Activity Report 2018

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]
Pablo Mesejo Santiago [Inria, Starting Research Position, until March 2018]

Faculty Member

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

Post-Doctoral Fellows

Simon Leglaive [Inria, since February 2018]
Mostafa Sadeghi [Inria, since August 2018]

PhD Students

Yutong Ban [Inria]
Guillaume Delorme [Inria]
Sylvain Guy [Univ Grenoble Alpes]
Stephane Lathuiliere [Inria, until May 2018]
Benoit Masse [Univ Grenoble Alpes, until November 2018]
Yihong Xu [Inria, since September 2018]

Technical staff

Soraya Arias [Inria]
Bastien Mourgue [Inria]
Guillaume Sarrazin [Inria]

Interns

Victor Bros [Inria, from Jun 2018 until Jul 2018]
Caroline Dam Hieu [Inria, from May 2018 until Aug 2018]
Fatbardha Hoxha [Inria, from Apr 2018 until Jul 2018]

Administrative Assistant

Nathalie Gillot [Inria]

Visiting Scientists

Christine Evers [Imperial College London, from Jul 2018 until Aug 2018]
Sharon Gannot [Bar Ilan University, from Jan 2018 until Feb 2018]
Tomislav Pribanic [University of Zagreb, from Apr 2018 until Aug 2018]

2. Overall Objectives

2.1. Audio-Visual Machine Perception

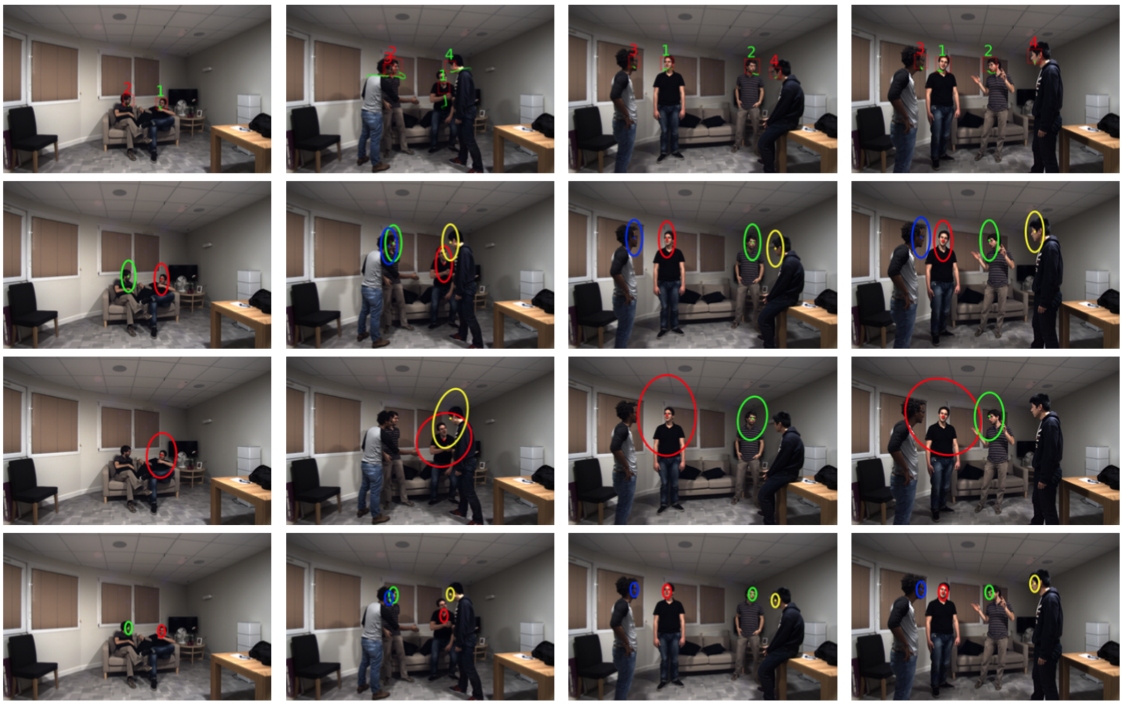

Figure 1. This figure illustrates the audio-visual multiple-person tracking that has been developed by the team [44], [56], [58]. The tracker is based on variational inference [4] and on supervised sound-source localization [9], [26]. Each person is identified with a digit. Green digits denote active 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 [22], [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 [8]. 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 [7]. 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 [16], [28]. 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 [17]. 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 [11].

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 [7] and audio-visual learning [9].

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 [31]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [29]. 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 [30]. 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 [18], [14],[13]. 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 [11].

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 [23]. 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 [21], [20]. 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 [6]. 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

- 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 started in February 2018 for a duration of 12 months. Software packages enabling companion robots to robustly interact with multiple users are being developed.
  Website: https://team.inria.fr/perception/projects/poc-vhialab/
- The 2018 winner of the prestigious ACM Special Interest Group on Multimedia (SIGMM) Rising Star Award is Perception team member Dr. Xavier Alameda-Pineda. The award is given in recognition of Xavier's contributions to multimodal social behavior understanding.
  Website: http://sigmm.org/news/sigmm_rising_star_award_2018
- A book was published by Academic Press (Elsevier), entitled “Multimodal Behavior Analysis in the Wild", co-edited by Xavier Alameda Pineda, Elisa Ricci (Fondazione Bruno Kessler and University of Trento) and Nicu Sebe (University of Trento). The book gathers 20 chapters written by 75 researchers from all over the world [53].

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.

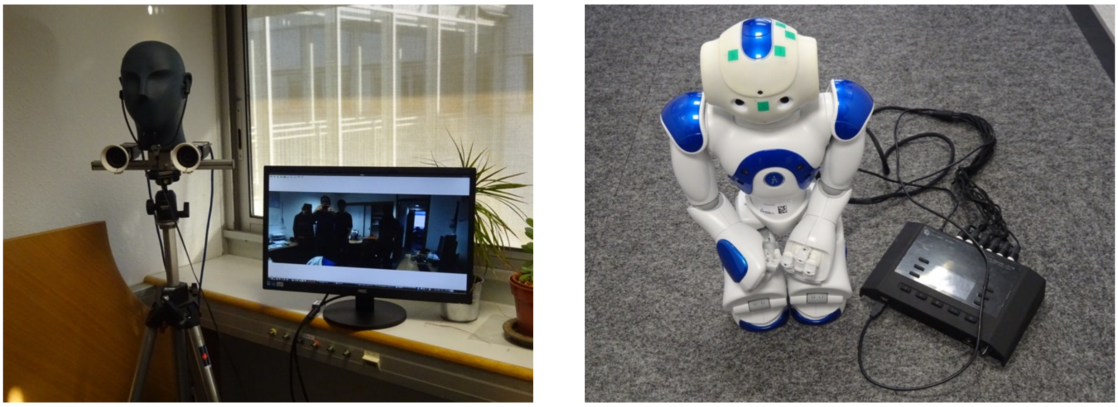

Figure 2. The Popeye+ audio-visual platform (left) delivers high-quality, high-resolution and wide-angle images at 30FPS. The NAO prototype used by PERCEPTION in the EARS STREP project has a twelve-channel spherical microphone array synchronized with a stereo camera pair.

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.

In order to manage the information flow gathered by all these sensors, we implemented several middleware packages. In 2012 we implemented Robotics Services Bus (RSB) developed by the University of Bielefeld. Subsequently (2015-2016) the PERCEPTION team developed 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. In 2017 we abandoned RSB and NAOLab and converted all our robotics software packages to ROS (Robotic Operating System).

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

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 [39]. 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 Dereverberation and Noise Reduction Using the Convolutive Transfer Function

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 [38].
Website: https://team.inria.fr/perception/research/ctf-dereverberation/.

6.3. Speech Enhancement with a Variational Auto-Encoder

We addressed the problem of enhancing speech signals in noisy mixtures using a source separation approach. We explored the use of neural networks as an alternative to a popular speech variance model based on supervised non-negative matrix factorization (NMF). More precisely, we use a variational auto-encoder as a speaker-independent supervised generative speech model, highlighting the conceptual similarities that this approach shares with its NMF-based counterpart. In order to be free of generalization issues regarding the noisy recording environments, we follow the approach of having a supervised model only for the target speech signal, the noise model being based on unsupervised NMF. We developed a Monte Carlo expectation-maximization algorithm for inferring the latent variables in the variational auto-encoder and estimating the unsupervised model parameters. Experiments show that the proposed method outperforms a semi-supervised NMF baseline and a state-of-the-art fully supervised deep learning approach.
Website: https://team.inria.fr/perception/research/ieee-mlsp-2018/.

6.4. 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 [36]. 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 [12], while several options for the speaking turns dynamics are fulfilled by a multi-case transition model. The modules that translate raw audio and visual data into image observations are also described in detail. The performance of the proposed method are tested on challenging datasets that are available from recent contributions which are used as baselines for comparison [36].
Websites:
https://team.inria.fr/perception/research/wdgmm/,
https://team.inria.fr/perception/research/speakerloc/,
https://team.inria.fr/perception/research/speechturndet/, and
https://team.inria.fr/perception/research/avdiarization/.

6.5. 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 [42].

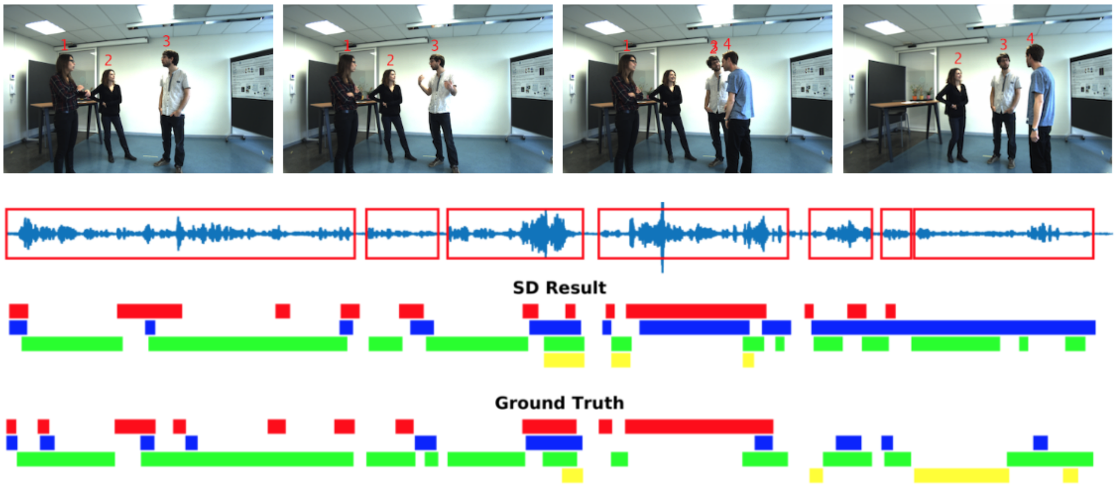

Figure 3. This figure illustrates the audiovisual tracking and diarization method that we have recently developed. First row: A number is associated with each tracked person. Second row: diarization result. Third row: the ground truth diarization. Fourth row: acoustic signal recorded by one of the two microphones.

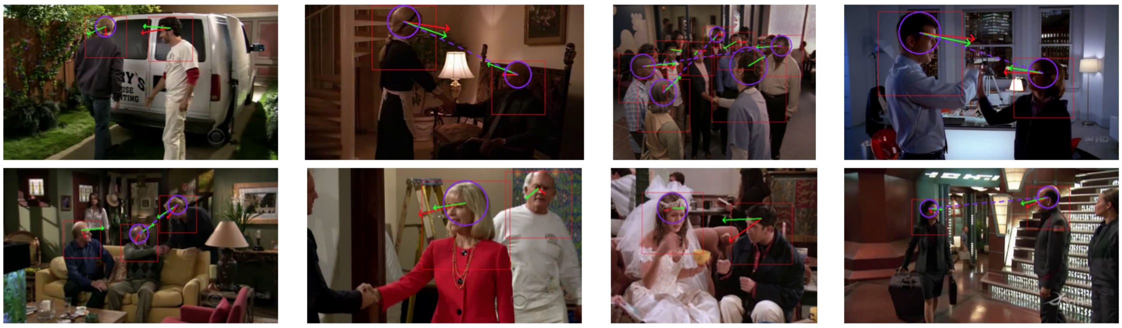

Figure 4. This figure shows some results obtained with the LAEO dataset. The top row shows results obtained with coarse head orientation and the bottom row shows results obtained with fine head orientation. Head orientations are shown with red arrows. The algorithm infers gaze directions (green arrows) and VFOAs (blue circles). People looking at each others are shown with a dashed blue line.

Website: https://team.inria.fr/perception/research/eye-gaze/.

6.6. Variational Bayesian Inference of Multiple-Person Tracking

We addressed the problem of tracking multiple speakers using audio information or via the fusion of visual and auditory information. 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 propose a variational inference model which amounts to approximate the joint distribution with a factorized distribution. The solutions take the form of closed-form expectation maximization procedures using Gaussian distributions [44], [58], [56] or the von Mises distribution for circular variables [55]. We described in detail the inference algorithms, we evaluate their performance and we compared them with several baseline methods. These experiments show that the proposed audio and audio-visual trackers perform well in informal meetings involving a time-varying number of people.
Websites:
https://team.inria.fr/perception/research/var-av-track/,
https://team.inria.fr/perception/research/audiotrack-vonm/.

6.7. High-Dimensional and Deep Regression

One of the most important achievements for the last years has been the development of high-dimensional to low-dimensional regression methods. The motivation for investigating this problem raised from several problems that appeared both in audio signal processing and in computer vision. Indeed, often the task in data-driven methods is to recover low-dimensional properties and associated parameterizations from high-dimensional observations. Traditionally, this can be formulated as either an unsupervised method (dimensionality reduction of manifold learning) or a supervised method (regression). We developed a learning methodology at the crossroads of these two alternatives: the output variable can be either fully observed or partially observed. This was cast into the framework of linear-Gaussian mixture models in conjunction with the concept of inverse regression. It gave rise to several closed-form and approximate inference algorithms [8]. The method is referred to as *Gaussian locally linear mapping*, or GLLiM. As already mentioned, high-dimensional regression is useful in a number of data processing tasks because the sensory data often lies in high-dimensional spaces. Each one of these tasks required a special-purpose version of our general framework. Sound-source localization was the first to benefit from our formulation. Nevertheless, the sparse nature of speech spectrograms required the development of a GLLiM version that is able to with full-spectrum sounds and to test with sparse-spectrum ones [9]. This could be immediately applied to audio-visual alignment and to sound-source separation and localization [7].

In conjunction with our computer vision work, high-dimensional regression is a very useful methodology since visual features, obtained either by hand-crafted feature extraction methods or using convolutional neural networks, lie in high-dimensional spaces. Such properties as object pose lie in low-dimensional spaces and must be extracted from features. We took such an approach and proposed a head pose estimator [10]. Visual tracking can also benefit from GLLiM. Indeed, it is not practical to track objects based on high-dimensional features. We therefore combined GLLiM with switching linear dynamic systems. In 2018 we proposed a robust deep regression method [46]. In parallel we thoroughly benchmarked and analyzed deep regression tasks using several CNN architectures [57].

6.8. Human-Robot Interaction

Audio-visual fusion raises interesting problems whenever it is implemented onto a robot. Robotic platforms have their own hardware and software constraints. In addition, commercialized robots have economical constraints which leads to the use of cheap components. A robot must be reactive to changes in its environment and hence it must take fast decisions. This often implies that most of the computing resources must be onboard of the robot.

Over the last decade we have tried to do our best to take these constraints into account. Starting from our scientific developments, we put a lot of efforts into robotics implementations. For example, the audio-visual fusion method described in [2] used a specific robotic middleware that allowed fast communication between the robot and an external computing unit. Subsequently we developed a powerful software package that enables distributed computing. We also put a lot of emphasis on the implementation of low-level audio and visual processing algorithms. In particular, our single- and multiple audio source methods were implemented in real time onto the humanoid robot NAO [25], [50]. The multiple person tracker [4] was also implemented onto our robotic platforms [5], e.g. Figure 5.

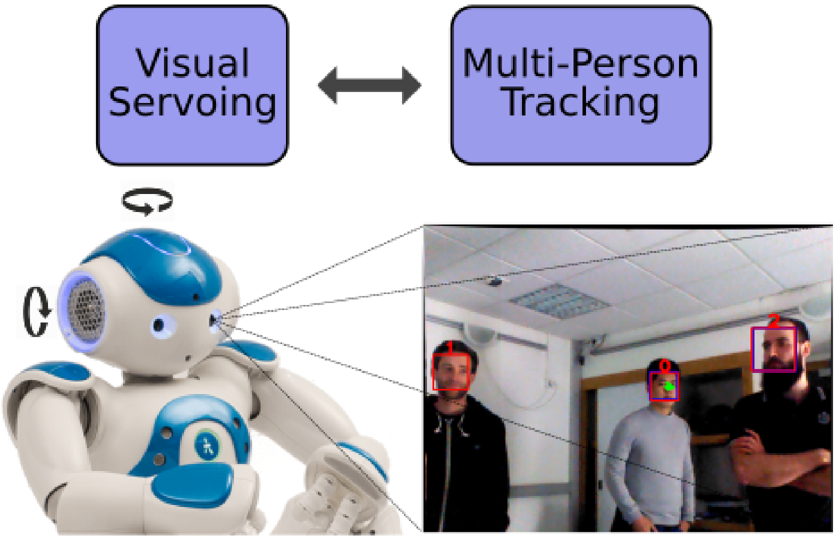

Figure 5. The multi-person tracking method is combined with a visual servoing module. The latter estimates the optimal robot commands and the expected impact of the tracked person locations. The multi-person tracking module refines the locations of the persons with the new observations and the information provided by the visual servoing.

More recently, we investigated the use of reinforcement learning (RL) as an alternative to sensor-based robot control [45], [37]. 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. 6. 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

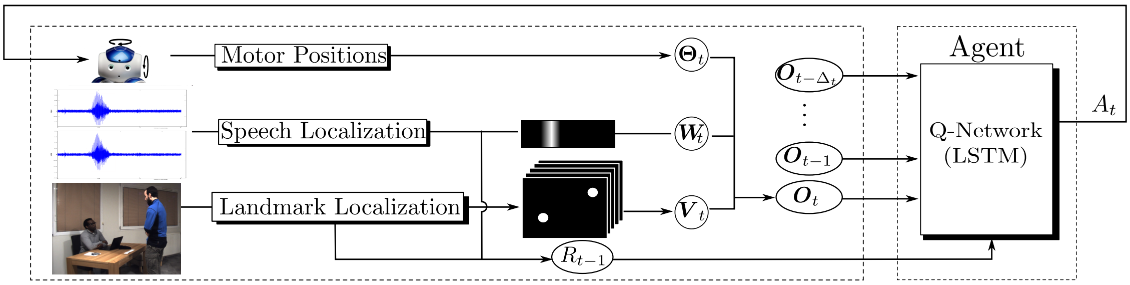

Figure 6. Overview of the proposed deep RL method for controlling the gaze of a robot. At each time index t, audio and visual data are represented as binary maps which, together with motor positions, form the set of observations $\mathbf{O}_t$*. A motor action* A_t *(rotate the head left, right, up, 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 [45], [37] for further details.*

a robot and users. RL avoids systematic recourse to annotated datasets for training. In [45], [37] we proposed the use of a simulated environment for pre-training the RL parameters, thus avoiding spending hours of tedious interaction.
Websites:
https://team.inria.fr/perception/research/deep-rl-for-gaze-control/,
https://team.inria.fr/perception/research/mot-servoing/.

6.9. Generation of Diverse Behavioral Data

We target the automatic generation of visual data depicting human behavior, and in particular how to design a method able to learn the generation of *data diversity*. In particular, we focus on smiles, because each smile is unique: one person surely smiles in different ways (e.g. closing/opening the eyes or mouth). We wonder if given one input image of a neutral face, we can generate multiple smile videos with distinctive characteristics. To tackle this one-to-many video generation problem, we propose a novel deep learning architecture named Conditional MultiMode Network (CMM-Net). To better encode the dynamics of facial expressions, CMM-Net explicitly exploits facial landmarks for generating smile sequences. Specifically, a variational auto-encoder is used to learn a facial landmark embedding. This single embedding is then exploited by a conditional recurrent network which generates a landmark embedding sequence conditioned on a specific expression (e.g. spontaneous smile), implemented as a Conditional LSTM. Next, the generated landmark embeddings are fed into a multi-mode recurrent landmark generator, producing a set of landmark sequences still associated to the given smile class but clearly distinct from each other, we call that a Multi-Mode LSTM. Finally, these landmark sequences are translated into face videos. Our experimental results, see Figure 7, demonstrate the effectiveness of our CMM-Net in generating realistic videos of multiple smile expressions [52].

6.10. 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

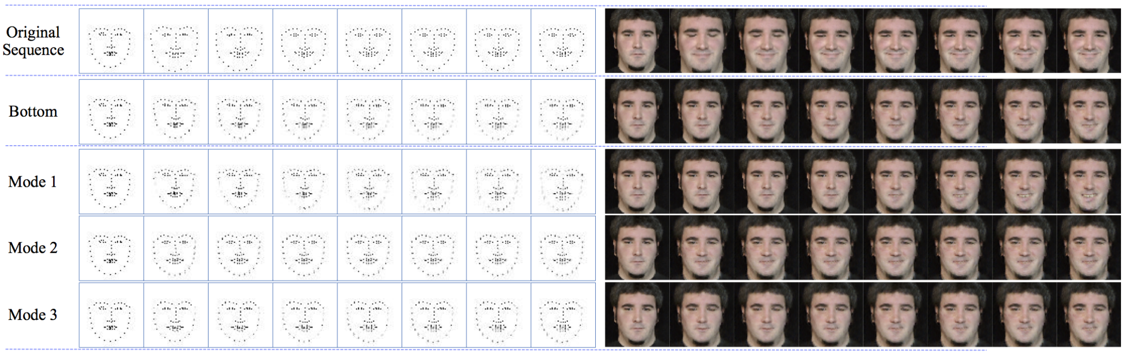

Figure 7. Multi-mode generation example with a sequence: landmarks (left) and associated face images (right) after the landmark-to-image decoding step based on Variational Auto-Encoders. The rows correspond to the original sequence (first), output of the Conditional LSTM (second), and output of the Multi-Mode LSTM (last three rows).

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 [35].
Website: https://team.inria.fr/perception/research/jrmpc/

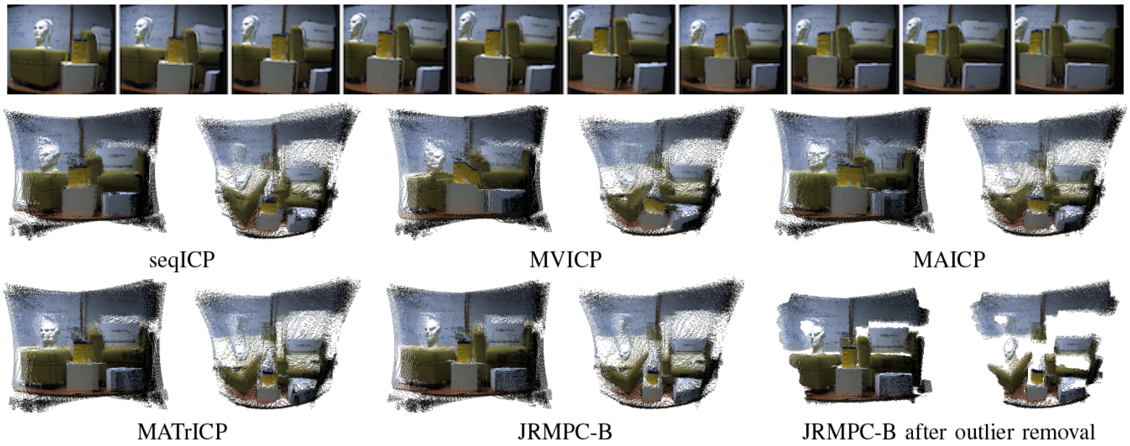

Figure 8. Integrated point clouds from the joint registration of 10 TOF images that record a static scene (EXBI data-set). Top: color images that roughly show the scene content of each range image (occlusions due to cameras baseline may cause texture artefacts). Bottom: front-view and top-view of integrated sets after joint registration. The results obtained with the proposed method (JRMPC-B) are compared with several other methods.

7. Partnerships and Cooperations

7.1. European Initiatives

7.1.1. 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/

7.1.2. VHIALab

Title: Vision and Hearing in Action Laboratory

EU framework: H2020

Type: ERC Proof of Concept

Duration: February 2018 - January 2019

Coordinator: Inria

Inria contact: Radu Horaud

The objective of VHIALab is the development and commercialization of software packages enabling a robot companion to easily and naturally interact with people. The methodologies developed in ERC VHIA propose state of the art solutions to human-robot interaction (HRI) problems in a general setting and based on audio-visual information. The ambitious goal of VHIALab will be to build software packages based on VHIA, thus opening the door to commercially available multi-party multi-modal human-robot interaction. The methodology investigated in VHIA may well be viewed as a generalization of existing single-user spoken dialog systems. VHIA enables a robot (i) to detect and to locate speaking persons, (ii) to track several persons over time, (iii) to recognize their behavior, and (iv) to extract the speech signal of each person for subsequent speech recognition and face-to-face dialog. These methods will be turned into software packages compatible with a large variety of companion robots. VHIALab will add a strong valorization potential to VHIA by addressing emerging and new market sectors. Industrial collaborations set up in VHIA will be strengthened.

7.2. International Research Visitors

7.2.1. Visits of International Scientists

- Professor Sharon Gannot, Bar Ilan University, Tel Aviv, Israel.
- Professor Tomislav Pribanic, University of Zagreb, Zagreb, Croatia.
- Doctor Christine Evers, Imperial College, London, United Kingdom.

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 organized several workshops in conjunction with IEEE CVPR'18, ECCV'18, and ACM Multimedia'18.

8.1.1.2. Reviewer

Xavier Alameda-PIneda was a reviewer for IEEE CVPR'18, NIPS'18, IEEE ICASSP'18, ACM Multimedia'18 and IEEE ICRA'18.

8.1.2. Journal

8.1.2.1. Member of the Editorial Boards

Radu Horaud is associated editor for the International Journal of Computer Vision and for the IEEE Robotics and Automation Letters.

8.1.2.2. Guest Editor

Xavier Alameda-Pineda was co-guest editor of a special issue of the ACM Transactions on Multimedia Computing Communications and Applications on “Multimodal Understanding of Social, Affective, and Subjective Attributes".

8.1.3. Invited Talks

- Radu Horaud gave an invited talk at the Multimodal Machine Perception Workshop, Google, San Francisco, and at SRI International, Menlo Park, USA, on "Audio-Visual Machine Perception for Human-Robot Interaction".
- Xavier Alameda-Pineda was invited to give a seminar at the University in May 2018 on “Audio-Visual Multiple Speaker Tracking with Robotic Platforms", and
- Xavier Alameda-Pineda gave an invited talk at the SOUND Workshop at Bar-Ilan University, Israel, December 2018 on “Multi-modal Automatic Detection of Social Attractors in Crowded Meetings".

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

- Laurent Girin is professor at Grenoble National Polytechnic Institute (G-INP) where he teaches signal processing and machine learning on the basis of a full professor service (192 hours/year)
- Xavier Alameda-Pineda is involved with the M2 course of the MSIAM Masters Program: “Fundamentals of probabilistic data mining, modeling seminars and projects”, for the practical sessions. Xavier is also preparing a doctoral course on “Learning with Multi-Modal data for Scene Understanding and Human-Robot Interaction” to be taught in spring 2019.

8.2.2. Supervision

- Radu Horaud has supervised the following PhD students: Israel Dejene-Gebru [32], Stéphane Lathuilière [33], Benoît Massé [34], Yutong Ban, Guillaume Delorme and Sylvain Guy.
- Xavier Alameda-PIneda has co-supervised Israel Dejene-Gebru, Yutong Ban, Guillaume Delorme and has supervised Yihong Xu.

8.2.3. Juries

Xavier Alameda-Pineda was reviewer and examiner of the PhD dissertations of Wei Wang (now post-doctoral fellow at EPFL) and of Dr. Dan Xu (now post-doctoral fellow at U. Oxford), both at University of Trento, Italy.

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] 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] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

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

[17] M. Hansard, R. Horaud. *A Differential Model of the Complex Cell*, in "Neural Computation", September 2011, vol. 23, n^{o} 9, pp. 2324-2357 [*DOI :* 10.1162/NECO_a_00163], http://hal.inria.fr/inria-00590266

[18] 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

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

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

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

[22] V. Khalidov, F. Forbes, R. Horaud. *Conjugate Mixture Models for Clustering Multimodal Data*, in "Neural Computation", February 2011, vol. 23, n^o 2, pp. 517-557 [*DOI :* 10.1162/NECO_a_00074], http://hal.inria.fr/inria-00590267

[23] D. Knossow, R. Ronfard, R. Horaud. *Human Motion Tracking with a Kinematic Parameterization of Extremal Contours*, in "International Journal of Computer Vision", September 2008, vol. 79, n^o 3, pp. 247-269 [*DOI :* 10.1007/s11263-007-0116-2], http://hal.inria.fr/inria-00590247

[24] 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^o 8, pp. 1408-1423 [*DOI :* 10.1109/TASLP.2016.2554286], https://hal.inria.fr/hal-01301762

[25] 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

[26] 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^o 11, pp. 2171 - 2186 [*DOI :* 10.1109/TASLP.2016.2598319], https://hal.inria.fr/hal-01349691

[27] 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^o 10, pp. 1997 - 2012, 16 pages, 4 figures, 4 tables [*DOI :* 10.1109/TASLP.2017.2740001], https://hal.inria.fr/hal-01413417

[28] M. Sapienza, M. Hansard, R. Horaud. *Real-time Visuomotor Update of an Active Binocular Head*, in "Autonomous Robots", January 2013, vol. 34, n^o 1, pp. 33-45 [*DOI :* 10.1007/s10514-012-9311-2], http://hal.inria.fr/hal-00768615

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

[30] A. Zaharescu, E. Boyer, R. Horaud. *Keypoints and Local Descriptors of Scalar Functions on 2D Manifolds*, in "International Journal of Computer Vision", October 2012, vol. 100, n^o 1, pp. 78-98 [*DOI :* 10.1007/s11263-012-0528-5], http://hal.inria.fr/hal-00699620

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[32] I. D. GEBRU. *Audio-Visual Analysis In the Framework of Humans Interacting with Robots*, Université Grenoble Alpes, April 2018, https://hal.inria.fr/tel-01774233

[33] S. LATHUILIÈRE. *Deep Regression Models and Computer Vision Applications for Multiperson Human-Robot Interaction*, Université Grenoble Alpes, May 2018, https://tel.archives-ouvertes.fr/tel-01801807

[34] B. MASSÉ. *Gaze Direction in the context of Social Human-Robot Interaction*, Université Grenoble - Alpes, October 2018, https://hal.inria.fr/tel-01936821

Articles in International Peer-Reviewed Journals

[35] 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^{o} 6, pp. 1397 - 1410, https://arxiv.org/abs/1609.01466 [*DOI :* 10.1109/TPAMI.2017.2717829], https://hal.inria.fr/hal-01413414

[36] 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^{o} 5, pp. 1086 - 1099, https://arxiv.org/abs/1603.09725 [*DOI :* 10.1109/TPAMI.2017.2648793], https://hal.inria.fr/hal-01413403

[37] 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", May 2018, https://arxiv.org/abs/1711.06834 [*DOI :* 10.1016/J.PATREC.2018.05.023], https://hal.inria.fr/hal-01643775

[38] X. LI, S. GANNOT, L. GIRIN, R. HORAUD. *Multichannel Identification and Nonnegative Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", May 2018, vol. 26, n^{o} 10, pp. 1755-1768, https://arxiv.org/abs/1711.07911 [*DOI :* 10.1109/TASLP.2018.2839362], https://hal.inria.fr/hal-01645749

[39] 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", January 2019, https://arxiv.org/abs/1711.07911 [*DOI :* 10.1109/TASLP.2019.2892412], https://hal.inria.fr/hal-01799809

[40] X. LI, L. GIRIN, R. HORAUD. *Expectation-Maximization for Speech Source Separation using Convolutive Transfer Function*, in "CAAI Transactions on Intelligent Technologies", January 2019 [*DOI :* 10.1049/TRIT.2018.1061], https://hal.inria.fr/hal-01982250

[41] R. T. MARRIOTT, A. PASHEVICH, R. HORAUD. *Plane-extraction from depth-data using a Gaussian mixture regression model*, in "Pattern Recognition Letters", July 2018, vol. 110, pp. 44-50, https://arxiv.org/abs/1710.01925 - 2 figures, 1 table [*DOI :* 10.1016/J.PATREC.2018.03.024], https://hal.inria.fr/hal-01663984

[42] 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^{o} 11, pp. 2711 - 2724, https://arxiv.org/abs/1703.04727 [*DOI :* 10.1109/TPAMI.2017.2782819], https://hal.inria.fr/hal-01511414

[43] D. XU, X. ALAMEDA-PINEDA, J. SONG, E. RICCI, N. SEBE. *Cross-Paced Representation Learning with Partial Curricula for Sketch-based Image Retrieval*, in "IEEE Transactions on Image Processing", September 2018, vol. 27, n^{o} 9, pp. 4410-4421 [*DOI :* 10.1109/TIP.2018.2837381], https://hal.inria.fr/hal-01803694

International Conferences with Proceedings

[44] Y. BAN, X. LI, X. ALAMEDA-PINEDA, L. GIRIN, R. HORAUD. *Accounting for Room Acoustics in Audio-Visual Multi-Speaker Tracking*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Alberta, Canada, IEEE, April 2018, pp. 6553-6557 [*DOI :* 10.1109/ICASSP.2018.8462100], https://hal.inria.fr/hal-01718114

[45] S. LATHUILIÈRE, B. MASSÉ, P. MESEJO, R. HORAUD. *Deep Reinforcement Learning for Audio-Visual Gaze Control*, in "IROS 2018 - IEEE/RSJ International Conference on Intelligent Robots and Systems", Madrid, Spain, October 2018, pp. 1-8 [*DOI :* 10.1109/IROS.2018.8594327], https://hal.inria.fr/hal-01851738

[46] S. LATHUILIÈRE, P. MESEJO, X. ALAMEDA-PINEDA, R. HORAUD. *DeepGUM: Learning Deep Robust Regression with a Gaussian-Uniform Mixture Model*, in "ECCV 2018 - European Conference on Computer Vision", Munich, Germany, September 2018, pp. 1-16, https://hal.inria.fr/hal-01851511

[47] S. LEGLAIVE, L. GIRIN, R. HORAUD. *A variance modeling framework based on variational autoencoders for speech enhancement*, in "MSLP 2018 - IEEE International Workshop on Machine Learning for Signal Processing", Aalborg, Denmark, IEEE, September 2018, pp. 1-6 [*DOI :* 10.1109/MLSP.2018.8516711], https://hal.inria.fr/hal-01832826

[48] X. LI, Y. BAN, L. GIRIN, X. ALAMEDA-PINEDA, R. HORAUD. *A Cascaded Multiple-Speaker Localization and Tracking System*, in "Proceedings of the LOCATA Challenge Workshop - a satellite event of IWAENC 2018", Tokyo, Japan, September 2018, pp. 1-5, https://hal.inria.fr/hal-01957137

[49] X. LI, S. GANNOT, L. GIRIN, R. HORAUD. *Multisource MINT Using the Convolutive Transfer Function*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Alberta, Canada, IEEE, April 2018, pp. 756-760 [*DOI :* 10.1109/ICASSP.2018.8462607], https://hal.inria.fr/hal-01718106

[50] X. LI, B. MOURGUE, L. GIRIN, S. GANNOT, R. HORAUD. *Online Localization of Multiple Moving Speakers in Reverberant Environments*, in "10th IEEE Workshop on Sensor Array and Multichannel Signal Processing (SAM 2018)", Sheffield, United Kingdom, IEEE, July 2018, pp. 405-409 [*DOI :* 10.1109/SAM.2018.8448423], https://hal.inria.fr/hal-01795462

[51] A. SIAROHIN, E. SANGINETO, S. LATHUILIÈRE, N. SEBE. *Deformable GANs for Pose-based Human Image Generation*, in "IEEE Conference on Computer Vision and Pattern Recognition", Salt Lake City, United States, June 2018, pp. 3408-3416, https://arxiv.org/abs/1801.00055 , https://hal.archives-ouvertes.fr/hal-01761539

[52] W. WANG, X. ALAMEDA-PINEDA, D. XU, P. FUA, E. RICCI, N. SEBE. *Every Smile is Unique: Landmark-Guided Diverse Smile Generation*, in "IEEE Conference on Computer Vision and Pattern Recognition", Salk

Lake City, United States, June 2018, pp. 7083-7092, https://arxiv.org/abs/1802.01873 , https://hal.inria.fr/hal-01759335

Scientific Books (or Scientific Book chapters)

[53] 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

[54] L. Girin, S. Gannot, X. Li. *Audio source separation into the wild*, in "Multimodal Behavior Analysis in the Wild", Computer Vision and Pattern Recognition, Academic Press (Elsevier), November 2018, pp. 53-78 [*DOI :* 10.1016/B978-0-12-814601-9.00022-5], https://hal.inria.fr/hal-01943375

Other Publications

[55] Y. Ban, X. Alameda-Pineda, C. Evers, R. Horaud. *Tracking Multiple Audio Sources with the von Mises Distribution and Variational EM*, December 2018, https://arxiv.org/abs/1812.08246 - Paper submitted to IEEE Signal Processing Letters, https://hal.inria.fr/hal-01969050

[56] Y. Ban, X. Alameda-Pineda, L. Girin, R. Horaud. *Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers*, December 2018, https://arxiv.org/abs/1809.10961 - Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, https://hal.inria.fr/hal-01950866

[57] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, R. Horaud. *A Comprehensive Analysis of Deep Regression*, March 2018, https://arxiv.org/abs/1803.08450 - Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, https://hal.inria.fr/hal-01754839

[58] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, R. Horaud. *Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environment*, July 2018, Submitted to Journal on Selected Topics in Signal Processing, https://hal.inria.fr/hal-01851985

[59] X. Li, L. Girin, S. Gannot, R. Horaud. *Multichannel Online Dereverberation based on Spectral Magnitude Inverse Filtering*, December 2018, https://arxiv.org/abs/1812.08471 - Paper submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing, https://hal.inria.fr/hal-01969041

[60] A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci, N. Sebe. *Increasing Image Memorability with Neural Style Transfer*, August 2018, working paper or preprint, https://hal.inria.fr/hal-01858389