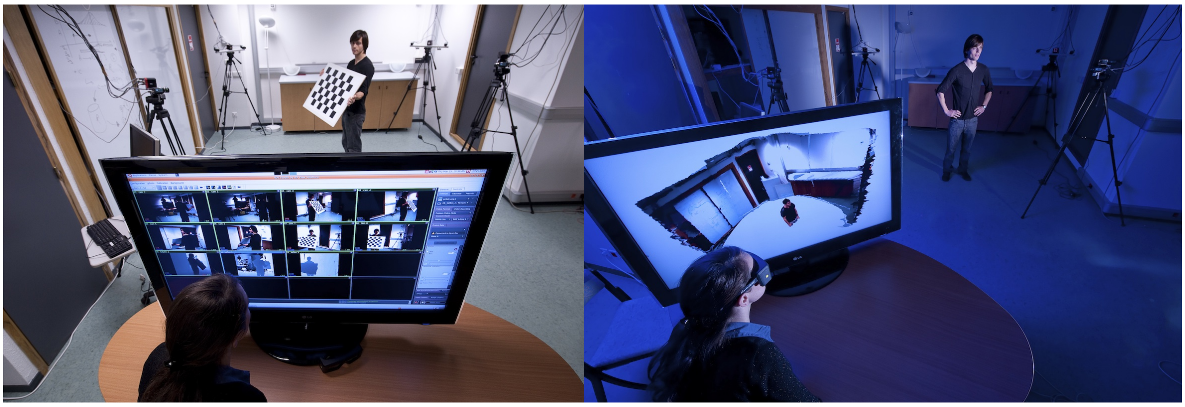

Figure 4. MIXCAM is a multiple-camera multiple-PC hardware/software platform that combines high-resolution color (RGB) cameras with low-resolution time-of-flight (TOF) cameras. The cameras are arranged in TOF-stereo “units", where each unit is composed of two RGB cameras and one TOF camera. Currently the system is composed of four such units, or a total of eight RGB and four TOF cameras. In 2015 we completed algorithms and software packages for the calibration if individual TOF cameras [3] and of the while system composed of four units, e.g. left image, [24]. The system allows high-resolution reconstruction of people, e.g. right image, [23].

These software developments were performed in collaboration with the Samsung Advanced Institute of Technology, Seoul, Korea. The multi-camera platform and the basic software modules are products of 4D Views Solutions SAS, a start-up company issued from the PERCEPTION group.

- Participants: Quentin Pelorson, Georgios Evangelidis, Soraya Arias, Radu Horaud.
- Contact: Radu Horaud
- URL: https://team.inria.fr/perception/mixcam-project/

5.5. NaoLAB

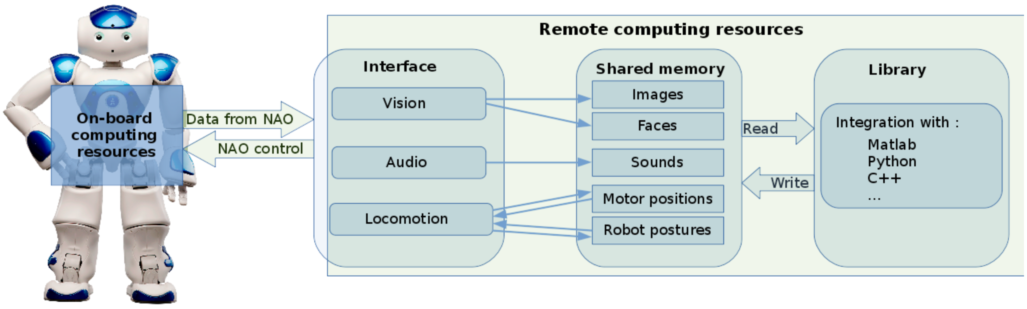

Figure 5. Overview of the proposed distributed architecture that allows fast development of interactive applications using the humanoid robot NAO.

FUNCTIONAL DESCRIPTION

NAOLab [27] is a middleware for the development of robotic applications in C, C++, Python and Matlab, using the humanoid robot NAO networked with a PC. NAOLab enables the joint use of NAO's on-board computing resources and external resources. More precisely, it allows the development of applications that combine embedded libraries, e.g. motion control, image/sound acquisition and transmission, etc., with external toolboxes, e.g. OpenCV, Matlab toolboxes, etc. The NAOLab toolbox has the following characteristic. The middleware complexity is transparent to the users. An user-friendly interface is provided through C++ and Python libraries extended with mex functions for Matlab. This enables the development of sophisticated audio and visual processing algorithms without the stringent constraints of the NAOqi SDK. NAOLab and NAOqi share the same modular approach, namely there are three categories of modules: vision, audio and motion. An interface (vision, audio, motion) is associated with each NAOqi module. Each interface deals with sensor-data access and actuator control. The role of these interfaces is twofold: (i) to feed the sensor data into a memory space that is subsequently shared with existing software or with software under development, and (ii) to send to the robot commands generated by the external modules.

- Participants: Fabien Badeig, Quentin Pelorson, Soraya Arias, Radu Horaud.
- Contact: Radu Horaud
- URL: https://team.inria.fr/perception/research/naolab/

6. New Results

6.1. Supervised Audio-Source Localization

We addressed the problem of localizing audio sources using binaural measurements. After proposing an unsupervised method [20], we proposed a supervised formulation that simultaneously localizes multiple sources at different locations [22]. The approach is intrinsically efficient because, contrary to prior work, it relies neither on source separation, nor on monaural segregation. The method starts with a training stage that establishes a locally-linear Gaussian regression [21] between the directional coordinates of all the sources and the auditory features extracted from binaural measurements. While fixed-length wide-spectrum sounds (white noise) are used for training to reliably estimate the model parameters, we show that the testing (localization) can be extended to variable-length sparse-spectrum sounds (such as speech), thus enabling a wide range of realistic applications. Indeed, we demonstrate that the method can be used for audio-visual fusion, namely to map speech signals onto images and hence to spatially align the audio and visual modalities, thus enabling to discriminate between speaking and non-speaking faces. We release a novel corpus of real-room recordings that allow quantitative evaluation of the co-localization method in the presence of one or two sound sources. Experiments demonstrate increased accuracy and speed relative to several state-of-the-art methods. More recently the method has been extended to an arbitrary number of microphones [35], [34]. Moreover, we have started to develop a method that extracts the direct path on an acoustic wave in order to enable robust audio-source localization in reverberant environments [40].

Websites:
https://team.inria.fr/perception/research/acoustic-learning/
https://team.inria.fr/perception/research/binaural-ssl/
https://team.inria.fr/perception/research/local-rtf/

6.2. Multichannel Audio-Source Separation

We address the problem of separating audio sources from time-varying convolutive mixtures. We proposed an unsupervised probabilistic framework based on the local complex-Gaussian model combined with non-negative matrix factorization. The time-varying mixing filters are modeled by a continuous temporal stochastic process. This model extends the case of static filters which corresponds to static audio sources. While static filters can be learn in advance, e.g. [37], time-varying filters cannot and therefore the problem is more complex. We present a variational expectation-maximization (VEM) algorithm that employs a Kalman smoother to estimate the time-varying mixing matrix, and that jointly estimates the source parameters. The sound sources are then separated by Wiener filters constructed with the estimators provided by the VEM algorithm. Extensive experiments on simulated data show that the proposed method outperforms a block-wise version of a state-of-the-art baseline method. This work is part of the PhD topic of Dionyssos Kounades Bastian and is conducted in collaboration with Sharon Gannot (Bar Ilan University) and Xavier Alameda Pineda (University of Trento). It received the best student paper award at WASPAA'15 [31]. An extended version has been submitted to IEEE Transactions on Audio, Speech, and Language Processing [39].

6.3. Audio-Visual Speaker Tracking and Recognition

Any multi-party conversation system benefits from speaker diarization, that is, the assignment of speech signals among the participants. More generally, in HRI and CHI scenarios it is important to recognize the speaker over time. We propose to address speaker diarization and speaker recognition using both audio and visual data. We cast the diarization problem into a tracking formulation whereby the active speaker is detected and tracked over time. A probabilistic tracker exploits the spatial coincidence of visual and auditory observations and infers a single latent variable which represents the identity of the active speaker. Visual and auditory observations are fused using our recently developed weighted-data mixture model [38], while several options for the speaking turns dynamics are fulfilled by a multi-case transition model. The modules that translate raw audio and visual data into image observations are also described in detail. The performance of the proposed trackers [29], [30] are tested on challenging data-sets that are available from recent contributions which are used as baselines for comparison. Currently we are developing a variational framework for the on-line tracking of multiple persons [36].

Websites:
https://team.inria.fr/perception/research/speakerloc/
https://team.inria.fr/perception/research/speechturndet/
https://team.inria.fr/perception/research/avdiarization/

6.4. Head Pose Estimation

Head pose estimation is an important task, because it provides information about cognitive interactions that are likely to occur. Estimating the head pose is intimately linked to face detection. We addressed the problem of head pose estimation with three degrees of freedom (pitch, yaw, roll) from a single image and in the presence of face detection errors. Pose estimation is formulated as a high-dimensional to low-dimensional mixture of linear regression problem [21]. We propose a method that maps HOG-based descriptors, extracted from face bounding boxes, to corresponding head poses. To account for errors in the observed bounding-box position, we learn regression parameters such that a HOG descriptor is mapped onto the union of a head pose and an offset, such that the latter optimally shifts the bounding box towards the actual position of the face in the image. The performance of the proposed method is assessed on publicly available datasets. The experiments that we carried out show that a relatively small number of locally-linear regression functions is sufficient to deal with the non-linear mapping problem at hand. Comparisons with state-of-the-art methods show that our method outperforms several other techniques. This work is part of the PhD of Vincent Drouard and it received the best student paper award (second place) at the IEEE ICIP'15 [28]. Currently we investigate a temporal extension of this model.

Website:

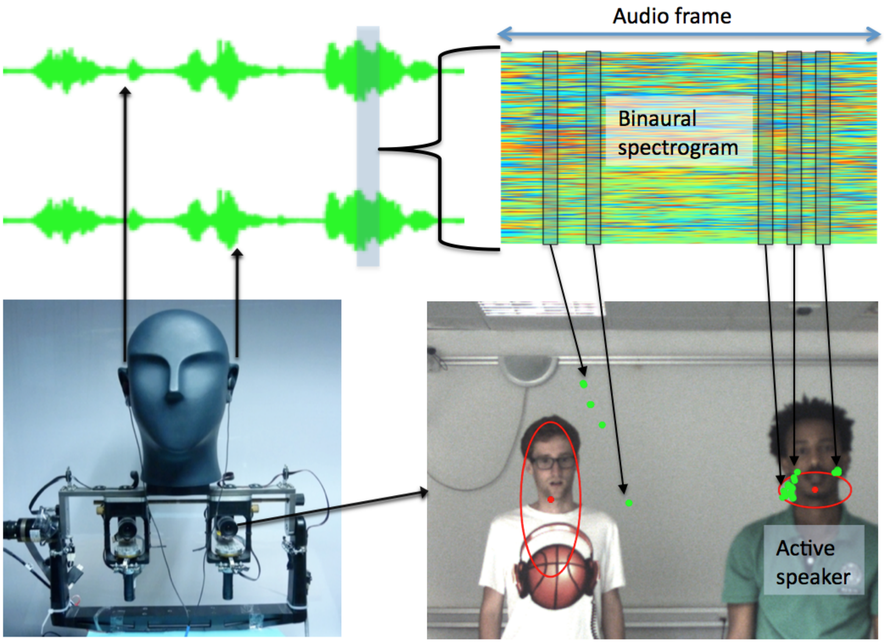

Figure 6. This figures illustrates the general principle of our audio-visual speaker tracking and diarization method. The auditory and visual data are recorded with two microphones and one camera. The audio signals are segmented into frames and each frame (vertical grey rectangle) is transformed into a binaural spectrogram [20]. This spectrogram is composed of a sequence of binaural vectors (vertical rectangles) and each binaural vector is mapped onto a sound-source direction which corresponds to a point in the image plane (green dots) [22]. The proposed audio-visual tracker associates people detected in the image sequence with these sound directions via audio-visual clustering [38] that is combined with an active-speaker transition model.

https://team.inria.fr/perception/research/head-pose/

6.5. High-Resolution Scene Reconstruction

We addressed the problem of range-stereo fusion for the construction of high-resolution depth maps. In particular, we combine low-resolution depth data with high-resolution stereo data, in a maximum a posteriori (MAP) formulation. Unlike existing schemes that build on MRF optimizers, we infer the disparity map from a series of local energy minimization problems that are solved hierarchically, by growing sparse initial disparities obtained from the depth data. The accuracy of the method is not compromised, owing to three properties of the data-term in the energy function. Firstly, it incorporates a new correlation function that is capable of providing refined correlations and disparities, via sub-pixel correction. Secondly, the correlation scores rely on an adaptive cost aggregation step, based on the depth data. Thirdly, the stereo and depth likelihoods are adaptively fused, based on the scene texture and camera geometry. These properties lead to a more selective growing process which, unlike previous seed-growing methods, avoids the tendency to propagate incorrect disparities. The proposed method gives rise to an intrinsically efficient algorithm, which runs at 3FPS on 2.0MP images on a standard desktop computer. The strong performance of the new method is established both by quantitative comparisons with state-of-the-art methods, and by qualitative comparisons using real depth-stereo data-sets [23]. This work is funded by the ANR project MIXCAM.

Website:
https://team.inria.fr/perception/research/dsfusion/

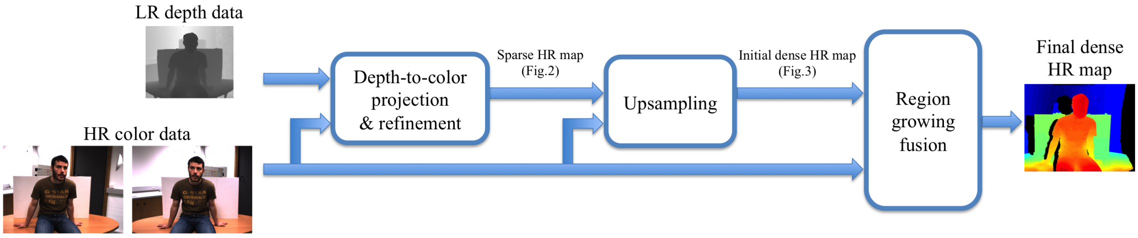

Figure 7. The pipeline of the proposed depth-stereo fusion method [23]. The low-resolution (LR) depth data are projected onto the color data and refined to yield a high-resolution (HR) sparse disparity map. Starting from these disparity seeds, an upsampling process provides an initial HR dense disparity map. Both the HR seeds and the initial dense disparity map are then used by the region-growing depth-stereo fusion to produce the final HR depth map. A prominent feature of our method is that fusion takes place at several data processing stages.

6.6. Hyper-Spectral Image Analysis

As an extension to our work on high-dimensional regression [21] we addressed the problem of analyzing hyper-spectral data. In particular we addressed the problem of recovering physical properties (parameters) form hyper-spectral low-resolution images, i.e. at large planetary scales. This involves resolving inverse problems which can be addressed within machine learning, with the advantage that, once a relationship between physical parameters and spectra has been established in a data-driven fashion, the learned relationship can be used to estimate physical parameters for new hyper-spectral observations. Within this framework, we propose a spatially-constrained and partially-latent regression method which maps high-dimensional inputs (hyper-spectral images) onto low-dimensional responses (physical parameters such as the local chemical composition of the soil). The proposed regression model comprises two key features. Firstly, it combines a Gaussian mixture of locally-linear mappings (GLLiM) with a partially-latent response model. While the former makes high-dimensional regression tractable, the latter enables to deal with physical parameters that cannot be observed or, more generally, with data contaminated by experimental artifacts that cannot be explained with noise models. Secondly, spatial constraints are introduced in the model through a Markov random field (MRF) prior which provides a spatial structure to the Gaussian-mixture hidden variables [19]. Experiments conducted on a database composed of remotely sensed observations collected from the Mars planet by the Mars Express orbiter demonstrate the effectiveness of the proposed model.

6.7. Gaussian Mixture Regression for Acoustic-Articulatory Inversion

The team expertise in latent-variable mixture models was applied to the problem of adaptation of an acoustic-articulatory model of a reference speaker to the voice of another speaker, using a limited amount of audio-only data [25]. In the context of pronunciation training, a virtual talking head displaying the internal speech articulators (e.g., the tongue) could be automatically animated by means of such a model using only the speaker's voice. In this study, the articulatory-acoustic relationship of the reference speaker is modeled by a gaussian mixture model (GMM). To address the speaker adaptation problem, we propose a new framework called cascaded Gaussian mixture regression (C-GMR), and derive two implementations. The first one, referred to as Split-C-GMR, is a straightforward chaining of two distinct GMRs: one mapping the acoustic features of the source speaker into the acoustic space of the reference speaker, and the other estimating the articulatory trajectories with the reference model. In the second implementation, referred to as Integrated-C-GMR, the two mapping steps are tied together in a single probabilistic model. For this latter model, we present the full derivation of the exact EM training algorithm, that explicitly exploits the missing data methodology of machine learning. Other adaptation schemes based on maximum-a posteriori (MAP), maximum likelihood linear regression (MLLR) and direct cross-speaker acoustic-to-articulatory GMR are also investigated. Experiments conducted on two speakers for different amount of adaptation data show the interest of the proposed C-GMR techniques. This work was done in collaboration with Thomas Hueber and Gérard Bailly from Gipsa Lab and with Xavier Alameda-Pineda from University of Trento and former team member.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

In 2015 we started a collaboration with Xerox Research Center India (XRCI), Bangalore. This three-year collaboration (2015-2017) is funded by a grant awarded by the **Xerox Foundation University Affairs Committee (UAC)** and the topic of the project is *Advanced and Scalable Graph Signal Processing Techniques*. The work is done in collaboration with EPI MISTIS and our Indian collaborators are Arijit Biswas and Anirban Mondal.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

8.1.1.1. MIXCAM

Type: ANR BLANC

Duration: March 2014 - February 2016

Coordinator: Radu Horaud

Partners: 4D View Solutions SAS

Abstract: Humans have an extraordinary ability to see in three dimensions, thanks to their sophisticated binocular vision system. While both biological and computational stereopsis have been thoroughly studied for the last fifty years, the film and TV methodologies and technologies have exclusively used 2D image sequences, including the very recent 3D movie productions that use two image sequences, one for each eye. This state of affairs is due to two fundamental limitations: it is difficult to obtain 3D reconstructions of complex scenes and glass-free multi-view 3D displays, which are likely to need real 3D content, are still under development. The objective of MIXCAM is to develop novel scientific concepts and associated methods and software for producing live 3D content for glass-free multi-view 3D displays. MIXCAM will combine (i) theoretical principles underlying computational stereopsis, (ii) multiple-camera reconstruction methodologies, and (iii) active-light sensor technology in order to develop a complete content-production and -visualization methodological pipeline, as well as an associated proof-of-concept demonstrator implemented on a multiple-sensor/multiple-PC platform supporting real-time distributed processing. MIXCAM plans to develop an original approach based on methods that combine color cameras with time-of-flight (TOF) cameras: TOF-stereo robust matching, accurate and efficient 3D reconstruction, realistic photometric rendering, real-time distributed processing, and the development of an advanced mixed-camera platform. The MIXCAM consortium is composed of two French partners (Inria and 4D View Solutions). The MIXCAM partners will develop scientific software that will be demonstrated using a prototype of a novel platform, developed by 4D Views Solutions, and which will be available at Inria, thus facilitating scientific and industrial exploitation.

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

8.2.1.1. EARS

Title: Embodied Audition for RobotS

Program: FP7

Duration: January 2014 - December 2016

Coordinator: Friedrich Alexander Universität Erlangen-Nünberg

Partners:

Aldebaran Roboticss (France)

Ben-Gurion University of the Negev (Israel)

Friedrich Alexander Universitat, Erlangen, Nurenberg (Germany)

Imperial College London (United Kingdom)

Humboldt-Universitat Zu Berlin (Germany)

Inria contact: Radu Horaud

The success of future natural intuitive human-robot interaction (HRI) will critically depend on how responsive the robot will be to all forms of human expressions and how well it will be aware of its environment. With acoustic signals distinctively characterizing physical environments and speech being the most effective means of communication among humans, truly humanoid robots must be able to fully extract the rich auditory information from their environment and to use voice communication as much as humans do. While vision-based HRI is well developed, current limitations in robot audition do not allow for such an effective, natural acoustic human-robot communication in real-world environments, mainly because of the severe degradation of the desired acoustic signals due to noise, interference and reverberation when captured by the robot's microphones. To overcome these limitations, EARS will provide intelligent 'ears' with close-to-human auditory capabilities and use it for HRI in complex real-world environments. Novel microphone arrays and powerful signal processing algorithms shall be able to localise and track multiple sound sources of interest and to extract and recognize the desired signals. After fusion with robot vision, embodied robot cognition will then derive HRI actions and knowledge on the entire scenario, and feed this back to the acoustic interface for further auditory scene analysis. As a prototypical application, EARS will consider a welcoming robot in a hotel lobby offering all the above challenges. Representing a large class of generic applications, this scenario is of key interest to industry and, thus, a leading European robot manufacturer will integrate EARS's results into a robot platform for the consumer market and validate it. In addition, the provision of open-source software and an advisory board with key players from the relevant robot industry should help to make EARS a turnkey project for promoting audition in the robotics world.

8.2.1.2. VHIA

Title: Vision and Hearing in Action

Program: FP7

Type: ERC

Duration: February 2014 - January 2019

Coordinator: Inria

Inria contact: Radu Horaud

The objective of VHIA is to elaborate a holistic computational paradigm of perception and of perception-action loops. We plan to develop a completely novel twofold approach: (i) learn from mappings between auditory/visual inputs and structured outputs, and from sensorimotor contingencies, and (ii) execute perception-action interaction cycles in the real world with a humanoid robot. VHIA will achieve a unique fine coupling between methodological findings and proof-of-concept implementations using the consumer humanoid NAO manufactured in Europe. The proposed multimodal approach is in strong contrast with current computational paradigms influenced by unimodal biological theories. These theories have hypothesized a modular view, postulating quasi-independent and parallel perceptual pathways in the brain. VHIA will also take a radically different view than today's audiovisual fusion models that rely on clean-speech signals and on accurate frontal-images of faces; These models assume that videos and sounds are recorded with hand-held or head-mounted sensors, and hence there is a human in the loop who intentionally supervises perception and interaction. Our approach deeply contradicts the belief that complex and expensive humanoids (often manufactured in Japan) are required to implement research ideas. VHIA's methodological program addresses extremely difficult issues: how to build a joint audiovisual space from heterogeneous, noisy, ambiguous and physically different visual and auditory stimuli, how to model seamless interaction, how to deal with high-dimensional input data, and how to achieve robust and efficient human-humanoid communication tasks through a well-thought tradeoff between offline training and online execution. VHIA bets on the high-risk idea that in the next decades, social robots will have a considerable economical impact, and there will be millions of humanoids, in our homes, schools and offices, which will be able to naturally communicate with us.

8.2.2. Inria International Partners

8.2.2.1. Informal International Partners

- Professor Sharon Gannot, Bar Ilan University, Tel Aviv, Israel,
- Professor Yoav Schechner, Technion, Haifa, Israel,
- Dr. Miles Hansard, Queen Mary University London,
- Dr. Thomas Hueber, Gipsa Lab, CNRS, Grenoble,
- Professor Daniel Gatica Perez, IDIAP Institute, Martigny, Switzerand,
- Professor Nicu Sebe, University of Trento, Trento, Italy,
- Professor Adrian Raftery, University of Washington, Seattle, USA.
- Dr. Zhengyou Zhang, Microsoft, Redmond WA, USA.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

- Professor Sharon Gannot (Bar Ilan University), February and October 2015.
- Dr. Romain Sérizel (Telecom Paris Tech), February 2015.
- Dr. Christine Evers (Imperial College), March 2015.
- Dr. Xavier Alameda-Pineda (University of Trento), November 2015.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events selection

9.1.1.1. Chair of conference program committees

Radu Horaud was program co-chair of ACM ICMI'15, November 2015, Seattle WA, USA.

9.1.2. Journal

9.1.2.1. Member of the editorial boards

Radu Horaud is a member of the following editorial boards:

- advisory board member of the *International Journal of Robotics Research, Sage*,
- associate editor of the *International Journal of Computer Vision, Kluwer*, and
- area editor of *Computer Vision and Image Understanding, Elsevier*.

9.1.3. Invited talks

Radu Horaud gave an invited talk at the IEEE ICCV Worskhop on 3D Reconstruction and Understanding with Video and Sound, Santiago de Chile, Chile, 17 December 2015.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

E-learning: MOOC on *Binaural Hearing for Robots*, May-June 2015, 5 hours over five weeks. Teacher: Radu Horaud.

Tutorial: *Embodied Audition for Robots* at the European Signal Processing Conference (EUSIPCO), Nice, France, 31 August 2015, 3 hours. Teachers: Radu Horaud, Heinrich Loellmann (Friedrich Alexander Universitat Erlangen) and Christine Evers (Imperial College London).

9.2.2. Supervision

- PhD in progress: Israel Dejene Gebru, October 2013, Radu Horaud and Sileye Ba.
- PhD in progress: Dionyssos Kounades-Bastian, October 2013, Radu Horaud and Laurent Girin.
- PhD in progress: Vincent Drouard, October 2014, Radu Horaud and Sileye Ba.
- PhD in progress: Benoit Massé, October 2014, Radu Horaud and Sileye Ba.
- PhD in progress: Stéphane Lathuilière, October 2014, Radu Horaud.
- PhD in progress: Yutong Ban, October 2015, Radu Horaud.

10. Bibliography

Major publications by the team in recent years

[1] X. Alameda-Pineda, R. Horaud. *A Geometric Approach to Sound Source Localization from Time-Delay Estimates*, in "IEEE Transactions on Audio, Speech and Language Processing", June 2014, vol. 22, n^{o} 6, pp. 1082-1095 [*DOI :* 10.1109/TASLP.2014.2317989], https://hal.inria.fr/hal-00975293

[2] N. ANDREFF, B. ESPIAU, R. HORAUD. *Visual Servoing from Lines*, in "International Journal of Robotics Research", 2002, vol. 21, n^o 8, pp. 679–700, http://hal.inria.fr/hal-00520167

[3] M. HANSARD, R. HORAUD, M. AMAT, G. EVANGELIDIS. *Automatic Detection of Calibration Grids in Time-of-Flight Images*, in "Computer Vision and Image Understanding", April 2014, vol. 121, pp. 108-118 [*DOI :* 10.1016/J.CVIU.2014.01.007], https://hal.inria.fr/hal-00936333

[4] M. HANSARD, R. HORAUD. *Cyclopean geometry of binocular vision*, in "Journal of the Optical Society of America A", September 2008, vol. 25, n^o 9, pp. 2357-2369 [*DOI :* 10.1364/JOSAA.25.002357], http://hal.inria.fr/inria-00435548

[5] M. HANSARD, R. HORAUD. *Cyclorotation Models for Eyes and Cameras*, in "IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics", March 2010, vol. 40, n^o 1, pp. 151-161 [*DOI :* 10.1109/TSMCB.2009.2024211], http://hal.inria.fr/inria-00435549

[6] M. HANSARD, R. HORAUD. *A Differential Model of the Complex Cell*, in "Neural Computation", September 2011, vol. 23, n^o 9, pp. 2324-2357 [*DOI :* 10.1162/NECO_A_00163], http://hal.inria.fr/inria-00590266

[7] M. HANSARD, S. LEE, O. CHOI, R. HORAUD. *Time of Flight Cameras: Principles, Methods, and Applications*, Springer Briefs in Computer Science, Springer, October 2012, 95 p. , http://hal.inria.fr/hal-00725654

[8] R. HORAUD, G. CSURKA, D. DEMIRDJIAN. *Stereo Calibration from Rigid Motions*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", December 2000, vol. 22, n^o 12, pp. 1446–1452 [*DOI :* 10.1109/34.895977], http://hal.inria.fr/inria-00590127

[9] R. HORAUD, F. FORBES, M. YGUEL, G. DEWAELE, J. ZHANG. *Rigid and Articulated Point Registration with Expectation Conditional Maximization*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", March 2011, vol. 33, n^o 3, pp. 587-602 [*DOI :* 10.1109/TPAMI.2010.94], http://hal.inria.fr/inria-00590265

[10] R. HORAUD, M. NISKANEN, G. DEWAELE, E. BOYER. *Human Motion Tracking by Registering an Articulated Surface to 3-D Points and Normals*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2009, vol. 31, n^o 1, pp. 158-163 [*DOI :* 10.1109/TPAMI.2008.108], http://hal.inria.fr/inria-00446898

[11] V. KHALIDOV, F. FORBES, R. HORAUD. *Conjugate Mixture Models for Clustering Multimodal Data*, in "Neural Computation", February 2011, vol. 23, n^o 2, pp. 517-557 [*DOI :* 10.1162/NECO_A_00074], http://hal.inria.fr/inria-00590267

[12] D. KNOSSOW, R. RONFARD, R. HORAUD. *Human Motion Tracking with a Kinematic Parameterization of Extremal Contours*, in "International Journal of Computer Vision", September 2008, vol. 79, n^o 3, pp. 247-269 [*DOI :* 10.1007/S11263-007-0116-2], http://hal.inria.fr/inria-00590247

[13] M. SAPIENZA, M. HANSARD, R. HORAUD. *Real-time Visuomotor Update of an Active Binocular Head*, in "Autonomous Robots", January 2013, vol. 34, n^o 1, pp. 33-45 [*DOI :* 10.1007/S10514-012-9311-2], http://hal.inria.fr/hal-00768615

[14] A. ZAHARESCU, E. BOYER, R. HORAUD. *Topology-Adaptive Mesh Deformation for Surface Evolution, Morphing, and Multi-View Reconstruction*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", April 2011, vol. 33, n^o 4, pp. 823-837 [*DOI :* 10.1109/TPAMI.2010.116], http://hal.inria.fr/inria-00590271

[15] A. ZAHARESCU, E. BOYER, R. HORAUD. *Keypoints and Local Descriptors of Scalar Functions on 2D Manifolds*, in "International Journal of Computer Vision", October 2012, vol. 100, n^{o} 1, pp. 78-98 [*DOI :* 10.1007/S11263-012-0528-5], http://hal.inria.fr/hal-00699620

[16] A. ZAHARESCU, R. HORAUD. *Robust Factorization Methods Using A Gaussian/Uniform Mixture Model*, in "International Journal of Computer Vision", March 2009, vol. 81, n^{o} 3, pp. 240-258 [*DOI :* 10.1007/S11263-008-0169-X], http://hal.inria.fr/inria-00446987

Publications of the year

Articles in International Peer-Reviewed Journals

[17] X. ALAMEDA-PINEDA, R. HORAUD. *Vision-Guided Robot Hearing*, in "International Journal of Robotics Research", April 2015, vol. 34, n^{o} 4-5, pp. 437-456 [*DOI :* 10.1177/0278364914548050], https://hal.inria.fr/hal-00990766

[18] F. CUZZOLIN, D. MATEUS, R. HORAUD. *Robust Temporally Coherent Laplacian Protrusion Segmentation of 3D Articulated Bodies*, in "International Journal of Computer Vision", March 2015, vol. 112, n^{o} 1, pp. 43-70 [*DOI :* 10.1007/S11263-014-0754-0], https://hal.archives-ouvertes.fr/hal-01053737

[19] A. DELEFORGE, F. FORBES, S. BA, R. HORAUD. *Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies*, in "IEEE Journal on Selected Topics in Signal Processing", September 2015, vol. 9, n^{o} 6, pp. 1037-1048 [*DOI :* 10.1109/JSTSP.2015.2416677], https://hal.inria.fr/hal-01136465

[20] A. DELEFORGE, F. FORBES, R. HORAUD. *Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds*, in "International Journal of Neural Systems", February 2015, vol. 25, n^{o} 1, 21 p. [*DOI :* 10.1142/S0129065714400036], https://hal.inria.fr/hal-00960796

[21] A. DELEFORGE, F. FORBES, R. HORAUD. *High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables*, in "Statistics and Computing", September 2015, vol. 25, n^{o} 5, pp. 893-911 [*DOI :* 10.1007/S11222-014-9461-5], https://hal.inria.fr/hal-00863468

[22] A. DELEFORGE, R. HORAUD, Y. Y. SCHECHNER, L. GIRIN. *Co-Localization of Audio Sources in Images Using Binaural Features and Locally-Linear Regression*, in "IEEE Transactions on Audio, Speech and Language Processing", April 2015, vol. 23, n^{o} 4, pp. 718-731 [*DOI :* 10.1109/TASLP.2015.2405475], https://hal.inria.fr/hal-01112834

[23] G. EVANGELIDIS, M. HANSARD, R. HORAUD. *Fusion of Range and Stereo Data for High-Resolution Scene-Modeling*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", November 2015, vol. 37, n^{o} 11, pp. 2178 - 2192 [*DOI :* 10.1109/TPAMI.2015.2400465], https://hal.archives-ouvertes.fr/hal-01110031

[24] M. HANSARD, G. EVANGELIDIS, Q. PELORSON, R. HORAUD. *Cross-Calibration of Time-of-flight and Colour Cameras*, in "Computer Vision and Image Understanding", April 2015, vol. 134, pp. 105-115 [*DOI :* 10.1016/J.CVIU.2014.09.001], https://hal.inria.fr/hal-01059891

[25] T. HUEBER, L. GIRIN, X. ALAMEDA-PINEDA, G. BAILLY. *Speaker-Adaptive Acoustic-Articulatory Inversion using Cascaded Gaussian Mixture Regression*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", December 2015, vol. 23, n^{o} 12, pp. 2246-2259 [*DOI :* 10.1109/TASLP.2015.2464702], https://hal.archives-ouvertes.fr/hal-01231197

[26] K. Kulkarni, G. Evangelidis, J. Cech, R. Horaud. *Continuous Action Recognition Based on Sequence Alignment*, in "International Journal of Computer Vision", March 2015, vol. 112, n$^{\text{o}}$ 1, pp. 90-114 [*DOI :* 10.1007/s11263-014-0758-9], https://hal.archives-ouvertes.fr/hal-01058732

International Conferences with Proceedings

[27] F. Badeig, Q. Pelorson, S. Arias, V. Drouard, I. D. Gebru, X. Li, G. Evangelidis, R. Horaud. *A Distributed Architecture for Interacting with NAO*, in "International Conference on Multimodal Interaction", Seattle, WA, United States, ACM, November 2015 [*DOI :* 10.1145/2818346.2823303], https://hal.inria.fr/hal-01201716

[28] *Best Paper*
V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, R. Horaud. *Head Pose Estimation via Probabilistic High-Dimensional Regression*, in "IEEE International Conference on Image Processing", Quebec City, Canada, Proceedings of the IEEE International Conference on Image Processing, September 2015, https://hal.inria.fr/hal-01163663.

[29] I. D. Gebru, S. Ba, G. Evangelidis, R. Horaud. *Audio-Visual Speech-Turn Detection and Tracking*, in "The 12-th International Conference on Latent Variable Analysis and Signal Separation", Liberec, Czech Republic, August 2015, pp. 143-151 [*DOI :* 10.1007/978-3-319-22482-4_17], https://hal.inria.fr/hal-01163659

[30] I. D. Gebru, S. Ba, G. Evangelidis, R. Horaud. *Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model*, in "ICCV Workshop on 3D Reconstruction and Understanding with Video and Sound", Santiago, Chile, December 2015, https://hal.archives-ouvertes.fr/hal-01220956

[31] *Best Paper*
D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, R. Horaud. *A Variational EM Algorithm for the Separation of Moving Sound Sources*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, NY, United States, IEEE Signal Processing Society, October 2015, https://hal.inria.fr/hal-01169764.

[32] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, R. Horaud. *An Inverse-Gamma Source Variance Prior with Factorized Parameterization for Audio Source Separation*, in "IEEE International Conference on Acoustic, Speech, and Signal Processing", Shangai, China, IEEE Signal Processing Society, March 2016, https://hal.inria.fr/hal-01253169

[33] X. Li, L. Girin, S. Gannot, R. Horaud. *Non-Stationary Noise Power Spectral Density Estimation Based on Regional Statistics*, in "IEEE International Conference on Audio, Speech and Signal Processing", Shangai, China, IEEE Signal Processing Society, March 2016, https://hal.inria.fr/hal-01250892

[34] X. Li, L. Girin, R. Horaud, S. Gannot. *Estimation of Relative Transfer Function in the Presence of Stationary Noise Based on Segmental Power Spectral Density Matrix Subtraction*, in "IEEE International Conference on Acoustics, Speech and Signal Processing", Brisbane, Australia, IEEE Signal Processing Society, April 2015, https://hal.archives-ouvertes.fr/hal-01119186

[35] X. Li, R. Horaud, L. Girin, S. Gannot. *Local Relative Transfer Function for Sound Source Localization*, in "The European Signal Processing Conference", Nice, France, August 2015, vol. The European Signal Processing Conference, https://hal.inria.fr/hal-01163675

References in notes

[36] S. Ba, X. Alameda-Pineda, A. Xompero, R. Horaud. *An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes*, in "CoRR", 2015, vol. abs/1509.01520, http://arxiv.org/abs/1509.01520

[37] A. Deleforge, F. Forbes, R. Horaud. *Variational EM for Binaural Sound-Source Separation and Localization*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing", Vancouver, Canada, IEEE, 2013, pp. 76-80 [*DOI :* 10.1109/ICASSP.2013.6637612], http://hal.inria.fr/hal-00823453

[38] I. D. Gebru, X. Alameda-Pineda, F. Forbes, R. Horaud. *EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis*, in "CoRR", 2015, vol. abs/1509.01509, http://arxiv.org/abs/1509.01509

[39] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, R. Horaud. *A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures*, in "CoRR", 2015, vol. abs/1510.04595, http://arxiv.org/abs/1510.04595

[40] X. Li, L. Girin, R. Horaud, S. Gannot. *Binaural Sound Source Localization based on Direct-Path Relative Transfer Function*, in "CoRR", 2015, vol. abs/1509.03205, http://arxiv.org/abs/1509.03205