Activity Report 2016

Project-Team PERCEPTION

Interpretation and Modeling of Images and Sounds

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team PERCEPTION

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

Keywords:

Computer Science and Digital Science:

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

Other Research Topics and Application Domains:

- 5.6. - Robotic systems

1. Members

Research Scientists

Radu Horaud [Team leader, Inria, Senior Researcher, HDR]
Xavier Alameda-Pineda [Inria, Researcher, from Dec 2016]
Georgios Evangelidis [Inria, Starting Research Position, until Jun 2016, granted by ERC VHIA]
Sileye Ba [Inria, Starting Research Position, until Jun 2016, granted by ERC VHIA]
Xiaofei Li [Inria, Starting Research Position, granted by ERC VHIA]
Pablo Mesejo Santiago [Inria, Starting Research Position, from Sep 2016, granted by ERC VHIA]

Faculty Member

Laurent Girin [Grenoble INP, Professor, HDR]

Engineers

Bastien Mourgue [Inria, from Dec 2016, granted by ERC VHIA]
Quentin Pelorson [Inria, until May 2016, granted by FP7 EARS]
Guillaume Sarrazin [Inria, from Jun 2016, granted by ERC VHIA]
Fabien Badeig [Inria, until Oct 2016, granted by ERC VHIA]

PhD Students

Yutong Ban [Inria, granted by ERC VHIA]
Israel Dejene Gebru [Inria, granted by Inria]
Vincent Drouard [Inria, granted by ERC VHIA]
Dionyssos Kounades [Inria, granted by FP7 EARS]
Stephane Lathuiliere [Inria, granted by ERC VHIA]
Benoit Masse [Inria, granted by ERC VHIA]

Visiting Scientists

Yuval Dorfan [Inria, from Jun 2016 until Jul 2016]
Sharon Gannot [Inria, until Oct 2016]
Rafael Munoz Salinas [Inria, from Jun 2016 until Aug 2016]

Administrative Assistant

Nathalie Gillot [Inria]

Others

Remi Juge [Inria, Intern, from Jul 2016]
Richard Thomas Marriott [Inria, Intern, from Feb until Sep 2016]
Alexander Pashevich [Inria, Intern, from Feb 2016 until Jun 2016]

2. Overall Objectives

2.1. Overall Objectives

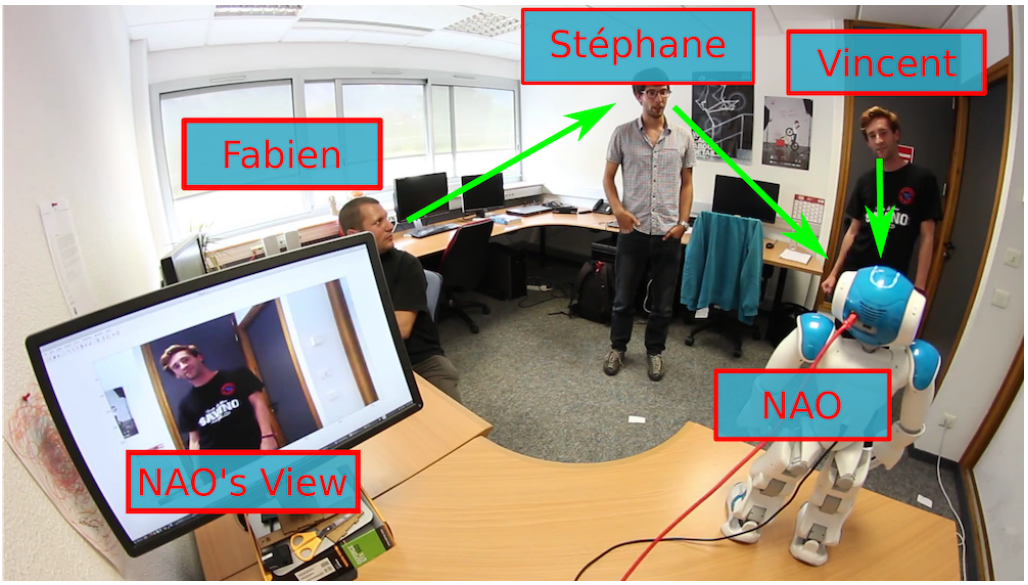

Figure 1. This figure illustrates the audio-visual multi-party human-robot interaction paradigm that the PERCEPTION team has developed in the recent past [18], [2], [25]. There are inter-person as well as person-robot interactions that must be properly detected and analyzed over time. This includes multiple-person tracking [24], person detection and head-pose estimation [42], sound-source separation and localization [5], [1], [28], [29], [44], and speaker diarization [26]. These developments are 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.

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 [18], [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 [6]. 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 [5]. 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 [12], [20]. 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 [13]. 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 [8].

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

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 [23]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [21]. 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 [22]. 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 [14], [10],[9]. 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 [8].

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 [19]. 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 [17], [16]. 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 [4]. 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

- **The three-year FP7 STREP project *Embodied Audition for Robots* successfully terminated in December 2016**. The project has addressed the problem of robot hearing, more precisely, the analysis of audio signals in complex environments: reverberant rooms, multiple users, and background noise. In collaboration with the project partners, PERCEPTION contributed to audio-source localization, audio-source separation, audio-visual alignment, and audio-visual disambiguation. The humanoid robot NAO has been used as a robotic platform and a new head (hardware and software) was developed: a stereoscopic camera pair, a spherical microphone array, and the associated synchronization, signal and image processing software modules.
- This year, PERCEPTION started a one year collaboration with the **Digital Media and Communications R&D Center, Samsung Electronics** (Seoul, Korea). The topic of this collaboration is *multi-modal speaker localization and tracking* (a central topic of the team) and is part of a strategic partnership between Inria and Samsung Electronics.

4.1.1. Awards

- **Antoine Deleforge** (former PhD student, PANAMA team), **Florence Forbes** (MISTIS team) and **Radu Horaud** received the **2016 Award for Outstanding Contributions in Neural Systems** for their paper: “Acoustic Space Learning for Sound-source Separation and Localization on Binaural Manifolds," International Journal of Neural Systems, volume 25, number 1, 2015. The Award for Outstanding Contributions in Neural Systems established by World Scientific Publishing Co. in 2010, is awarded annually to the most innovative paper published in the previous volume/year of the International Journal of Neural Systems.
- **Xavier Alameda-Pineda** and his co-authors from the University of Trento received the **Intel Best Scientific Paper Award** (Track: Image, Speech, Signal and Video Processing) for their paper “Multi-Paced Dictionary Learning for Cross-Domain Retrieval and Recognition" presented at the 23rd IEEE International Conference on Pattern Recognition, Cancun, Mexico, December 2016 .

BEST PAPER AWARD:

[41]
D. XU, J. SONG, X. ALAMEDA-PINEDA, E. RICCI, N. SEBE. *Multi-Paced Dictionary Learning for Cross-Domain Retrieval and Recognition*, in "IEEE International Conference on Pattern Recognition", Cancun, Mexico, December 2016, https://hal.inria.fr/hal-01416419

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, Radu Horaud and Manuel Yguel
- 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: Patrice Horaud, Pierre Arquier, Quentin Pelorson, Michel Amat, Miles Hansard, Georgios Evangelidis, Soraya Arias, Radu Horaud, Richard Broadbridge and Clement Menier
- 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: Quentin Pelorson, Fabien Badeig 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: Jordi Sanchez-Riera, Soraya Arias, Jan Cech and Radu Horaud
- Contact: Soraya Arias
- URL: https://code.humavips.eu/projects/stereomatch

5.5. Platforms

5.5.1. Audio-Visual Head Popeye+

In 2016 we upgraded our audio-visual platform, from Popeye to Popeye+. Popeye+ has two high-definitions 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.
Website:
https://team.inria.fr/perception/projects/popeye/
https://team.inria.fr/perception/projects/popeye-plus/
https://team.inria.fr/perception/avtrack1/.

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 a social robot 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.

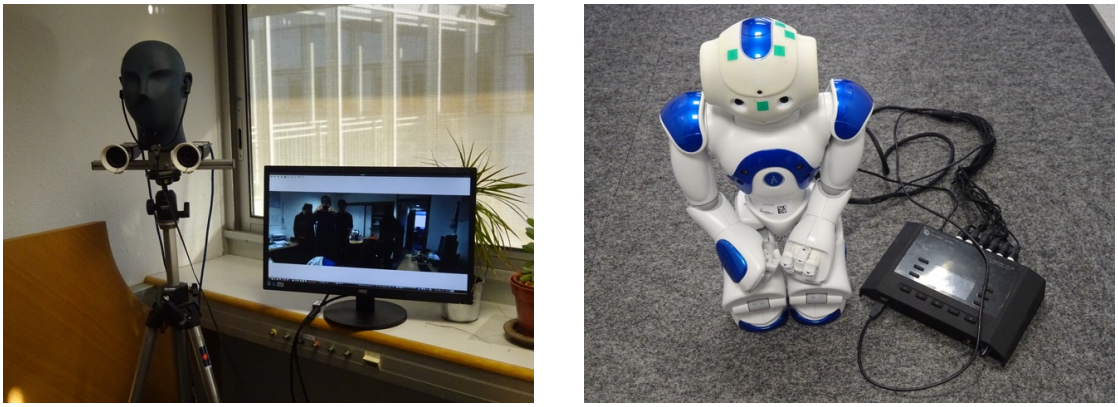

Figure 2. The Popeye+ audio-visual platform (left) delivers high-quality, high-resolution and wide-angle images at 30FPS. The NAO prototype used by PERCEPTION in the EARS STREP project has a twelve-channel spherical microphone array synchronized with a stereo camera pair.

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

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

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

6. New Results

6.1. Audio-Source Localization

In previous years we have developed several *supervised* sound-source localization algorithms. The general principle of these algorithms was based on the learning of a mapping (regression) between binaural feature vectors and source locations [5], [7]. While fixed-length wide-spectrum sounds (white noise) are used for training to reliably estimate the model parameters, we show that the testing (localization) can be extended to variable-length sparse-spectrum sounds (such as speech), thus enabling a wide range of realistic applications. Indeed, we demonstrate that the method can be used for audio-visual fusion, namely to map speech signals onto images and hence to spatially align the audio and visual modalities, thus enabling to discriminate between speaking and non-speaking faces. We released a novel corpus of real-room recordings that allow quantitative evaluation of the co-localization method in the presence of one or two sound sources. Experiments demonstrate increased accuracy and speed relative to several state-of-the-art methods. During the period 2015-2016 we extended this method to an arbitrary number of microphones based on the *relative transfer function – RTF* (between any channel and a reference channel). Then we extended this work and developed a novel transfer function that contains the direct path between the source and the microphone array, namely the *direct-path relative transfer function* [29], [36].

Websites:

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

6.2. Audio-Source Separation

We address the problem of separating audio sources from time-varying convolutive mixtures. We proposed an unsupervised probabilistic framework based on the local complex-Gaussian model combined with non-negative matrix factorization [33], [28]. The time-varying mixing filters are modeled by a continuous temporal stochastic process. This model extends the case of static filters which corresponds to static audio sources. While static filters can be learnt in advance, e.g. [5], time-varying filters cannot and therefore the problem is more complex. We present a variational expectation-maximization (VEM) algorithm that employs a Kalman smoother to estimate the time-varying mixing matrix, and that jointly estimates the source parameters. The sound sources are then separated by Wiener filters constructed with the estimators provided by the VEM algorithm. Extensive experiments on simulated data show that the proposed method outperforms a block-wise version of a state-of-the-art baseline method. This work is part of the PhD topic of Dionyssos Kounades Bastian and is conducted in collaboration with Sharon Gannot (Bar Ilan University) and Xavier Alameda Pineda (University of Trento). Our journal paper [28] is an extended version of a paper presented at IEEE WASPAA in 2015 which received the best student paper award.

Website:
https://team.inria.fr/perception/research/vemove/
https://team.inria.fr/perception/research/nmfig/

6.3. Single-Channel Audio Processing

While most of our audio scene analysis work involves microphone arrays, it is important to develop single-channel (one microphone) signal processing methods as well. In particular, it is important to detect speech signal (or voice) in the presence of various types of noise (stationary or non-stationary). In this context, we developed the following methods [39], [37]:

- Statistical likelihood ratio test is a widely used voice activity detection (VAD) method, in which the likelihood ratio of the current temporal frame is compared with a threshold. A fixed threshold is always used, but this is not suitable for various types of noise. In this work, an adaptive threshold is proposed as a function of the local statistics of the likelihood ratio. This threshold represents the upper bound of the likelihood ratio for the non-speech frames, whereas it remains generally lower than the likelihood ratio for the speech frames. As a result, a high non-speech hit rate can be achieved, while maintaining speech hit rate as large as possible.
- Estimating the noise power spectral density (PSD) is essential for single channel speech enhancement algorithms. We propose a noise PSD estimation approach based on regional statistics which consist of four features representing the statistics of the past and present periodograms in a short-time period. We show that these features are efficient in characterizing the statistical difference between noise PSD and noisy-speech PSD. We therefore propose to use these features for estimating the speech presence probability (SPP). The noise PSD is recursively estimated by averaging past spectral power values with a time-varying smoothing parameter controlled by the SPP. The proposed method exhibits good tracking capability for non-stationary noise, even for abruptly increasing noise level.

Website:
https://team.inria.fr/perception/research/noise-psd/

6.4. Tracking Multiple Persons

Object tracking is an ubiquitous problem in computer vision with many applications in human-machine and human-robot interaction, augmented reality, driving assistance, surveillance, etc. Although thoroughly investigated, tracking multiple persons remains a challenging and an open problem. In this work, an online variational Bayesian model for multiple-person tracking is proposed. This yields a variational expectation-maximization (VEM) algorithm. The computational efficiency of the proposed method is made possible thanks to closed-form expressions for both the posterior distributions of the latent variables and for the estimation of the model parameters. A stochastic process that handles person birth and person death enables the tracker to handle a varying number of persons over long periods of time [24], [30].

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

6.5. Audio-Visual Speaker Detection, Localization, and Diarization

Any multi-party conversation system benefits from speaker diarization, that is, the assignment of speech signals among the participants. More generally, in HRI and CHI scenarios it is important to recognize the speaker over time. We propose to address speaker detection, localization and diarization using both audio and visual data. We cast the diarization problem into a tracking formulation whereby the active speaker is detected and tracked over time. A probabilistic tracker exploits the spatial coincidence of visual and auditory observations and infers a single latent variable which represents the identity of the active speaker. Visual and auditory observations are fused using our recently developed weighted-data mixture model [25], 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 data-sets that are available from recent contributions which are used as baselines for comparison [26].

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

6.6. Head Pose Estimation and Tracking

Head pose estimation is an important task, because it provides information about cognitive interactions that are likely to occur. Estimating the head pose is intimately linked to face detection. We addressed the problem of head pose estimation with three degrees of freedom (pitch, yaw, roll) from a single image and in the presence of face detection errors. Pose estimation is formulated as a high-dimensional to low-dimensional mixture of linear regression problem [6]. We propose a method that maps HOG-based descriptors, extracted from face bounding boxes, to corresponding head poses. To account for errors in the observed bounding-box position, we learn regression parameters such that a HOG descriptor is mapped onto the union of a head pose and an offset, such that the latter optimally shifts the bounding box towards the actual position of the face in the image. The performance of the proposed method is assessed on publicly available datasets. The experiments that we carried out show that a relatively small number of locally-linear regression functions is sufficient to deal with the non-linear mapping problem at hand. Comparisons with state-of-the-art methods show that our method outperforms several other techniques [42]. This work is part of the PhD of Vincent Drouard and it received the best student paper award (second place) at the IEEE ICIP'15. Currently we investigate a temporal extension of this model.

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

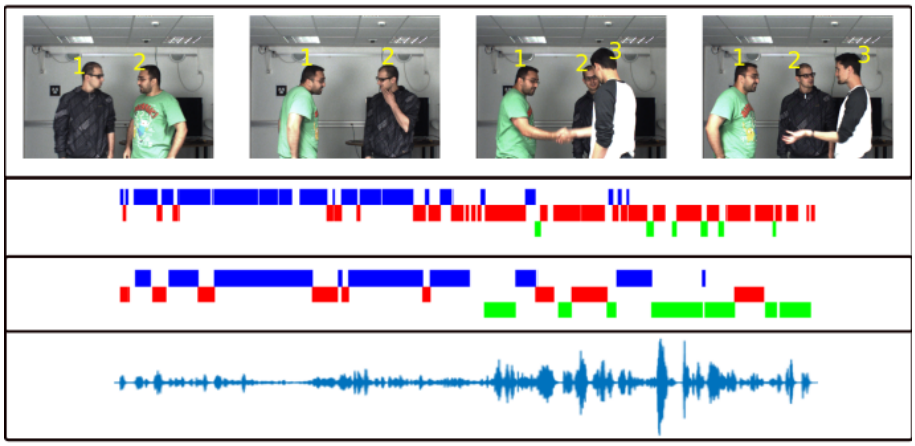

Figure 3. This figure illustrates the audiovisual tracking and diarization method that we have recently developed. First row: A number is associated with each tracked person. Second row: diarization result. Third row: the ground truth diarization. Fourth row: acoustic signal recorded by one of the two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

6.7. Estimation of Eye Gaze and of Visual Focus of Attention

We address the problem of estimating the visual focus of attention (VFOA), e.g. who is looking at whom? This is of particular interest in human-robot interactive scenarios, e.g. when the task requires to identify targets of interest and to track them over time. We make the following contributions. We propose a Bayesian temporal model that links VFOA to eye-gaze direction and to head orientation. Model inference is cast into a switching Kalman filter formulation, which makes it tractable. The model parameters are estimated via training based on manual annotations. The method is tested and benchmarked using a publicly available dataset. We show that both eye-gaze and VFOA of several persons can be reliably and simultaneously estimated and tracked over time from observed head poses as well as from people and object locations [40].

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

6.8. High-Resolution Scene Reconstruction

We addressed the problem of range-stereo fusion for the construction of high-resolution depth maps. In particular, we combine time-of-flight (low resolution) depth [27] data with high-resolution stereo data, in a maximum a posteriori (MAP) formulation. Unlike existing schemes that build on MRF optimizers, we infer the disparity map from a series of local energy minimization problems that are solved hierarchically, by growing sparse initial disparities obtained from the depth data. The accuracy of the method is not compromised, owing to three properties of the data-term in the energy function. Firstly, it incorporates a new correlation function that is capable of providing refined correlations and disparities, via sub-pixel correction. Secondly, the correlation scores rely on an adaptive cost aggregation step, based on the depth data. Thirdly, the stereo and depth likelihoods are adaptively fused, based on the scene texture and camera geometry. These properties lead to a more selective growing process which, unlike previous seed-growing methods, avoids the tendency to propagate incorrect disparities. The proposed method gives rise to an intrinsically efficient algorithm, which runs at 3FPS on 2.0MP images on a standard desktop computer. The strong performance of the new method is established both by quantitative comparisons with state-of-the-art methods, and by qualitative comparisons using real depth-stereo data-sets [8]. This work is funded by the ANR project MIXCAM.

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

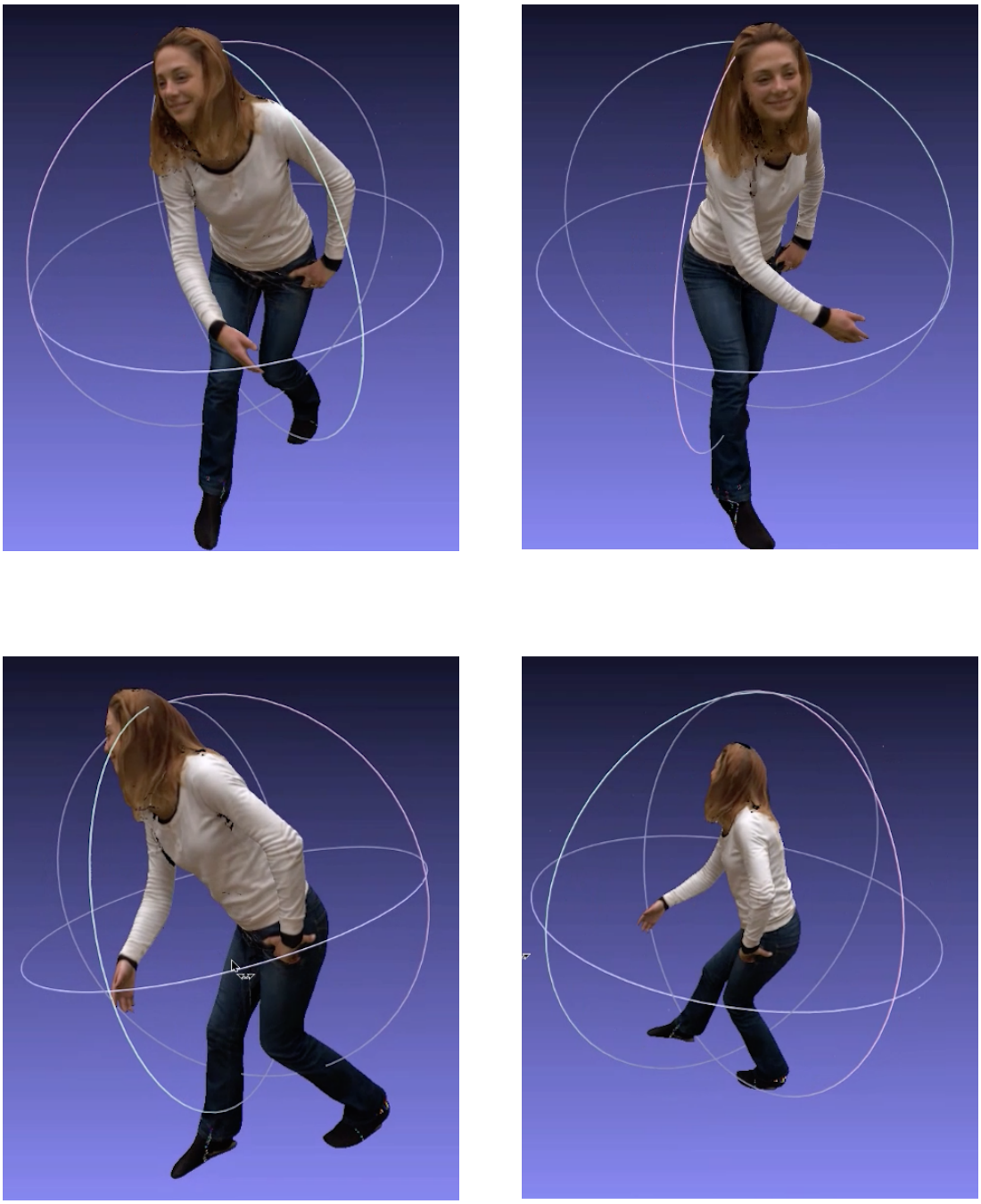

Figure 4. Four views of a 3D person reconstructed with our algorithm. In this example we used a large number of high-resolution cameras and the rendering was performed by the software of 4D View Solutions.

6.9. Registration of Multiple Point Sets

We have also addressed the rigid registration problem of multiple 3D point sets. While the vast majority of state-of-the-art techniques build on pairwise registration, we proposed a generative model that explains jointly registered multiple sets: back-transformed points are considered realizations of a single Gaussian mixture model (GMM) whose means play the role of the (unknown) scene points. Under this assumption, the joint registration problem is cast into a probabilistic clustering framework. We formally derive an expectation-maximization procedure that robustly estimates both the GMM parameters and the rigid transformations that map each individual cloud onto an under-construction reference set, that is, the GMM means. GMM variances carry rich information as well, thus leading to a noise- and outlier-free scene model as a by-product. A second version of the algorithm is also proposed whereby newly captured sets can be registered online. A thorough discussion and validation on challenging data-sets against several state-of-the-art methods confirm the potential of the proposed model for jointly registering real depth data [43].

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

- In December, PERCEPTION started a one year collaboration with the **Digital Media and Communications R&D Center, Samsung Electronics** (Seoul, Korea). The topic of this collaboration is *multi-modal speaker localization and tracking* (a central topic of the team) and is part of a strategic partnership between Inria and Samsung Electronics.
- Over the past six years we have collaborated with Aldebaran Robotics (now SoftBank). This collaboration was part of two EU STREP projects, HUMAVIPS (2010-2012) and EARS (2014-2016). This enabled our team to establish strong connections with SoftBank, to design a stereoscopic camera head and to jointly develop several demonstrators using three different generations of the NAO robot.
  Website: https://team.inria.fr/perception/nao/
- In 2015 we started a collaboration with Xerox Research Center India (XRCI), Bangalore. This three-year collaboration (2015-2017) is funded by a grant awarded by the **Xerox Foundation University Affairs Committee (UAC)** and the topic of the project is *Advanced and Scalable Graph Signal Processing Techniques*. The work is done in collaboration with EPI MISTIS and our Indian collaborators are Arijit Biswas and Anirban Mondal.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

8.1.1.1. MIXCAM

Type: ANR BLANC

Duration: March 2014 - February 2016

Coordinator: Radu Horaud

Partners: 4D View Solutions SAS

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

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

8.2.1.1. EARS

Title: Embodied Audition for RobotS

Program: FP7

Duration: January 2014 - December 2016

Coordinator: Friedrich Alexander Universität Erlangen-Nünberg

Partners:

Aldebaran Robotics (France)

Ben-Gurion University of the Negev (Israel)

Friedrich Alexander Universität Erlangen-Nünberg (Germany)

Imperial College of Science, Technology and Medicine (United Kingdom)

Humboldt-Universitat Zu Berlin (Germany)

Inria contact: Radu Horaud

The success of future natural intuitive human-robot interaction (HRI) will critically depend on how responsive the robot will be to all forms of human expressions and how well it will be aware of its environment. With acoustic signals distinctively characterizing physical environments and speech being the most effective means of communication among humans, truly humanoid robots must be able to fully extract the rich auditory information from their environment and to use voice communication as much as humans do. While vision-based HRI is well developed, current limitations in robot audition do not allow for such an effective, natural acoustic human-robot communication in real-world environments, mainly because of the severe degradation of the desired acoustic signals due to noise, interference and reverberation when captured by the robot's microphones. To overcome these limitations, EARS will provide intelligent 'ears' with close-to-human auditory capabilities and use it for HRI in complex real-world environments. Novel microphone arrays and powerful signal processing algorithms shall be able to localise and track multiple sound sources of interest and to extract and recognize the desired signals. After fusion

with robot vision, embodied robot cognition will then derive HRI actions and knowledge on the entire scenario, and feed this back to the acoustic interface for further auditory scene analysis. As a prototypical application, EARS will consider a welcoming robot in a hotel lobby offering all the above challenges. Representing a large class of generic applications, this scenario is of key interest to industry and, thus, a leading European robot manufacturer will integrate EARS's results into a robot platform for the consumer market and validate it. In addition, the provision of open-source software and an advisory board with key players from the relevant robot industry should help to make EARS a turnkey project for promoting audition in the robotics world.

8.2.1.2. VHIA

Title: Vision and Hearing in Action

Program: FP7

Type: ERC

Duration: February 2014 - January 2019

Coordinator: Inria

Inria contact: Radu Horaud

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

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

- Professor Sharon Gannot, Bar Ilan University, Tel Aviv, Israel,
- Dr. Miles Hansard, Queen Mary University London, UK,
- Professor Nicu Sebe, University of Trento, Trento, Italy,
- Professor Adrian Raftery, University of Washington, Seattle, USA,
- Dr. Rafael Munoz-Salinas, University of Cordoba, Spain,
- Dr. Noam Shabatai, Ben Gourion University of the Negev, Israel.
- Dr. Christine Evers, Imperial College of Science and Medecine, UK.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

- Professor Sharon Gannot, Bar Ilan University, Tel Aviv, Israel,
- Yuval Dorfan, Bar Ilan University, Tel Aviv, Israel,
- Dr. Rafael Munoz-Salinas, University of Cordoba, Spain,
- Dr. Noam Shabatai, Ben Gourion University of the Negev, Israel.
- Dr. Christine Evers, Imperial College of Science and Medecine, UK.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Journal

9.1.1.1. Member of the Editorial Boards

Radu Horaud is a member of the following editorial boards:

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

9.1.2. Invited Talks

- Xavier Alameda-Pineda gave invited talks Polytechnic University of Catalunya (May, Barcelona), Telecom ParisTech (May), Columbia University (June, New York, USA), and Carnegie Mellon University (June, Pittsburgh, USA).
- Radu Horaud gave invited talks at the Working Group on Model Based Clustering (July, Paris), at Google Research (July, Mountain View, USA), SRI International (July, Menlo Park, USA), and Amazon (July, Seattle, USA).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Tutorial: *Multimodal Human Behaviour Analysis in the Wild: Recent Advances and Open Problems* at the IEEE ICPR'16 Conference, December 2016, 4 hours. Teachers: Xavier Alameda-Pineda, Nicu Sebe and Elisa Ricci (University of Trento).

9.2.2. Supervision

PhD in progress: Israel Dejene Gebru, October 2013, Radu Horaud and Xavier Alameda-Pineda.

PhD in progress: Dionyssos Kounades-Bastian, October 2013, Radu Horaud, Laurent Girin, and Xavier Alameda-Pineda.

PhD in progress: Vincent Drouard, October 2014, Radu Horaud and Sileye Ba.

PhD in progress: Benoit Massé, October 2014, Radu Horaud and Sileye Ba.

PhD in progress: Stéphane Lathuilière, October 2014, Radu Horaud.

PhD in progress: Yutong Ban, October 2015, Radu Horaud and Laurent Girin

10. Bibliography

Major publications by the team in recent years

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

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

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

[4] 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

[5] 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, 21p p. [*DOI :* 10.1142/S0129065714400036], https://hal.inria.fr/hal-00960796

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

Articles in International Peer-Reviewed Journals

[24] 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

[25] 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

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

[27] R. HORAUD, M. HANSARD, G. EVANGELIDIS, M. CLÉMENT. *An Overview of Depth Cameras and Range Scanners Based on Time-of-Flight Technologies*, in "Machine Vision and Applications Journal", October 2016, vol. 27, n° 7, pp. 1005–1020 [*DOI :* 10.1007/s00138-016-0784-4], https://hal.inria.fr/hal-01325045

[28] 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", April 2016, vol. 24, n° 8, pp. 1408-1423 [*DOI :* 10.1109/TASLP.2016.2554286], https://hal.inria.fr/hal-01301762

[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

International Conferences with Proceedings

[30] Y. BAN, S. BA, X. ALAMEDA-PINEDA, R. HORAUD. *Tracking Multiple Persons Based on a Variational Bayesian Model*, in "Computer Vision – ECCV 2016 Workshops", Amsterdam, Netherlands, Lecture Notes in Computer Science, Springer, October 2016, vol. Volume 9914, pp. 52-67 [*DOI :* 10.1007/978-3-319-48881-3_5], https://hal.inria.fr/hal-01359559

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

[32] L. GIRIN, R. BADEAU. *On the Use of Latent Mixing Filters in Audio Source Separation*, in "13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)", Grenoble, France, Proc. 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017), February 2017, https://hal.archives-ouvertes.fr/hal-01400965

[33] D. KOUNADES-BASTIAN, L. GIRIN, X. ALAMEDA-PINEDA, S. GANNOT, R. HORAUD. *An Inverse-Gamma Source Variance Prior with Factorized Parameterization for Audio Source Separation*, in "International Conference on Acoustics, Speech and SIgnal Processing", Shanghai, China, IEEE Signal Processing Society, March 2016, pp. 136-140 [*DOI :* 10.1109/ICASSP.2016.7471652], https://hal.inria.fr/hal-01253169

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

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

[36] 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

[37] X. LI, L. GIRIN, S. GANNOT, R. HORAUD. *Non-Stationary Noise Power Spectral Density Estimation Based on Regional Statistics*, in "International Conference on Acoustics, Speech and SIgnal Processing", Shanghai, China, IEEE Signal Processing Society, March 2016, pp. 181-185 [*DOI :* 10.1109/ICASSP.2016.7471661], https://hal.inria.fr/hal-01250892

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

[39] X. LI, R. HORAUD, L. GIRIN, S. GANNOT. *Voice Activity Detection Based on Statistical Likelihood Ratio With Adaptive Thresholding*, in "International Workshop on Acoustic Signal Enhancement", Xi'an, China, IEEE, September 2016, 5 p. [*DOI :* 10.1109/IWAENC.2016.7602911], https://hal.inria.fr/hal-01349776

[40] B. MASSÉ, S. BA, R. HORAUD. *Simultaneous Estimation of Gaze Direction and Visual Focus of Attention for Multi-Person-to-Robot Interaction*, in "International Conference on Multimedia and Expo", Seattle, United States, IEEE Signal Processing Society, July 2016, pp. 1-6 [*DOI :* 10.1109/ICME.2016.7552986], https://hal.inria.fr/hal-01301766

[41] *Best Paper*
D. XU, J. SONG, X. ALAMEDA-PINEDA, E. RICCI, N. SEBE. *Multi-Paced Dictionary Learning for Cross-Domain Retrieval and Recognition*, in "IEEE International Conference on Pattern Recognition", Cancun, Mexico, December 2016, https://hal.inria.fr/hal-01416419.

Research Reports

[42] V. DROUARD, R. HORAUD, A. DELEFORGE, S. BA, G. EVANGELIDIS. *Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regression*, Inria Grenoble - Rhone-Alpes, April 2016, 11 pages, 4 figures, 3 tables, https://hal.inria.fr/hal-01413406

[43] G. EVANGELIDIS, R. HORAUD. *Joint Registration of Multiple Point Sets*, Inria Grenoble - Rhone-Alpes, September 2016, 14 pages, 10 figures, 4 tables, https://hal.inria.fr/hal-01413414

[44] X. LI, L. GIRIN, R. HORAUD, S. GANNOT. *Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization*, Inria Grenoble - Rhone-Alpes, November 2016, 13 pages, 3 figures, 3 tables, https://hal.inria.fr/hal-01413417