Activity Report 2019

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. Team, Visitors, External Collaborators

Research Scientists

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

Faculty Member

Laurent Girin [Institut polytechnique de Grenoble, Professor, HDR]

Post-Doctoral Fellows

Simon Leglaive [Inria, Post-Doctoral Fellow, until Aug 2019]
Mostafa Sadeghi [Inria, Post-Doctoral Fellow]

PhD Students

Anand Ballou [Univ Grenoble Alpes, PhD Student, from Nov 2019]
Yutong Ban [Inria, PhD Student, until May 2019]
Xiaoyu Bie [Univ Grenoble Alpes, PhD Student, from Dec 2019]
Guillaume Delorme [Inria, PhD Student]
Wen Guo [Univ Grenoble Alpes, PhD Student, from Oct 2019]
Louis Airale [Univ Grenoble Alpes, PhD Student, from Oct 2019]
Sylvain Guy [Univ Grenoble Alpes, PhD Student]
Yihong Xu [Inria, PhD Student]

Technical staff

Soraya Arias [Inria, Engineer]
Alex Auternaud [Inria, Engineer, from Nov 2019]
Bastien Mourgue [Inria, Engineer, until May 2019]
Guillaume Sarrazin [Inria, Engineer, until May 2019]

Interns and Apprentices

Shiva Shankar Arumugam [Inria, until Sep 2019]
Lucas Dislaire [Inria, from Feb 2019 until Jul 2019]

Hassan Eskandar [Inria, from Feb 2019 until Jul 2019]
Matthieu Laurendeau [Inria, from Jun 2019 until Sep 2019]
Elsa Marie [Inria, from Feb 2019 until Jul 2019]
Adrien Raison [Inria, from May 2019 until Jul 2019]
Vadim Sushko [Inria, from Feb 2019 until Jul 2019]

Administrative Assistant

Nathalie Gillot [Inria, Administrative Assistant]

2. Overall Objectives

2.1. Audio-Visual Machine Perception

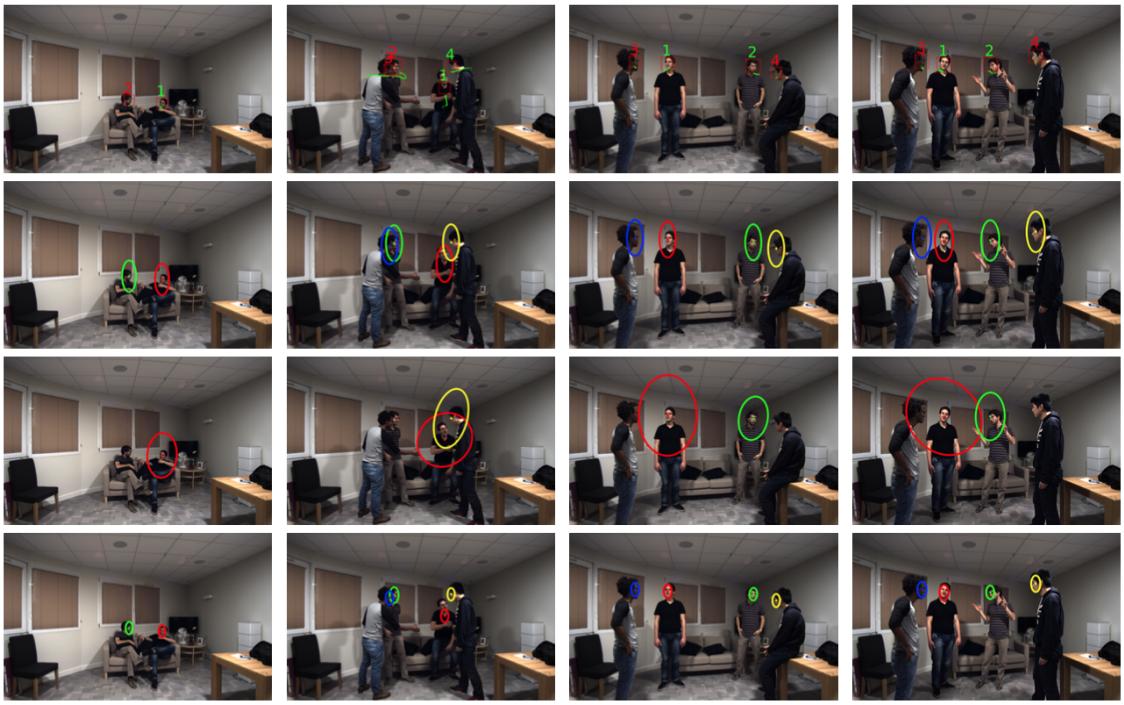


Figure 1. This figure illustrates the audio-visual multiple-person tracking that has been developed by the team [38], [41]. The tracker is based on variational inference [5] and on supervised sound-source localization [10], [29]. Each person is identified with a digit. Green digits denote active speakers while red digits denote silent persons. The next rows show the covariances (uncertainties) associated with the visual (second row), audio (third row) and dynamic (fourth row) contributions for tracking a varying number of persons. Notice the large uncertainty associated with audio and the small uncertainty associated with the dynamics of the tracker. In the light of this example, one may notice the complementary roles played by vision and audio: vision data are more accurate while audio data provide speaker information. 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/lito-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 [25], [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 [9]. 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 [8]. 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 [19], [32]. 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 [20]. 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 [12].

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 [8] and audio-visual learning [10].

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 [35]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [33]. 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 [34]. 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 [21], [17],[16]. 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 [12].

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 [26]. 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 [24], [23]. 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 [7]. 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

4.1.1. IEEE Senior Member.

Xavier Alameda-Pineda has become an IEEE Senior Member on February 1st, 2019. The grade of Senior Member requires experience reflecting professional maturity as an engineer, scientist, educator, technical executive, or originator in IEEE-designated fields for a total of 10 years and have demonstrated 5 years of significant performance.

4.1.2. H2020 Project SPRING

(1 Januray 2020 – 31 December 2023) is a research and innovation action (RIA) with eight partners: Inria Grenoble (coordinator), Università degli Studi di Trento, Czech Technical University Prague, Heriot-Watt University Edinburgh, Bar-Ilan University Tel Aviv, ERM Automatismes Industriels Carpentras, PAL Robotics Barcelona, and Hôpital Broca Paris.. The main objective of SPRING (Socially Pertinent Robots in Gerontological Healthcare) is the development of socially assistive robots with the capacity of performing multimodal multiple-person interaction and open-domain dialogue. In more detail:

- The scientific objective of SPRING is to develop a novel paradigm and novel concept of socially-aware robots, and to conceive innovative methods and algorithms for computer vision, audio processing, sensor-based control, and spoken dialog systems based on modern statistical- and deep-learning to ground the required social robot skills.

- The technological objective of SPRING is to create and launch a brand new generation of robots that are flexible enough to adapt to the needs of the users, and not the other way around.
- The experimental objective of SPRING is twofold: to validate the technology based on HRI experiments in a gerontology hospital, and to assess its acceptability by patients and medical staff.

Website: https://spring-h2020.eu/

4.1.3. ANR JCJC Project ML3RI

(1 March 2020 – 28 February 2024) has been awarded to Xavier Alameda-Pineda. Multi-person robot interaction in the wild (i.e. unconstrained and using only the robot's resources) is nowadays unachievable because of the lack of suitable machine perception and decision-taking models. *Multi-Modal Multi-person Low-Level Learning models for Robot Interaction* (ML3RI) has the ambition to develop the capacity to understand and react to low-level behavioral cues, which is crucial for autonomous robot communication. The main scientific impact of ML3RI is to develop new learning methods and algorithms, thus opening the door to study multi-party conversations with robots. In addition, the project supports open and reproducible research.

4.1.4. MIAI Chair.

The Multidisciplinary Institue in Artificial Intelligence (MIAI) is one of the four AI French institutes launched in 2019 by the French government. MIAI is structured around several chairs, each chair gathering 3-6 researchers as well as postdocs and PhD students. Team members Radu Horaud and Xavier Alameda-Pineda are co-chairs of the *Audio-visual machine perception and interaction for companion robots* chair. The development of methods and algorithms for enabling socially-aware robot behavior with the specific goal of interacting with humans is the core topic. The emphasis is put on unsupervised and weakly supervised learning with audio and visual data, based on Bayesian methods, deep learning and reinforcement learning. It is planned to develop challenging proof-of-concept implementations and demonstrators.

5. New Software and Platforms

5.1. 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 Radu Horaud
- Contact: Radu Horaud
- URL: https://team.inria.fr/perception/research/naolab/

5.2. 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: Antoine Deleforge, Jordi Sanchez-Riera, Radu Horaud and Xavier Alameda-pineda
- Contact: Radu Horaud

5.3. Audio Cue Extractor Library

FUNCTIONAL DESCRIPTION: This module extracts auditory cues from the raw audio streams. The interaural time difference (ITD) is estimated using cross-correlation methods.

- Participants: Antoine Deleforge, Radu Horaud and Soraya Arias
- Contact: Soraya Arias

5.4. Audiovisual Robots and 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/

5.5. GLLiM

Gaussian Locally Linear Mapping

KEYWORDS: Regression - Machine learning - Gaussian mixture

SCIENTIFIC DESCRIPTION: GLLiM is a flexible tool for probabilistic non-linear regression using Gaussian mixtures. Using an inverse regression strategy with a reduced number of parameters, it is particularly suited for high- to low-dimensional regression tasks. It also enables the modeling of additional unobserved non-linear effects on input data. The method was published in [Deleforge et al., IJNS 2015]. The toolbox include an example of application to head pose estimation from synthetic images.

- Participant: Antoine Deleforge
- Contact: Antoine Deleforge
- Publication: hal-00863468, version 3
- URL: https://team.inria.fr/perception/gllim_toolbox/

5.6. Litbot

Live together with robots

KEYWORDS: Speaker Localization - Audio tracking - Visual tracking - NAO Robot - Computer vision - Signal processing

SCIENTIFIC DESCRIPTION: Litbot stands for "Live together with robots". This library aims to provide algorithms and associated software packages to perform audio-visual speaker localization and tracking with a consumer robot (in particular a NAO robot). The scope of this project is two-fold. The first is to develop the robust speaker localization and tracking algorithm in the presence of other audio-visual sources like TV. The second is to modify or optimize the original algorithm to be fit into real-time system. This library benefits from the work done with Online Multiple Sound-Source Localization package developed by X. Li.

FUNCTIONAL DESCRIPTION: This project develops algorithms and associated software packages to perform audio-visual speaker localization and tracking with a consumer robot. This version of the litbot library provides new functions to integrate the Samsung robotic platform to handle ROS middleware (robotic defacto standard) and modifies and optimizes tracking and audio localization processes (better handling of the residual noise signals, performance improved to match real time).

- Participants: Xiaofei Li, Yutong Ban, Soraya Arias, Radu Horaud, Guillaume Sarrazin and Bastien Mourgue
- Contact: Radu Horaud

5.7. Online Multiple Sound-Source Localization

KEYWORDS: Audio signal processing - Multiple sound-source localization - Matlab - Direct-path RTF

FUNCTIONAL DESCRIPTION: This project tackles multiple sound-source localization in noisy and reverberant environments, using binaural recordings of an acoustic scene. It provides Matlab routines to estimate multiple sound source (such as speakers) locations based on direct-path relative transfer function (DP-RTF) estimation.

- Participants: Xiaofei Li and Radu Horaud
- Contact: Radu Horaud
- Publications: Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization - Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization

5.8. RMP

RoMPers

KEYWORDS: Middleware - Robotics - NAO Robot

SCIENTIFIC DESCRIPTION: Robot Middleware developed by Perception. It follows the development done on RobotHandler and NAOLab. Its goal is to provide an abstraction which allows an easy access to robot sensors. In the same time, this high level access is independant of the robot.

FUNCTIONAL DESCRIPTION: Robot Middleware developed by Perception. It follows the development done on RobotHandler and NAOLab. Its goal is to provide an abstraction which allows an easy access to robot sensors. In the same time, this high level access is independant of the robot. And it also provides tools for sensor calibration (audio, video), video annotation, etc

- Participant: Guillaume Sarrazin
- Contact: Soraya Arias

5.9. SE-VAE-alpha-stable

KEYWORDS: Audio signal processing - Speech processing - Deep learning - Neural networks

FUNCTIONAL DESCRIPTION: This software provides an iterative algorithm for enhancing a speech signal in a noisy monophonic recording. The algorithm is detailed in the following paper: "Speech enhancement with variational autoencoders and alpha-stable distributions" Simon Leglaive, Umut Simsekli, Antoine Liutkus, Laurent Girin, Radu Horaud IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Brighton, UK, May 2019

- Contact: Simon Leglaive

5.10. Sound recognition library

FUNCTIONAL DESCRIPTION: This recognition module is based on supervised learning.

- Participants: Maxime Janvier and Radu Horaud
- Contact: Radu Horaud

5.11. SE-VAE-NMF

KEYWORDS: Audio signal processing - Speech processing - Deep learning - Neural networks

FUNCTIONAL DESCRIPTION: This software provides an iterative algorithm for enhancing a speech signal in a noisy monophonic recording. The algorithm is detailed in the following paper: "A variance modeling framework based on variational autoencoders for speech enhancement" Simon Leglaive, Laurent Girin, Radu Horaud Proc. of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, Denmark, September 2018

- Contact: Simon Leglaive
- Publication: hal-01832826v1

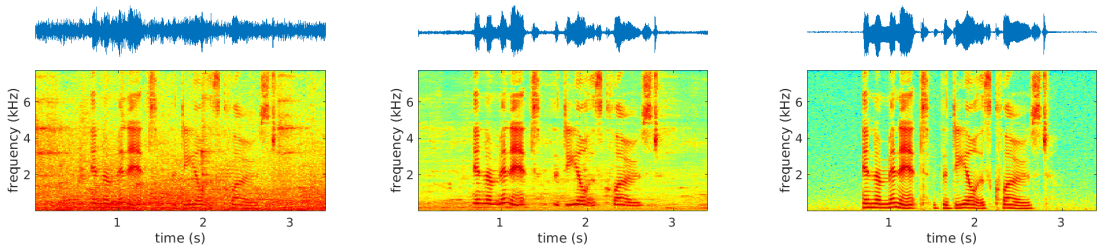
6. New Results

6.1. Multichannel Speech Separation and Enhancement Using the Convolutive Transfer Function

We addressed the problem of speech separation and enhancement from multichannel convolutive and noisy mixtures, *assuming known mixing filters*. We proposed to perform the speech separation and enhancement tasks in the short-time Fourier transform domain, using the convolutive transfer function (CTF) approximation [43], [44]. Compared to time-domain filters, CTF has much less taps, consequently it has less near-common zeros among channels and less computational complexity. The work proposes three speech-source recovery methods, namely: (i) the multichannel inverse filtering method, i.e. the multiple input/output inverse theorem (MINT), is exploited in the CTF domain, and for the multi-source case, (ii) a beamforming-like multichannel inverse filtering method applying single source MINT and using power minimization, which is suitable whenever the source CTFs are not all known, and (iii) a constrained Lasso method, where the sources are recovered by minimizing the ℓ_1-norm to impose their spectral sparsity, with the constraint that the ℓ_2-norm fitting cost, between the microphone signals and the mixing model involving the unknown source signals, is less than a tolerance. The noise can be reduced by setting a tolerance onto the noise power. Experiments under various acoustic conditions are carried out to evaluate the three proposed methods. The comparison between them as well as with the baseline methods is presented.

6.2. Speech Denoising and Enhancement with LTSMs

We have started to address the problems of multichannel speech denoising [45] and enhancement [51] in the short-time Fourier transform (STFT) domain and in the framework of sequence-to-sequence deep learning. In the case of denoising, the magnitude of noisy speech is mapped onto the noise power spectral density. In the case of speech enhancement, the noisy speech is mapped onto clean speech. A long short-time memory (LSTM) network takes as input a sequence of STFT coefficients associated with a frequency bin of multichannel noisy-speech signals. The network's output is a sequence of single-channel cleaned speech at the same frequency bin. We propose several clean-speech network targets, namely, the magnitude ratio mask, the complex ideal ratio mask, the STFT coefficients and spatial filtering [54]. A prominent feature of the proposed model is that the same LSTM architecture, with identical parameters, is trained across frequency bins. The proposed method is referred to as narrow-band deep filtering. This choice stays in contrast with traditional wide-band speech enhancement methods. The proposed deep filter is able to discriminate between speech and noise by exploiting their different temporal and spatial characteristics: speech is non-stationary and spatially coherent while noise is relatively stationary and weakly correlated across channels. This is similar in spirit with unsupervised techniques, such as spectral subtraction and beamforming. We describe extensive experiments with both mixed signals (noise is added to clean speech) and real signals (live recordings). We empirically evaluate the proposed architecture variants using speech enhancement and speech recognition metrics, and we compare our results with the results obtained with several state of the art methods. In the light of these experiments we conclude that narrow-band deep filtering has very good performance, and excellent generalization capabilities in terms of speaker variability and noise type, e.g. Figure 2.



Noisy speech input Filtered output Ground-truth speech signal

Figure 2. An example of narrow-band deep filtering for speech enhancement [54]. Waveforms and spectrograms of the noisy (unprocessed) input, the filtered output and the ground-truth clean-speech. Four microphones were used in this example. The signal-to-noise ratio in this example is 0 dB.

Website: https://team.inria.fr/perception/research/mse-lstm/.

6.3. Multichannel Speech Enhancement with Variational Auto-Encoder

We addressed speaker-independent multichannel speech enhancement in unknown noisy environments. Our work is based on a well-established multichannel local Gaussian modeling framework. We propose to use a neural network for modeling the speech spectro-temporal content. The parameters of this supervised model are learned using the framework of variational autoencoders. The noisy recording environment is supposed to be unknown, so the noise spectro-temporal modeling remains unsupervised and is based on non-negative matrix factorization (NMF). We develop a Monte Carlo expectation-maximization algorithm and we experimentally show that the proposed approach outperforms its NMF-based counterpart, where speech is modeled using supervised NMF [49].
Website: https://team.inria.fr/perception/research/icassp-2019-mvae/

6.4. Audio-visual Speech Enhancement with Conditional Variational Auto-Encoder

Variational auto-encoders (VAEs) are deep generative latent variable models that can be used for learning the distribution of complex data. VAEs have been successfully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. One advantage of this generative approach is that it does not require pairs of clean and noisy speech signals at training. In this work, we propose audio-visual variants of VAEs for single-channel and speaker-independent speech enhancement. We developed a conditional VAE (CVAE) where the audio speech generative process is conditioned on visual information of the lip region, e.g. Figure 3. At test time, the audio-visual speech generative model is combined with a noise model, based on nonnegative matrix factorization, and speech enhancement relies on a Monte Carlo expectation-maximization algorithm. Experiments were conducted with the recently published NTCD-TIMIT dataset. The results confirm that the proposed audio-visual CVAE effectively fuse audio and visual information, and it improves the speech enhancement performance compared with the audio-only VAE model, especially when the speech signal is highly corrupted by noise. We also showed that the proposed unsupervised audio-visual speech enhancement approach outperforms a state-of-the-art supervised deep learning method [55].
Website: https://team.inria.fr/perception/research/av-vae-se/

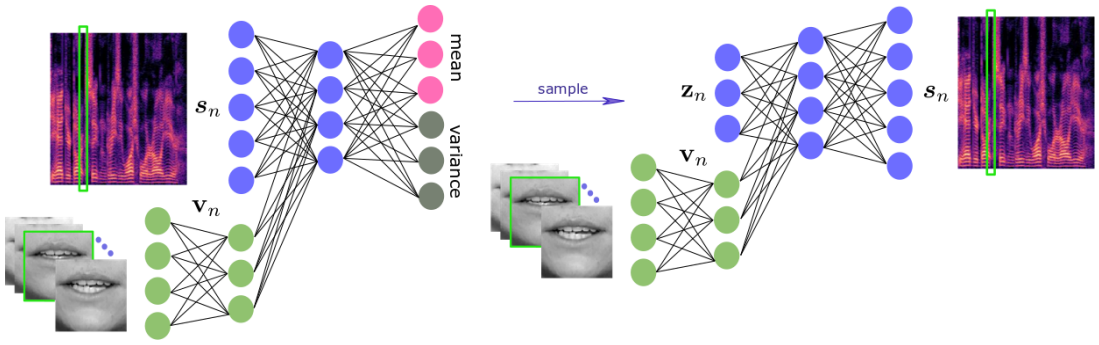


Figure 3. We proposed a conditional variational auto-encoder architecture for fusing audio and visual data for speech enhancement [55].

6.5. Variational Bayesian Inference of Audio-visual Speaker Tracking

We addressed the problem of tracking multiple speakers via the fusion of visual and auditory information [36]. We proposed to exploit the complementary nature of these two modalities in order to accurately estimate smooth trajectories of the tracked persons, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status – either speaking or silent – of each tracked person along time, e.g. Figure 1. We proposed to cast the problem at hand into a generative audio-visual fusion (or association) model formulated as a latent-variable temporal graphical model. This may well be viewed as the problem of maximizing the posterior joint distribution of a set of continuous and discrete latent variables given the past and current observations, which is intractable. We proposed a variational inference model which amounts to approximate the joint distribution with a factorized distribution. The solution takes the form of closed-form expectation maximization procedures using Gaussian distributions [38]. We described in detail the inference algorithm, we evaluated its performance and we compared the results with several baseline methods. These experiments show that the proposed audio-visual tracker performs well in informal meetings involving a time-varying number of people. Real-time versions of the algorithm were implemented on our robotic platform [47].
Website: https://team.inria.fr/perception/research/var-av-track/.

6.6. Detection, Localization and Tracking of Multiple Audio Sources

We addressed the problem of online detection, localization and tracking of multiple moving speakers in reverberant environments [36]. The work has the following contributions. We used the direct-path relative transfer function (DP-RTF), an inter-channel feature that encodes acoustic information robust against reverberation, and we proposed an online algorithm well suited for estimating DP-RTFs associated with moving audio sources. Another crucial ingredient of the proposed method is its ability to properly assign DP-RTFs to audio-source directions. Towards this goal, we adopted a maximum-likelihood formulation and we proposed to use the exponentiated gradient (EG) to efficiently update source-direction estimates starting from their currently available values. The problem of multiple-speaker tracking is computationally intractable because the number of possible associations between observed source directions and physical speakers grows exponentially with time. We adopt a Bayesian framework and we proposed two variational approximations of the posterior filtering distributions associated with multiple speaker tracking, as well as two efficient variational expectation maximization (VEM) solvers [41], [37]. The proposed online localization and tracking methods were thoroughly evaluated using two datasets that contain recordings performed in real environments.
Websites:

https://team.inria.fr/perception/research/audiotrack-vonm/
https://team.inria.fr/perception/research/multi-speaker-tracking/.

6.7. The Kinovis Multiple-Speaker Tracking Datasets

The Kinovis multiple speaker tracking (Kinovis-MST) datasets contain live acoustic recordings of multiple moving speakers in a reverberant environment. The data were recorded in the Kinovis multiple-camera laboratory at Inria Grenoble Rhône-Alpes. The room size is $10.2 \times 9.9 \times 5.6$ meters with T60 = 0.53 seconds. The data were recorded with four microphones embedded into the head of a NAO robot. Because there is a fan located inside the robot head nearby the microphones, there is a fair amount of stationary and spatially correlated microphone noise. The signal-to-noise ratio of the microphone signals is of approximately 2.7 dB. The recordings contain between one and three moving participants that speak naturally, hence the number of active speech sources varies over time. The robot-to-speaker distance ranges between 1.5 and 3.5 meters. Ground-truth trajectories and speech activity information were obtained in the following way. Participants were wearing optical markers placed on their heads such that the Kinovis motion capture system provides accurate 3D trajectories for each participant. Moreover, an infrared marker is placed on the participants' foreheads. This enables the identification of each participant over time. Whenever time a participant is silent, he/she hides his/her infrared marker, thus allowing speaking/silent annotations of the recordings.
Website: https://team.inria.fr/perception/the-kinovis-mst-dataset/.

6.8. Deep Regression

Deep learning revolutionized data science, and recently its popularity has grown exponentially, as did the amount of papers employing deep networks. Vision tasks, such as human pose estimation, did not escape from this trend. There is a large number of deep models, where small changes in the network architecture, or in the data pre-processing, together with the stochastic nature of the optimization procedures, produce notably different results, making extremely difficult to sift methods that significantly outperform others. This situation motivates the current study, in which we perform a systematic evaluation and statistical analysis of vanilla deep regression, i.e. convolutional neural networks with a linear regression top layer. This is the first comprehensive analysis of deep regression techniques. We perform experiments on four vision problems, and report confidence intervals for the median performance as well as the statistical significance of the results, if any. Surprisingly, the variability due to different data pre-processing procedures generally eclipses the variability due to modifications in the network architecture. Our results reinforce the hypothesis according to which, in general, a general-purpose network (e.g. VGG-16 or ResNet-50) adequately tuned can yield results close to the state-of-the-art without having to resort to more complex and ad-hoc regression models, [40].
Website: https://team.inria.fr/perception/research/deep-regression/.

6.9. Deep Reinforcement Learning for Audio-Visual Robot Control

More recently, we investigated the use of reinforcement learning (RL) as an alternative to sensor-based robot control. The robotic task consists of turning the robot head (gaze control) towards speaking people. The method is more general in spirit than visual (or audio) servoing because it can handle an arbitrary number of speaking or non speaking persons and it can improve its behavior online, as the robot experiences new situations. An overview of the proposed method is shown in Fig. 4. The reinforcement learning formulation enables a robot to learn where to look for people and to favor speaking people via a trial-and-error strategy.

Past, present and future HRI developments require datasets for training, validation, test as well as for benchmarking. HRI datasets are challenging because it is not easy to record realistic interactions between a robot and users. RL avoids systematic recourse to annotated datasets for training. In [39] we proposed the use of a simulated environment for pre-training the RL parameters, thus avoiding spending hours of tedious interaction.
Website: https://team.inria.fr/perception/research/deep-rl-for-gaze-control/.

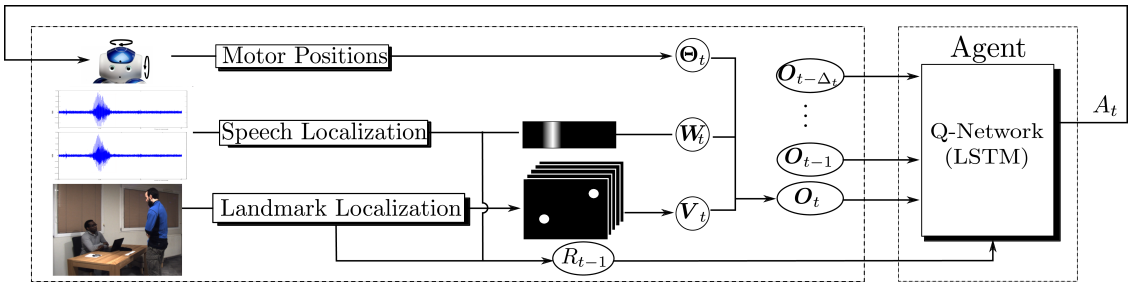


Figure 4. Overview of the proposed deep RL method for controlling the gaze of a robot. At each time index t, audio and visual data are represented as binary maps which, together with motor positions, form the set of observations $\mathbf{O}_t$. A motor action A_t (rotate the head left, right, up, down, or stay still) is selected based on past and present observations via maximization of current and future rewards. The rewards R are based on the number of visible persons as well as on the presence of speech sources in the camera field of view. We use a deep Q-network (DQN) model that can be learned both off-line and on-line. Please consult [39] for further details.

7. Partnerships and Cooperations

7.1. European Initiatives

7.1.1. Collaborations with Major European Organizations

Universitat Politècnica de Catalunya (UPC), Spain

Physical complex Interactions and Multi-person Pose Estimation (PIMPE) is three year project financed by IDEX. The scientific challenges of PIMPE are the followings: (i) Modeling multi-person interactions in full-body pose estimation, (ii) Estimating human poses in complex multi-person physical interactions, and (iii) Generating controlled and realistic multi-person complex pose images.

7.2. International Research Visitors

7.2.1. Research Stays Abroad

Xavier Alameda-Pineda spent three months at the University of Verona, Italy.

Yihong Xu (Ph.D. student) spent three months at the Technical University Munich, Germany.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events: Organisation

8.1.1.1. Member of the Organizing Committees

Xavier Alameda-Pineda

- 10th International Workshop on Human Behavior Understanding, in conjunction with the 2019 International Conference on Computer Vision (ICCV).
- 1st Workshop on Fairness, Accountability and Transparency in Multimedia, in conjunction with the 2019 ACM International Conference on Multimedia (ACM MM).
- 3rd Workshop on Understanding Subjective Properties of Data, Focus on Fashion and Subjective Search, in conjunction with the 2019 International Conference on Computer Vision and Pattern Recognition (CVPR).

8.1.2. Scientific Events: Selection

8.1.2.1. Member of the Conference Program Committees

- Area Chair of the 2019 ACM International Conference on Multimedia (ACM MM).
- Area Chair of the 2019 International Conference on Image Analysis and Processing (ICIAP).

8.1.2.2. Reviewer

Xavier Alameda-Pineda: CVPR 2019, NeurIPS 2019, ICCV 2019, ACII 2019, ICLR 2019, ICML 2019, ICIP 2019.

8.1.3. Journal

8.1.3.1. Member of the Editorial Boards

Xavier Alameda-Pineda is Associated Editor of the ACM Transactions on Multimedia Computing Communications and Applications

8.1.3.2. Reviewer - Reviewing Activities

Xavier Alameda-Pineda: TPAMI, TASLP, TMM.

8.1.4. Invited Talks

Xavier Alameda-Pineda:

- Significancy & Robustness in Deep Regression (July 2019) at University of Trento
- Probabilistic and deep methods for human behavior understanding (July 2019) at Media Integration and Communication Center

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Master : Xavier Alameda-Pineda, Fundamentals of Probabilistic Data Mining, 18h, M2, UGA, France.

Master : Xavier Alameda-Pineda, Category Learning and Object Recognition, 11h, M2, UGA, France.

Master : Xavier Alameda-Pineda, Advanced Learning Models, 13.5h, M2, UGA, France.

8.2.2. Supervision

PhD: Yutong Ban, Audio-visual multiple-speaker tracking for robot perception [36], Université Grenoble Alpes, May 2019, Xavier Alameda-Pineda and Radu Horaud,

PhD in progress: Guillaume Delorme, Deep Person Re-identification, October 2017, Xavier Alameda-Pineda and Radu Horaud,

PhD in progress: Yihong Xu, Deep Multiple-person Tracking, October 2018, Xavier Alameda-Pineda and Radu Horaud,

PhD in progress: Wen Guo, Deep Human Pose, October 2019, Xavier Alameda-Pineda and Radu Horaud,

PhD in progress: Anand Ballou, Deep Reinforcement Learning for Robot Control, November 2019, Xavier Alameda-Pineda and Radu Horaud,

PhD in progress: Louis Airale, Data Generation for Deep Multimodal Interaction Algorithms, October 2019, Xavier Alameda-Pineda and Dominique Vaufreydaz,

PhD in progress: Xiaoyu Bie, Deep Generative Methods for Audio and Vision, December 2019, Xavier Alameda-Pineda and Laurent Girin.

8.2.3. Juries

Xavier Alameda-Pineda belonged to the following PhD Juries as “examinateur”:

- Yutong Ban (University Grenoble-Alpes)
- Irtiza Hasan (University of Verona)
- Theodoros Tsesmelis (University of Verona)

9. 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] X. ALAMEDA-PINEDA, E. RICCI, N. SEBE. *Multimodal behavior analysis in the wild: Advances and challenges*, Academic Press (Elsevier), December 2018, https://hal.inria.fr/hal-01858395

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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^{o} 3, pp. 1428–1440 [*DOI :* 10.1109/TIP.2017.2654165], https://hal.inria.fr/hal-01413406

[12] 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° 11, pp. 2178–2192 [*DOI :* 10.1109/TPAMI.2015.2400465], https://hal.archives-ouvertes.fr/hal-01110031

[13] 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 2018, vol. 40, n° 6, pp. 1397–1410, https://arxiv.org/abs/1609.01466 [*DOI :* 10.1109/TPAMI.2017.2717829], https://hal.inria.fr/hal-01413414

[14] 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° 12, pp. 2402–2415 [*DOI :* 10.1109/TPAMI.2016.2522425], https://hal.inria.fr/hal-01261374

[15] 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", July 2018, vol. 40, n° 5, pp. 1086–1099, https://arxiv.org/abs/1603.09725 [*DOI :* 10.1109/TPAMI.2017.2648793], https://hal.inria.fr/hal-01413403

[16] 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

[17] 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

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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] X. LI, L. GIRIN, F. BADEIG, R. HORAUD. *Reverberant Sound Localization with a Robot Head Based on Direct-Path Relative Transfer Function*, in "IEEE/RSJ International Conference on Intelligent Robots and Systems", Daejeon, South Korea, IEEE, October 2016, pp. 2819–2826 [*DOI :* 10.1109/IROS.2016.7759437], https://hal.inria.fr/hal-01349771

[29] 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

[30] 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, 16 pages, 4 figures, 4 tables [*DOI :* 10.1109/TASLP.2017.2740001], https://hal.inria.fr/hal-01413417

[31] 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", November 2018, vol. 40, n° 11, pp. 2711–2724, https://arxiv.org/abs/1703.04727 [*DOI :* 10.1109/TPAMI.2017.2782819], https://hal.inria.fr/hal-01511414

[32] 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

[33] 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

[34] 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

[35] 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

[36] Y. BAN. *Audio-visual multiple-speaker tracking for robot perception*, Université Grenoble Alpes, May 2019, https://tel.archives-ouvertes.fr/tel-02163418

Articles in International Peer-Reviewed Journals

[37] Y. BAN, X. ALAMEDA-PINEDA, C. EVERS, R. HORAUD. *Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM*, in "IEEE Signal Processing Letters", June 2019, vol. 26, n° 6, pp. 798 - 802, https://arxiv.org/abs/1812.08246 [*DOI :* 10.1109/LSP.2019.2908376], https://hal.inria.fr/hal-01969050

[38] Y. BAN, X. ALAMEDA-PINEDA, L. GIRIN, R. HORAUD. *Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", November 2019, vol. 42, pp. 1-17, https://arxiv.org/abs/1809.10961 [*DOI :* 10.1109/TPAMI.2019.2953020], https://hal.inria.fr/hal-01950866

[39] S. LATHUILIÈRE, B. MASSÉ, P. MESEJO, R. HORAUD. *Neural Network Based Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction*, in "Pattern Recognition Letters", February 2019, vol. 118, pp. 61-71, https://arxiv.org/abs/1711.06834 [*DOI :* 10.1016/j.patrec.2018.05.023], https://hal.inria.fr/hal-01643775

[40] S. LATHUILIÈRE, P. MESEJO, X. ALAMEDA-PINEDA, R. HORAUD. *A Comprehensive Analysis of Deep Regression*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", April 2019, vol. 41, pp. 1-17, https://arxiv.org/abs/1803.08450 [*DOI :* 10.1109/TPAMI.2019.2910523], https://hal.inria.fr/hal-01754839

[41] X. LI, Y. BAN, L. GIRIN, X. ALAMEDA-PINEDA, R. HORAUD. *Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments*, in "IEEE Journal of Selected Topics in Signal Processing", March 2019, vol. 13, n° 1, pp. 88-103, https://arxiv.org/abs/1809.10936 [*DOI :* 10.1109/JSTSP.2019.2903472], https://hal.inria.fr/hal-01851985

[42] X. LI, L. GIRIN, S. GANNOT, R. HORAUD. *Multichannel Online Dereverberation based on Spectral Magnitude Inverse Filtering*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", May 2019, vol. 27, n° 9, pp. 1365-1377, https://arxiv.org/abs/1812.08471 [*DOI :* 10.1109/TASLP.2019.2919183], https://hal.inria.fr/hal-01969041

[43] X. LI, L. GIRIN, S. GANNOT, R. HORAUD. *Multichannel Speech Separation and Enhancement Using the Convolutive Transfer Function*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", March 2019, vol. 27, n° 3, pp. 645-659, https://arxiv.org/abs/1711.07911 [*DOI :* 10.1109/TASLP.2019.2892412], https://hal.inria.fr/hal-01799809

[44] X. LI, L. GIRIN, R. HORAUD. *Expectation-Maximization for Speech Source Separation using Convolutive Transfer Function*, in "CAAI Transactions on Intelligent Technologies", March 2019, vol. 4, n^{o} 1, pp. 47 - 53 [*DOI :* 10.1049/TRIT.2018.1061], https://hal.inria.fr/hal-01982250

[45] X. LI, S. LEGLAIVE, L. GIRIN, R. HORAUD. *Audio-noise Power Spectral Density Estimation Using Long Short-term Memory*, in "IEEE Signal Processing Letters", June 2019, vol. 26, n^{o} 6, pp. 918-922, https://arxiv.org/abs/1904.05166 [*DOI :* 10.1109/LSP.2019.2911879], https://hal.inria.fr/hal-02100059

Invited Conferences

[46] F. FORBES, A. DELEFORGE, R. HORAUD, E. PERTHAME. *Robust non-linear regression approach for generalized inverse problems in a high dimensional setting*, in "AIP 2019 - Applied Inverse Problem conference", Grenoble, France, July 2019, https://hal.archives-ouvertes.fr/hal-02415115

International Conferences with Proceedings

[47] X. ALAMEDA-PINEDA, S. ARIAS, Y. BAN, G. DELORME, L. GIRIN, R. HORAUD, X. LI, B. MOURGUE, G. SARRAZIN. *Audio-Visual Variational Fusion for Multi-Person Tracking with Robots*, in "ACMMM 2019 - 27th ACM International Conference on Multimedia", Nice, France, ACM Press, October 2019, pp. 1059-1061 [*DOI :* 10.1145/3343031.3350590], https://hal.inria.fr/hal-02354514

[48] L. GIRIN, F. ROCHE, T. HUEBER, S. LEGLAIVE. *Notes on the use of variational autoencoders for speech and audio spectrogram modeling*, in "DAFx 2019 - 22nd International Conference on Digital Audio Effects", Birmingham, United Kingdom, February 2019, pp. 1-8, https://hal.archives-ouvertes.fr/hal-02349385

[49] S. LEGLAIVE, L. GIRIN, R. HORAUD. *Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization*, in "ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brighton, United Kingdom, IEEE, May 2019, pp. 101-105 [*DOI :* 10.1109/ICASSP.2019.8683704], https://hal.inria.fr/hal-02005102

[50] S. LEGLAIVE, U. SIMSEKLI, A. LIUTKUS, L. GIRIN, R. HORAUD. *Speech enhancement with variational autoencoders and alpha-stable distributions*, in "ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing", Brighton, United Kingdom, IEEE, 2019, pp. 541-545, https://arxiv.org/abs/1902.03926 [*DOI :* 10.1109/ICASSP.2019.8682546], https://hal.inria.fr/hal-02005106

[51] X. LI, R. HORAUD. *Multichannel Speech Enhancement Based on Time-frequency Masking Using Subband Long Short-Term Memory*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, NY, United States, October 2019, pp. 1-5, https://hal.inria.fr/hal-02264247

[52] B. MASSÉ, S. LATHUILIÈRE, P. MESEJO, R. HORAUD. *Extended Gaze Following: Detecting Objects in Videos Beyond the Camera Field of View*, in "FG 2019 - 14th IEEE International Conference on Automatic Face and Gesture Recognition", Lille, France, IEEE, May 2019, pp. 1-8 [*DOI :* 10.1109/FG.2019.8756555], https://hal.inria.fr/hal-02054236

Other Publications

[53] S. LEGLAIVE, X. ALAMEDA-PINEDA, L. GIRIN, R. HORAUD. *A Recurrent Variational Autoencoder for Speech Enhancement*, October 2019, https://arxiv.org/abs/1910.10942 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02329000

[54] X. LI, R. HORAUD. *Narrow-band Deep Filtering for Multichannel Speech Enhancement*, November 2019, https://arxiv.org/abs/1911.10791 - working paper or preprint, https://hal.inria.fr/hal-02378413

[55] M. SADEGHI, S. LEGLAIVE, X. ALAMEDA-PINEDA, L. GIRIN, R. HORAUD. *Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoder*, November 2019, https://arxiv.org/abs/1908.02590 - Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing, https://hal.inria.fr/hal-02364900