Activity Report 2017

Project-Team PERCEPTION

Interpretation and Modelling of Images and Videos

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:

A3.4. - Machine learning and statistics
A5.1. - Human-Computer Interaction
A5.3. - Image processing and analysis
A5.4. - Computer vision
A5.7. - Audio modeling and processing
A5.10.2. - Perception
A5.10.5. - Robot interaction (with the environment, humans, other robots)
A9.2. - Machine learning
A9.5. - Robotics

Other Research Topics and Application Domains:

B5.6. - Robotic systems

1. Personnel

Research Scientists

Radu Patrice Horaud [Team leader, Inria, Senior Researcher, HDR]
Xavier Alameda-Pineda [Inria, Researcher]
Xiaofei Li [Inria, Starting Research Position]
Pablo Mesejo Santiago [Inria, Starting Research Position]

PhD Students

Yutong Ban [Inria]
Israel Dejene Gebru [Inria, until Jan 2017]
Guillaume Delorme [Inria, from Sep 2017]
Vincent Drouard [Inria]
Sylvain Guy [Univ Grenoble Alpes, from Oct 2017]
Dionyssos Kounades [Inria, until Mar 2017]
Stephane Lathuiliere [Inria]
Benoit Masse [Inria]

Technical staff

Dionyssos Kounades [Inria, from Apr 2017 until Sep 2017]
Bastien Mourgue [Inria]
Guillaume Sarrazin [Inria]

Interns

Guillaume Delorme [Inria, from Apr 2017 until Aug 2017]
Divya Grover [Inria, from Mar 2017 until Aug 2017]
Sylvain Guy [Inria, from Feb 2017 until Jul 2017]
Duc Anh Luu [Inria, from Jul 2017 until Sep 2017]

Administrative Assistant

Nathalie Gillot [Inria]

Visiting Scientists

Sharon Gannot [Bar Ilan University, from Feb 2017 until Oct 2017]
Oscar David Gomez Lopez [University of Granada, from Dec 2017]

External Collaborator

Laurent Girin [Institut Polytechnique de Grenoble, HDR]

2. Overall Objectives

2.1. Overall Objectives

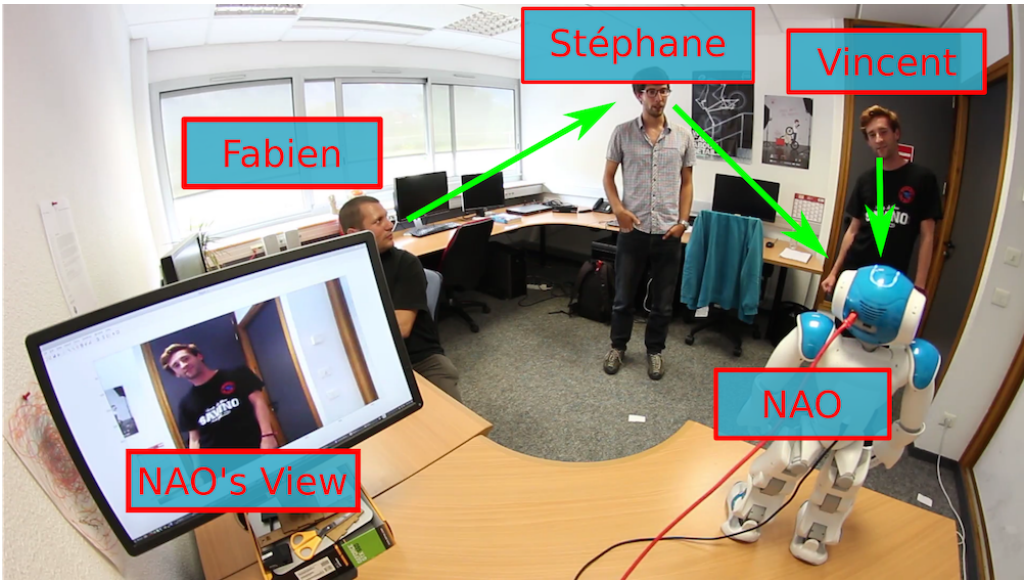

Figure 1. This figure illustrates the audio-visual multi-party human-robot interaction paradigm that the PERCEPTION team has developed in the recent past [20], [2], [10]. There are inter-person as well as person-robot interactions that must be properly detected and analyzed over time. This includes multiple-person tracking [4], person detection and head-pose estimation [30], sound-source separation and localization [6], [1], [22], [23],[35], and speaker diarization [33]. These developments have been supported by the European Union via the FP7 STREP project "Embodied Audition for Robots" (EARS) and the ERC advanced grant "Vision and Hearing in Action" (VHIA).

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.
Video: https://team.inria.fr/perception/demos/nao-video/

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 [20], [2]. 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 [7]. 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 [6]. The European projects HUMAVIPS (2010-2013) coordinated by R. Horaud and EARS (2014-2017), 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 [14], [24]. 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 [15]. 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 [9].

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 [2]. 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 [6] and audio-visual learning [8].

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 [27]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [25]. 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 [26]. 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 [16], [12],[11]. 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 [9].

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 [21]. 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 [19], [18]. 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 [5]. 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

- In collaboration with several partners, PERCEPTION completed the three year EU STREP project EARS (2014-2017). PERCEPTION contributed to audio-source localization using microphone arrays and to the disambiguation of audio information using vision, in particular to discriminate between speaking and silent persons.
  Website: https://robot-ears.eu/
- PERCEPTION started and completed a one year collaboration (December 2016 – November 2017) with **Samsung Electronics Digital Media and Communications R&D Center**, Seoul, Korea. The topic of this collaboration, fully funded by Samsung, was *multi-modal methodologies for human-robot interaction* (a central topic of the team) and is part of a strategic partnership between Inria and Samsung Electronics. A follow-up of this collaboration is under preparation and it is planned to start soon (February 2018).
- As an ERC Advanced Grant holder, Radu Horaud was awarded a Proof of Concept grant for his project Vision and Hearing in Action Laboratory (VHIALab). The project will develop software packages enabling companion robots to robustly interact with multiple users.
  Website: https://team.inria.fr/perception/projects/poc-vhialab/

4.1.1. Awards

- Israel Dejene Gebru (PhD student) and his co-authors, Christine Evers, Patrick Naylor (both from Imperial College London) and Radu Horaud, received the best paper award at the IEEE Fifth Joint Workshop on Hands-free Speech Communication and Microphone Arrays, San Francisco, USA, 1-3 March 2017, for their paper Audio-visual Tracking by Density Approximation in a Sequential Bayesian Filtering Framework.
- Yutong Ban (PhD student) and his co-authors, Xavier Alameda-Pineda, Fabien Badeig, and Radu Horaud, were among the five finalists of the "Novel Technology Paper Award for Amusement Culture" at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada, September 2017, for their paper Tracking a Varying Number of People with a Visually-Controlled Robotic Head.

BEST PAPERS AWARDS:

[41]
I. GEBRU, C. EVERS, P. NAYLOR, R. HORAUD. *Audio-visual Tracking by Density Approximation in a Sequential Bayesian Filtering Framework*, in "IEEE Workshop on Hands-free Speech Communication and Microphone Arrays", San Francisco, CA, United States, IEEE Signal Processing Society, March 2017, Best Paper Award [*DOI :* 10.1109/HSCMA.2017.7895564], https://hal.inria.fr/hal-01452167

[38]
Y. BAN, X. ALAMEDA-PINEDA, F. BADEIG, S. BA, R. HORAUD. *Tracking a Varying Number of People with a Visually-Controlled Robotic Head*, in "IEEE/RSJ International Conference on Intelligent Robots and Systems", Vancouver, Canada, September 2017, https://hal.inria.fr/hal-01542987

5. New Software and Platforms

5.1. ECMPR

Expectation Conditional Maximization for the Joint Registration of Multiple Point Sets
FUNCTIONAL DESCRIPTION: Rigid registration of two or several point sets based on probabilistic matching between point pairs and a Gaussian mixture model

- Participants: Florence Forbes, Manuel Yguel and Radu Horaud
- Contact: Patrice Horaud
- URL: https://team.inria.fr/perception/research/jrmpc/

5.2. Mixcam

Reconstruction using a mixed camera system
KEYWORDS: Computer vision - 3D reconstruction
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 2011-2014 contains the calibration of TOF cameras, alignment between TOF and color cameras, TOF-stereo fusion, and image-based rendering. 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: Clément Ménier, Georgios Evangelidis, Michel Amat, Miles Hansard, Patrice Horaud, Pierre Arquier, Quentin Pelorson, Radu Horaud, Richard Broadbridge and Soraya Arias
- Contact: Patrice Horaud
- URL: https://team.inria.fr/perception/mixcam-project/

5.3. NaoLab

Distributed middleware architecture for interacting with NAO

FUNCTIONAL DESCRIPTION: This software provides a set of librairies and tools to simply the control of NAO robot from a remote machine. The main challenge is to make easy prototuping applications for NAO ising C++ and Matlab programming environments. Thus NaoLab provides a prototyping-friendly interface to retrieve sensor date (video and sound streams, odometric data...) and to control the robot actuators (head, arms, legs...) from a remote machine.This interface is available on Naoqi SDK, developed by Aldebarab company, Naoqi SDK is needed as it provides the tools to acess the embedded NAO services (low-level motor command, sensor data access...)

- Authors: Fabien Badeig, Quentin Pelorson and Patrice Horaud
- Contact: Patrice Horaud
- URL: https://team.inria.fr/perception/research/naolab/

5.4. Stereo matching and recognition library

KEYWORD: Computer vision
FUNCTIONAL DESCRIPTION: Library providing stereo matching components to rectify stereo images, to retrieve faces from left and right images, to track faces and method to recognise simple gestures

- Participants: Jan Cech, Jordi Sanchez-Riera, Radu Horaud and Soraya Arias
- Contact: Soraya Arias
- URL: https://code.humavips.eu/projects/stereomatch

5.5. Platforms

5.5.1. Audio-Visual Head Popeye+

In 2016 our audio-visual platform was upgraded from Popeye to Popeye+. Popeye+ has two high-definition cameras with a wide field of view. We also upgraded the software libraries that perform synchronized acquisition of audio signals and color images. Popeye+ has been used for several datasets.
Websites:
https://team.inria.fr/perception/projects/popeye/
https://team.inria.fr/perception/projects/popeye-plus/
https://team.inria.fr/perception/avtrack1/
https://team.inria.fr/perception/avdiar/

5.5.2. NAO Robots

The PERCEPTION team selected the companion robot NAO for experimenting and demonstrating various audio-visual skills as well as for developing the concept of social robotics that is able to recognize human presence, to understand human gestures and voice, and to communicate by synthesizing appropriate behavior. The main challenge of our team is to enable human-robot interaction in the real world.

The humanoid robot NAO is manufactured by SoftBank Robotics Europe. Standing, the robot is roughly 60 cm tall, and 35cm when it is sitting. Approximately 30 cm large, NAO includes two CPUs. The first one, placed in the torso, together with the batteries, controls the motors and hence provides kinematic motions with 26 degrees of freedom. The other CPU is placed in the head and is in charge of managing the proprioceptive sensing, the communications, and the audio-visual sensors (two cameras and four microphones, in our case). NAO's on-board computing resources can be accessed either via wired or wireless communication protocols.

NAO's commercially available head is equipped with two cameras that are arranged along a vertical axis: these cameras are neither synchronized nor a significant common field of view. Hence, they cannot be used in combination with stereo vision. Within the EU project HUMAVIPS, Aldebaran Robotics developed a binocular camera system that is arranged horizontally. It is therefore possible to implement stereo vision algorithms on NAO. In particular, one can take advantage of both the robot's cameras and microphones. The cameras deliver VGA sequences of image pairs at 12 FPS, while the sound card delivers the audio signals arriving from all four microphones and sampled at 48 kHz. Subsequently, Aldebaran developed a second binocular camera system to go into the head of NAO v5.

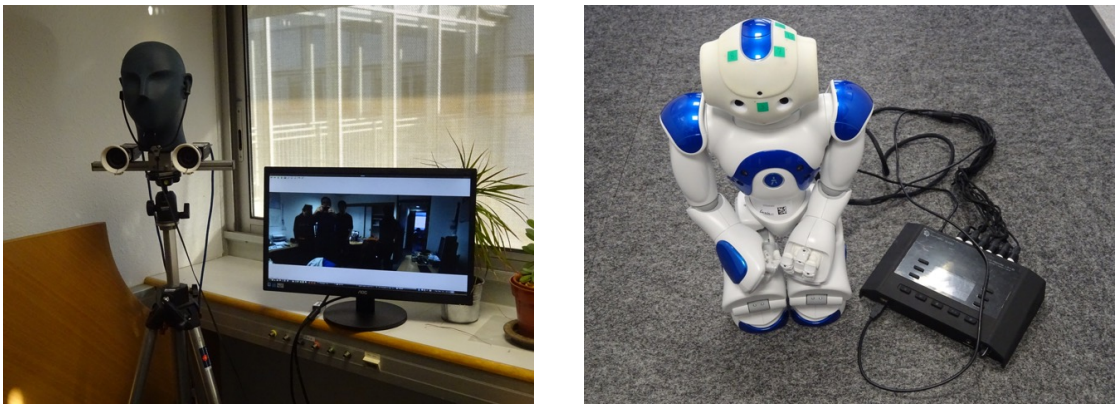

Figure 2. The Popeye+ audio-visual platform (left) delivers high-quality, high-resolution and wide-angle images at 30FPS. The NAO prototype used by PERCEPTION in the EARS STREP project has a twelve-channel spherical microphone array synchronized with a stereo camera pair.

In order to manage the information flow gathered by all these sensors, we implemented our software on top of the Robotics Services Bus (RSB). RSB is a platform-independent event-driven middleware specifically designed for the needs of distributed robotic applications. Several RSB tools are available, including real-time software execution, as well as tools to record the event/data flow and to replay it later, so that application development can be done off-line. RSB events are automatically equipped with several time stamps for introspection and synchronization purposes. RSB was chosen because it allows our software to be run on a remote PC platform, neither with performance nor deployment restrictions imposed by the robot's CPUs. Moreover, the software packages can be easily reused for other robots.

Recently (2015-2016) the PERCEPTION team started the development of NAOLab, a middleware for hosting robotic applications in C, C++, Python and Matlab, using the computing power available with NAO, augmented with a networked PC. More recently, NAOLab was renamed RMP (Robotics Middleware for Perception).

Websites:
https://team.inria.fr/perception/nao/
https://team.inria.fr/perception/research/naolab/

6. New Results

6.1. Audio-Source Localization

In previous years we have developed several *supervised* sound-source localization algorithms. The general principle of these algorithms was based on the learning of a mapping (regression) between binaural feature vectors and source locations [6], [8]. While fixed-length wide-spectrum sounds (white noise) are used for training to reliably estimate the model parameters, we showed 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 demonstrated that the method could 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. This year we released 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. During the period 2015-2016 we extended this method to an arbitrary number of microphones based on the *relative transfer function – RTF* (between any channel and a reference channel). In the period 2016-2017 we extended this work and developed a novel transfer function that contains the direct path between the source and the microphone array, namely the *direct-path relative transfer function* [23], [35].

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

6.2. Audio-Source Separation

We addressed the problem of separating audio sources from both static and time-varying convolutive mixtures. We proposed an unsupervised probabilistic framework based on the local complex-Gaussian model combined with non-negative matrix factorization [22]. The time-varying mixing filters are modeled by a continuous temporal stochastic process. This model extended the case of static filters which corresponds to static audio sources. While static filters can be learnt in advance, e.g. [6], time-varying filters cannot and therefore the problem is more complex. We developed 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. In 2017 we extended this method to incorporate the concept of diarization. Indeed, audio sources such as speaking persons do not emit continuously, bet merely take "turns". We formally modeled speech turn-taking within a combined separation and diarization formulation [45], [44]. We also started to investigate the use of the convolutive transfer function for audio-source separation [49], [48], [54].
Websites:
https://team.inria.fr/perception/research/vemove/
https://team.inria.fr/perception/research/nmfig/
https://team.inria.fr/perception/research/dnd/

6.3. Speech Dereverberation and Noise Reduction

We address the problems of blind multichannel identification and equalization for *joint speech dereverberation and noise reduction*. The standard time-domain cross-relation methods are hardly applicable for blind room impulse response identification due to the near-common zeros of the long impulse responses. We extend the cross-relation formulation to the short-time Fourier transform (STFT) domain, in which the time-domain impulse response is approximately represented by the convolutive transfer function (CTF) with much less coefficients. For the oversampled STFT, CTFs suffer from the common zeros caused by the non-flat-top STFT window. To overcome this, we propose to identify CTFs using the STFT framework with oversampled signals and critically sampled CTFs, which is a good trade-off between the frequency aliasing of the signals and the common zeros problem of CTFs. The phases of the identified CTFs are inaccurate due to the frequency aliasing of the CTFs, and thus only their magnitudes are used. This leads to a non-negative multichannel equalization method based on a non-negative convolution model between the STFT magnitude of the source signal and the CTF magnitude. To recover the STFT magnitude of the source signal and to reduce the additive noise, the ℓ_2-norm fitting error between the STFT magnitude of the microphone signals and the non-negative convolution is constrained to be less than a noise power related tolerance. Meanwhile, the ℓ_1-norm of the STFT magnitude of the source signal is minimized to impose the sparsity [53].
Website: https://team.inria.fr/perception/research/ctf-dereverberation/

6.4. Acoustic-Articulatory Mapping

In this series of studies, we tackle the problem of adapting an acoustic-articulatory inversion model of a reference speaker to the voice of another source speaker. We exploited the framework of Gaussian mixture regressors (GMR) with missing data. To address speaker adaptation, we previously proposed a general framework called Cascaded-GMR (C-GMR) which decomposes the adaptation process into two consecutive steps: spectral conversion between source and reference speaker and acoustic-articulatory inversion of converted spectral trajectories. In particular, we proposed the Integrated C-GMR technique (IC-GMR) in which both steps are tied together in the same probabilistic model. In [34], [43], we extend the C-GMR framework with another model called Joint-GMR (J-GMR). Contrary to the IC-GMR, this model aims at exploiting all potential acoustic-articulatory relationships, including those between the source speaker's acoustics and the reference speaker's articulation. We present the full derivation of the exact Expectation-Maximization (EM) training algorithm for the J-GMR. It exploits the missing data methodology of machine learning to deal with limited adaptation data. We provide an extensive evaluation of the J-GMR on both synthetic acoustic-articulatory data and on the multi-speaker MOCHA EMA database. We compare the J-GMR performance to other models of the C-GMR framework, notably the IC-GMR, and discuss their respective merits. We also exploited the IC-GMR framework with visual data to provide visual biofeedback [32]. Visual biofeedback is the process of gaining awareness of physiological functions through the display of visual information. As speech is concerned, visual biofeedback usually consists in showing a speaker his/her own articulatory movements, which has proven useful in applications such as speech therapy or second language learning. We automatically animate an articulatory tongue model from ultrasound images. We benchmarked several GMR-based techniques on a multispeaker database. The IC-GMR approach is able (i) to maintain good mapping performance while minimizing the amount of adaptation data (and thus limiting the duration of the enrollment session), and (ii) to generalize to articulatory configurations not seen during enrollment better than the plain GMR approach. As a result, the GMR appears to be a good mapping technique for non-linear regression tasks, and in particular for those requiring adaptation (either using J-GMR or IC-GMR).

6.5. Visual Tracking of Multiple Persons

Object tracking is an ubiquitous problem in computer vision with many applications in human-machine and human-robot interaction, augmented reality, driving assistance, surveillance, etc. Although thoroughly investigated, tracking multiple persons remains a challenging and an open problem. In this work, an online variational Bayesian model for multiple-person tracking is proposed. This yields a variational expectation-maximization (VEM) algorithm. The computational efficiency of the proposed method is made possible thanks to closed-form expressions for both the posterior distributions of the latent variables and for the estimation of the model parameters. A stochastic process that handles person birth and person death enables the tracker to handle a varying number of persons over long periods of time [4]. The method was combined with visual servoing and implemented on our robot platform (Fig. 3) [38].
Websites:
https://team.inria.fr/perception/research/ovbt/
https://team.inria.fr/perception/research/mot-servoing/

6.6. Audio-Visual Speaker Tracking and Diarization

We are particularly interested in modeling the interaction between an intelligent device and a group of people. For that purpose we develop audio-visual person tracking methods [33], [41], [52], [39]. As the observed persons are supposed to carry out a conversation, we also include speaker diarization into our tracking methodology. 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 [10], 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

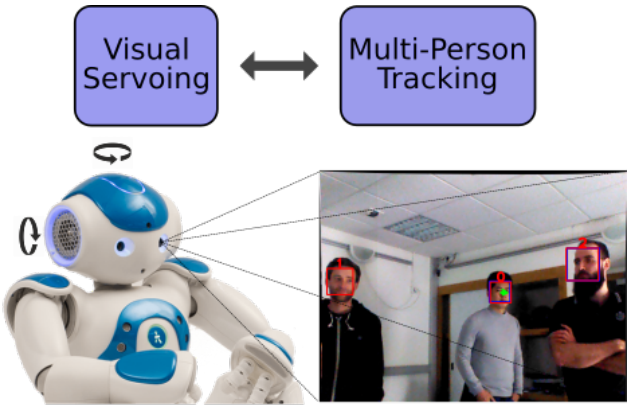

Figure 3. The multi-person tracking method is combined with a visual servoing module. The latter estimates the optimal robot commands and the expected impact of the tracked person locations. The multi-person tracking module refines the locations of the persons with the new observations and the information provided by the visual servoing.

of the proposed method are tested on challenging datasets that are available from recent contributions which are used as baselines for comparison [33].
Websites:
https://team.inria.fr/perception/research/wdgmm/
https://team.inria.fr/perception/research/speakerloc/
https://team.inria.fr/perception/research/speechturndet/
https://team.inria.fr/perception/research/avdiarization/

6.7. Head Pose Estimation and Tracking

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 [7]. 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 [30]. This work is part of the PhD of Vincent Drouard [28] that received the best student paper award (second place) at the IEEE ICIP'15.

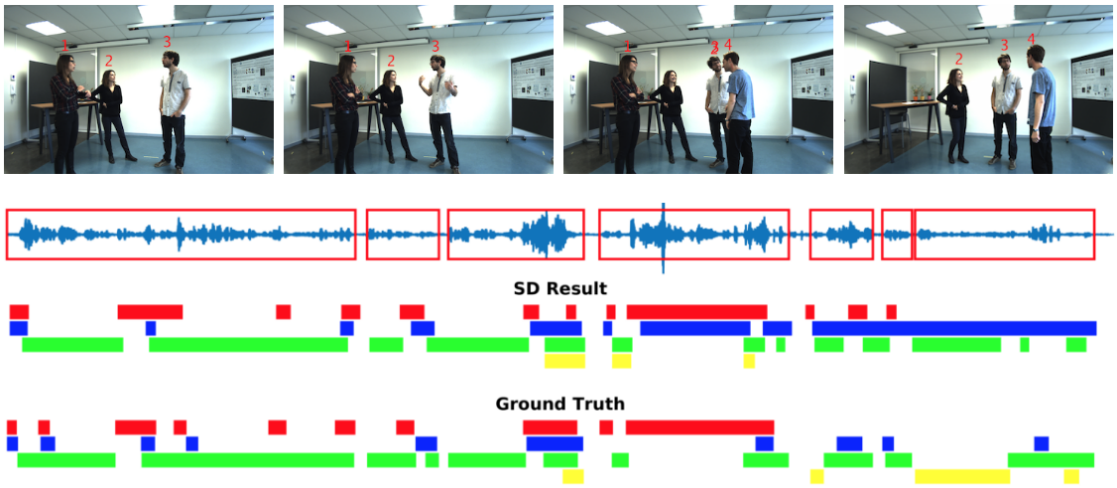

Figure 4. This figure illustrates the audiovisual tracking and diarization method that we have recently developed. First row: A number is associated with each tracked person. Second row: diarization result. Third row: the ground truth diarization. Fourth row: acoustic signal recorded by one of the two microphones.

In 2017 we extended this work and we proposed a head-pose tracker based on a switching Kalman filter (SKF) formalism. The SKF governs the temporal predictive distribution of the pose parameters (modeled as continuous latent variables) conditioned by the discrete variables associated with the mixture of linear inverse-regression formulation of [7]. We formally derived the equations of the proposed switching linear regression model, we proposed an approximation that is both identifiable and computationally tractable, we designed an EM procedure to estimate the SKF parameters in closed-form, and we carried out experiments and comparisons with other methods using recently released datasets [40].
Websites:
https://team.inria.fr/perception/research/head-pose/
https://team.inria.fr/perception/research/head-pose-tracking/

6.8. Tracking Eye Gaze and of Visual Focus of Attention

The visual focus of attention (VFOA) has been recognized as a prominent conversational cue. We are interested in estimating and tracking the VFOAs associated with multi-party social interactions. We note that in this type of situations the participants either look at each other or at an object of interest; therefore their eyes are not always visible. Consequently both gaze and VFOA estimation cannot be based on eye detection and tracking. We propose a method that exploits the correlation between eye gaze and head movements. Both VFOA and gaze are modeled as latent variables in a Bayesian switching state-space model (also named switching Kalman filter). The proposed formulation leads to a tractable learning method and to an efficient online inference procedure that simultaneously tracks gaze and visual focus. The method is tested and benchmarked using two publicly available datasets, Vernissage and LAEO, that contain typical multi-party human-robot and human-human interactions [36].
Website:
https://team.inria.fr/perception/research/eye-gaze/.

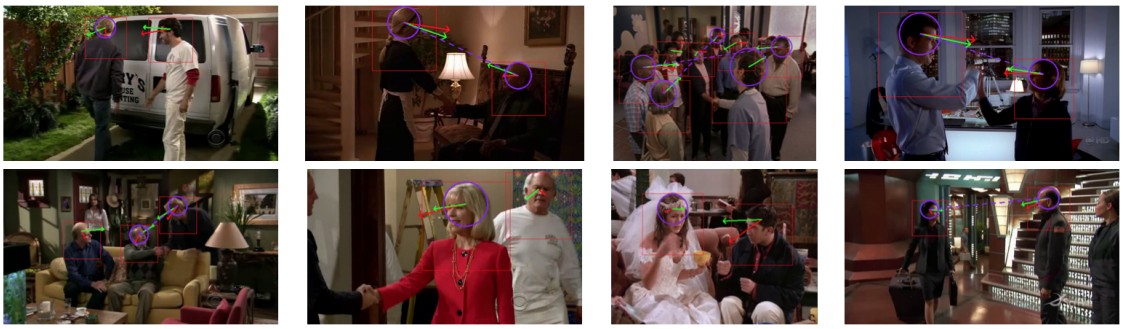

Figure 5. This figure shows some results obtained with the LAEO dataset. The top row shows results obtained with coarse head orientation and the bottom row shows results obtained with fine head orientation. Head orientations are shown with red arrows. The algorithm infers gaze directions (green arrows) and VFOAs (blue circles). People looking at each others are shown with a dashed blue line.

6.9. Attention-Gated Conditional Random Fields

Recent works have shown that exploiting multi-scale representations deeply learned via convolutional neural networks (CNN) is of tremendous importance for accurate contour detection. We present [51] a novel approach for predicting contours which advances the state of the art in two fundamental aspects, i.e. multi-scale feature generation and fusion. Different from previous works directly considering multi-scale feature maps obtained from the inner layers of a primary CNN architecture, we introduce a hierarchical deep model which produces more rich and complementary representations. Furthermore, to refine and robustly fuse the representations learned at different scales, the novel Attention-Gated Conditional Random Fields (AG-CRFs) are proposed. The experiments ran on two publicly available datasets (BSDS500 and NYUDv2) demonstrate the effectiveness of the latent AG-CRF model and of the overall hierarchical framework.

6.10. Pooling Local Virality

In our overly-connected world, the automatic recognition of virality - the quality of an image or video to be rapidly and widely spread in social networks - is of crucial importance, and has recently awaken the interest of the computer vision community. Concurrently, recent progress in deep learning architectures showed that global pooling strategies allow the extraction of activation maps, which highlight the parts of the image most likely to contain instances of a certain class. We extended this concept by introducing a pooling layer that learns the size of the support area to be averaged: the learned top-N average (LENA) pooling [37]. We hypothesize that the latent concepts (feature maps) describing virality may require such a rich pooling strategy. We assess the effectiveness of the LENA layer by appending it on top of a convolutional siamese architecture and evaluate its performance on the task of predicting and localizing virality. We report experiments on two publicly available datasets annotated for virality and show that our method outperforms state-of-the-art approaches.

6.11. Registration of Multiple Point Sets

We have also addressed the rigid registration problem of multiple 3D point sets. While the vast majority of state-of-the-art techniques build on pairwise registration, we proposed a generative model that explains jointly registered multiple sets: back-transformed points are considered realizations of a single Gaussian mixture model (GMM) whose means play the role of the (unknown) scene points. Under this assumption, the joint registration problem is cast into a probabilistic clustering framework. We formally derive an expectation-maximization procedure that robustly estimates both the GMM parameters and the rigid transformations

that map each individual cloud onto an under-construction reference set, that is, the GMM means. GMM variances carry rich information as well, thus leading to a noise- and outlier-free scene model as a by-product. A second version of the algorithm is also proposed whereby newly captured sets can be registered online. A thorough discussion and validation on challenging data-sets against several state-of-the-art methods confirm the potential of the proposed model for jointly registering real depth data [31].
Website:
https://team.inria.fr/perception/research/jrmpc/

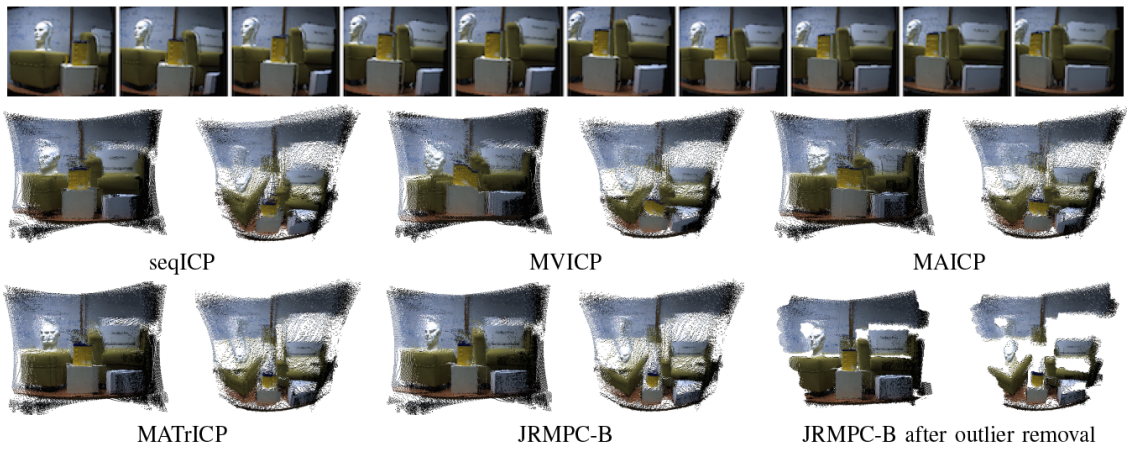

Figure 6. Integrated point clouds from the joint registration of 10 TOF images that record a static scene (EXBI data-set). Top: color images that roughly show the scene content of each range image (occlusions due to cameras baseline may cause texture artefacts). Bottom: front-view and top-view of integrated sets after joint registration. The results obtained with the proposed method (JRMPC-B) are compared with several other methods.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

From December 2016 to November 2017 the PERCEPTION team had a collaborative project with Samsung's Digital Media and Communication R&D Center. The collaboration was fully funded by Samsung Electronics. The topic of this collaboration was *multi-modal approach to human-robot interaction*.

8. Partnerships and Cooperations

8.1. European Initiatives

8.1.1. FP7 & H2020 Projects

8.1.1.1. VHIA

Title: Vision and Hearing in Action

EU framework: FP7

Type: ERC Advanced Grant

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.

Website: https://team.inria.fr/perception/projects/erc-vhia/

8.2. International Initiatives

8.2.1. Inria International Partners

8.2.1.1. Informal International Partners

- Bar Ilan University, Israel (prof. Sharon Gannot and his team)
- University of Trento, Italy (prof. Nicu Sebe and prof. Elisa Ricci)
- Dr. Rafael Munoz-Salinas and prof. Manuel Marin-Jimenez, University of Cordoba, Spain,
- Dr. Christine Evers and prof. Patrick Naylor, Imperial College of Science and Medecine, UK.
- Dr. Miriam Redi, Wikimedia Foundation, UK.
- Prof. Shih-Fu Chang, Columbia University, USA.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

- Prof. Sharon Gannot (Bar Ilan University)
- Oscar David Gomez Lopez (University of Granada)

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Xavier Alameda-Pineda was area chair of IEEE/CVF *International Conference on Computer Vision* 2017.

9.1.1.2. Member of the Organizing Committees

Xavier Alameda-Pineda co-organized a special session at ACM *International Conference on Multimedia Retrieval* 2017, and a workshop at ACM *International Conference on Multimedia* 2017.

9.1.2. Scientific Events Selection

9.1.2.1. Reviewer

In 2017, Xavier Alameda-Pineda reviewed for IEEE Conference on *Computer Vision and Pattern Recognition* 2017, *Advances on Neural Information Processing Systems* 2017, IEEE International *Conference on Audio Speech and Signal Processing* 2018 and *International Conference on Learning Representations*.
Xavier was awarded with the CVPR 2017 **Outstanding Reviewer Award**.

9.1.3. Journal

9.1.3.1. 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*.

9.1.3.2. Reviewer - Reviewing Activities

Xavier Alameda-Pineda regularly acts as reviewer for IEEE *Transactions on Pattern Analysis and Machine Intelligence*, IEEE *Transactions on Audio, Speech, and Language Processing*, IEEE *Transactions on Multimedia*, IEEE *Transactions on Image Procesing*, IEEE *Transactions on Affective Computing*, *International Journal on Computer Vision* and *Computer Vision and Image Understanding*.

9.1.4. Invited Talks

- Xavier Alameda-Pineda gave an invited talk at GIPSA-Lab on Multi-speaker tracking with auditory data.
- Radu Horaud gave invited talks at two IEEE ICCV Workshops:
  https://www.msf-workshop.com/
  https://mvr3d.github.io/.

9.2. Teaching - Supervision - Juries

9.2.1. Supervision

PhD defended February 2017: Dionyssos Kounades-Bastian, October 2013, Radu Horaud, Laurent Girin, and Xavier Alameda-Pineda.

PhD defended in December 2017: 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 and Laurent Girin

PhD in progress: Guillaume Delorme, September 2017, Radu Horaud and Xavier Alameda-Pineda

PhD in progress: Sylvain Guy, October 2017, Radu Horaud and Laurent Girin

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] 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

[3] 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

[4] S. BA, X. ALAMEDA-PINEDA, A. XOMPERO, R. HORAUD. *An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes*, in "Computer Vision and Image Understanding", December 2016, vol. 153, pp. 64–76 [*DOI :* 10.1016/J.CVIU.2016.07.006], https://hal.inria.fr/hal-01349763

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] I. D. GEBRU, X. ALAMEDA-PINEDA, F. FORBES, R. HORAUD. *EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", December 2016, vol. 38, n^{o} 12, pp. 2402 - 2415 [*DOI :* 10.1109/TPAMI.2016.2522425], https://hal.inria.fr/hal-01261374

[11] 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

[12] 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

[13] 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

[14] 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° 1, pp. 151-161 [*DOI :* 10.1109/TSMCB.2009.2024211], http://hal.inria.fr/inria-00435549

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

[16] 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

[17] 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° 12, pp. 1446–1452 [*DOI :* 10.1109/34.895977], http://hal.inria.fr/inria-00590127

[18] 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° 3, pp. 587-602 [*DOI :* 10.1109/TPAMI.2010.94], http://hal.inria.fr/inria-00590265

[19] 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° 1, pp. 158-163 [*DOI :* 10.1109/TPAMI.2008.108], http://hal.inria.fr/inria-00446898

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

[21] 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° 3, pp. 247-269 [*DOI :* 10.1007/s11263-007-0116-2], http://hal.inria.fr/inria-00590247

[22] 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 "IEEE/ACM Transactions on Audio, Speech and Language Processing", August 2016, vol. 24, n° 8, pp. 1408-1423 [*DOI :* 10.1109/TASLP.2016.2554286], https://hal.inria.fr/hal-01301762

[23] X. LI, L. GIRIN, R. HORAUD, S. GANNOT. *Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", November 2016, vol. 24, n° 11, pp. 2171 - 2186 [*DOI :* 10.1109/TASLP.2016.2598319], https://hal.inria.fr/hal-01349691

[24] M. SAPIENZA, M. HANSARD, R. HORAUD. *Real-time Visuomotor Update of an Active Binocular Head*, in "Autonomous Robots", January 2013, vol. 34, n° 1, pp. 33-45 [*DOI :* 10.1007/s10514-012-9311-2], http://hal.inria.fr/hal-00768615

[25] 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° 4, pp. 823-837 [*DOI :* 10.1109/TPAMI.2010.116], http://hal.inria.fr/inria-00590271

[26] 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° 1, pp. 78-98 [*DOI :* 10.1007/s11263-012-0528-5], http://hal.inria.fr/hal-00699620

[27] A. ZAHARESCU, R. HORAUD. *Robust Factorization Methods Using A Gaussian/Uniform Mixture Model*, in "International Journal of Computer Vision", March 2009, vol. 81, n° 3, pp. 240-258 [*DOI :* 10.1007/s11263-008-0169-x], http://hal.inria.fr/inria-00446987

Publications of the year

Doctoral Dissertations and Habilitation Theses

[28] V. DROUARD. *From Images and Sounds to Face Localization and Tracking*, Université Grenoble Alpes, December 2017, https://hal.inria.fr/tel-01667740

[29] D. KOUNADES-BASTIAN. *Some Contributions to Audio Source Separation and Diarisation of Multichannel Convolutive Mixtures*, Université Grenoble - Alpes, February 2017, https://hal.inria.fr/tel-01543101

Articles in International Peer-Reviewed Journals

[30] V. DROUARD, R. HORAUD, A. DELEFORGE, S. BA, G. EVANGELIDIS. *Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions*, in "IEEE Transactions on Image Processing", March 2017, vol. 26, n° 3, pp. 1428 - 1440, https://arxiv.org/abs/1603.09732 [*DOI :* 10.1109/TIP.2017.2654165], https://hal.inria.fr/hal-01413406

[31] G. EVANGELIDIS, R. HORAUD. *Joint Alignment of Multiple Point Sets with Batch and Incremental Expectation-Maximization*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", June 2017, vol. XX, https://arxiv.org/abs/1609.01466 - 14 pages, 12 figures, 5 tables [*DOI :* 10.1109/TPAMI.2017.2717829], https://hal.inria.fr/hal-01413414

[32] D. FABRE, T. HUEBER, L. GIRIN, X. ALAMEDA-PINEDA, P. BADIN. *Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract*, in "Speech Communication", October 2017, vol. 93, pp. 63 - 75 [*DOI :* 10.1016/j.specom.2017.08.002], https://hal.archives-ouvertes.fr/hal-01578315

[33] I. GEBRU, S. BA, X. LI, R. HORAUD. *Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2017, vol. 39, https://arxiv.org/abs/1603.09725 - 14 pages [*DOI :* 10.1109/TPAMI.2017.2648793], https://hal.inria.fr/hal-01413403

[34] L. GIRIN, T. HUEBER, X. ALAMEDA-PINEDA. *Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", March 2017, vol. 25, n° 3, pp. 662-673 [*DOI :* 10.1109/TASLP.2017.2651398], https://hal.archives-ouvertes.fr/hal-01485540

[35] X. LI, L. GIRIN, R. HORAUD, S. GANNOT. *Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", October 2017, vol. 25, n° 10, pp. 1997 - 2012, https://arxiv.org/abs/1611.01172 - 16 pages, 4 figures, 4 tables [*DOI :* 10.1109/TASLP.2017.2740001], https://hal.inria.fr/hal-01413417

[36] B. MASSÉ, S. BA, R. HORAUD. *Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", December 2017, vol. PP, n° 99, pp. 1-15, https://arxiv.org/abs/1703.04727 [*DOI :* 10.1109/TPAMI.2017.2782819], https://hal.inria.fr/hal-01511414

International Conferences with Proceedings

[37] X. ALAMEDA-PINEDA, A. PILZER, D. XU, N. SEBE, E. RICCI. *Viraliency: Pooling Local Virality*, in "IEEE Conference on Computer Vision and Pattern Recognition", Honolulu, Hawaii, United States, July 2017, https://hal.inria.fr/hal-01558137

[38] *Best Paper*
Y. BAN, X. ALAMEDA-PINEDA, F. BADEIG, S. BA, R. HORAUD. *Tracking a Varying Number of People with a Visually-Controlled Robotic Head*, in "IEEE/RSJ International Conference on Intelligent Robots and Systems", Vancouver, Canada, September 2017, https://hal.inria.fr/hal-01542987.

[39] Y. BAN, L. GIRIN, X. ALAMEDA-PINEDA, R. HORAUD. *Exploiting the Complementarity of Audio and Visual Data in Multi-Speaker Tracking*, in "ICCV Workshop on Computer Vision for Audio-Visual Media", Venezia, Italy, October 2017, https://hal.inria.fr/hal-01577965

[40] V. DROUARD, S. BA, R. HORAUD. *Switching Linear Inverse-Regression Model for Tracking Head Pose*, in "IEEE Winter Conference on Applications of Computer Vision", Santa Rosa, CA, United States, March 2017 [*DOI :* 10.1109/WACV.2017.142], https://hal.inria.fr/hal-01430727

[41] *Best Paper*
I. GEBRU, C. EVERS, P. NAYLOR, R. HORAUD. *Audio-visual Tracking by Density Approximation in a Sequential Bayesian Filtering Framework*, in "IEEE Workshop on Hands-free Speech Communication and Microphone Arrays", San Francisco, CA, United States, IEEE Signal Processing Society, March 2017, Best Paper Award [*DOI :* 10.1109/HSCMA.2017.7895564], https://hal.inria.fr/hal-01452167.

[42] L. GIRIN, R. BADEAU. *On the Use of Latent Mixing Filters in Audio Source Separation*, in "13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)", Grenoble, France, Proc. 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017), Springer, February 2017, vol. 10169, pp. 225-235 [*DOI :* 10.1007/978-3-319-53547-0_22], https://hal.archives-ouvertes.fr/hal-01400965

[43] L. GIRIN, T. HUEBER, X. ALAMEDA-PINEDA. *Adaptation of a Gaussian Mixture Regressor to a New Input Distribution: Extending the C-GMR Framework*, in "LVA ICA 2017- International Conference on Latent Variable Analysis and Signal Separation", Grenoble, France, February 2017, https://hal.inria.fr/hal-01646098

[44] D. KOUNADES-BASTIAN, L. GIRIN, X. ALAMEDA-PINEDA, S. GANNOT, R. HORAUD. *An EM Algorithm for Joint Source Separation and Diarisation of Multichannel Convolutive Speech Mixtures*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01430761

[45] D. KOUNADES-BASTIAN, L. GIRIN, X. ALAMEDA-PINEDA, R. HORAUD, S. GANNOT. *Exploiting the Intermittency of Speech for Joint Separation and Diarization*, in "IEEE Workshop on Applications of Signal

Processing to Audio and Acoustics", New Paltz, NY, United States, October 2017, https://hal.inria.fr/hal-01568813

[46] S. LATHUILIÈRE, G. EVANGELIDIS, R. HORAUD. *Recognition of Group Activities in Videos Based on Single- and Two-Person Descriptors*, in "IEEE Winter Conference on Applications of Computer Vision", Santa Rosa, CA, United States, March 2017 [*DOI :* 10.1109/WACV.2017.31], https://hal.inria.fr/hal-01430732

[47] S. LATHUILIÈRE, R. JUGE, P. MESEJO, R. MUÑOZ-SALINAS, R. HORAUD. *Deep Mixture of Linear Inverse Regressions Applied to Head-Pose Estimation*, in "IEEE Conference on Computer Vision and Pattern Recognition", Honolulu, Hawaii, United States, IEEE Computer Society, July 2017, https://hal.inria.fr/hal-01504847

[48] X. LI, L. GIRIN, R. HORAUD. *An EM Algorithm for Audio Source Separation Based on the Convolutive Transfer Function*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, NY, United States, October 2017, https://hal.inria.fr/hal-01568818

[49] X. LI, L. GIRIN, R. HORAUD. *Audio Source Separation Based on Convolutive Transfer Function and Frequency-Domain Lasso Optimization*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01430754

[50] H. LÖLLMANN, A. MOORE, P. NAYLOR, B. RAFAELY, R. HORAUD, A. MAZEL, W. KELLERMANN. *Microphone Array Signal Processing for Robot Audition*, in "IEEE Workshop on Hands-free Speech Communication and Microphone Arrays", San Francisco, United States, IEEE Signal Processing Society, March 2017 [*DOI :* 10.1109/HSCMA.2017.7895560], https://hal.inria.fr/hal-01485322

[51] D. XU, W. OUYANG, X. ALAMEDA-PINEDA, E. RICCI, X. WANG, N. SEBE. *Learning Deep Structured Multi-Scale Features using Attention-Gated CRFs for Contour Prediction*, in "Advances in Neural Information Processing Systems", Long Beach, United States, December 2017, https://hal.inria.fr/hal-01646112

Other Publications

[52] S. LATHUILIÈRE, B. MASSÉ, P. MESEJO, R. HORAUD. *Neural Network Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction*, November 2017, https://arxiv.org/abs/1711.06834 - 14 pages, https://hal.inria.fr/hal-01643775

[53] X. LI, S. GANNOT, R. HORAUD. *Blind MultiChannel Identification and Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function*, November 2017, https://arxiv.org/abs/1706.03652 - 13 pages, 5 figures, 5 tables, https://hal.inria.fr/hal-01568835

[54] X. LI, L. GIRIN, S. GANNOT, R. HORAUD. *Multichannel Source Separation and Speech Enhancement Using the Convolutive Transfer Function*, November 2017, https://arxiv.org/abs/1711.07911 - 13 pages, 5 figures, https://hal.inria.fr/hal-01645749

[55] R. T. MARRIOTT, A. PASHEVICH, R. HORAUD. *Plane-extraction from depth-data using a Gaussian mixture regression model*, December 2017, https://arxiv.org/abs/1710.01925 - 10 pages, 2 figures, 1 table, https://hal.inria.fr/hal-01663984

[56] K. TOMBRE, L. QUAN, R. HORAUD, P. GROS, C. SCHMID, P. STURM. *In Memoriam Roger Mohr*, Société Informatique de France, September 2017, pp. 91-98, Article qui rappelle la carrière scientifique de Roger Mohr, https://hal.inria.fr/hal-01598085