IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2014

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

Keywords: Computer Vision, Auditory Signal Processing, Machine Learning, Audio-Visual Fusion, Human-Robot Interaction

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

1. Members

Research Scientists
Radu Horaud [Team leader, Inria, Senior Researcher, HdR]
Sileye Ba [Inria, Startring Research Position, from Jun 2014]
Georgios Evangelidis [Inria, Startring Research Position]

Faculty Member
Laurent Girin [INP Grenoble, Professor, HdR]

Engineers
Pierre Arquier [Inria, granted by ANR MIXCAM project]
Quentin Pelorson [Inria, granted by ANR MIXCAM project]

PhD Students
Israel–Dejene Gebru [Inria]
Vincent Drouard [Inria]
Maxime Janvier [Inria, until Aug 2014]
Dionyssos Kounades-Bastian [Inria]
Kaustubh Kulkarni [Inria]
Stéphane Lathuiliere [Inria, from Oct 2014]
Benoît Massé [Inria, from Oct 2014]

Post-Doctoral Fellows
Xiaofei Li [Inria, from Feb 2014, granted by FP7 EARS project]
Xavier Alameda-Pineda [Inria, from Jan 2014 until Jun 2014, granted by ERC VHIA project]

Visiting Scientists
Sharon Gannot [Inria, professor, Feb 2014]
Manuel Jesus Marin Jimenez [Inria, professor, from Jul 2014 until Aug 2014]

Administrative Assistant
Nathalie Gillot [Inria]

Others
Cosmin George Alexandru [Inria, student internship, from Feb 2014 until Aug 2014]
Konstantinos Georgiadis [Inria, student internship, from Feb 2014 until Aug 2014]
Adrian Ioan Postelnicu [Inria, student internship, from Feb 2014 until Aug 2014]
Alessio Xompero [Inria, student internship, from Mar 2014 until Aug 2014]

2. Overall Objectives

2.1. Overall Objectives

The main objective of the PERCEPTION team is to study the fundamental role played by audio-visual perception in human-robot interaction.

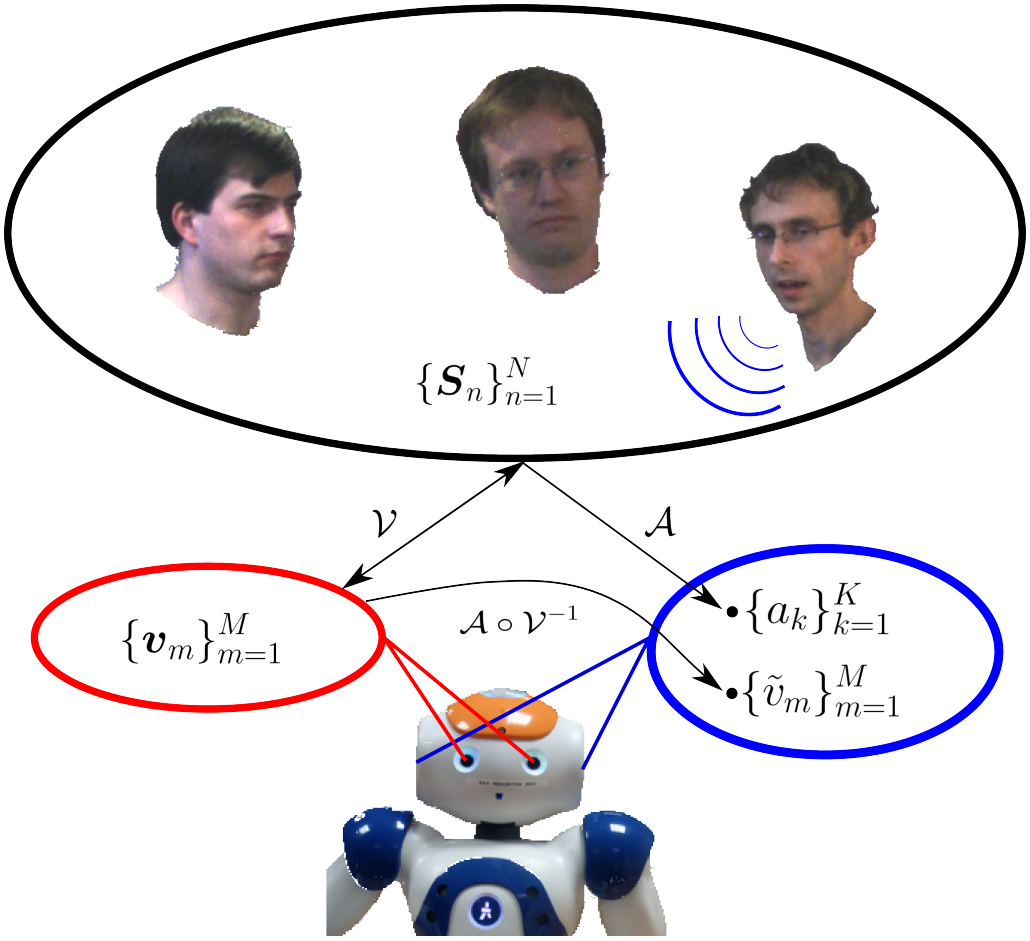

Figure 1. This figure illustrates the general principle of the latent-variable mixture models for audio-visual data analysis that the PERCEPTION team have developed [9], [16]. Audiovisual events ($\mathbf{s}$), e.g., speaking faces, are observed with two cameras and two microphones, hence two types of observations are available: 3D binocular features ($\mathbf{v}$) and 1D binaural features (a). By combining the inverse visual mapping with the direct auditory mapping, $\mathcal{A} \circ \mathcal{V}^{-1}$, it is possible to project 3D visual features onto an 1D auditory space, to represent visual and auditory data in the same space, and to properly cluster and classify them.

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

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

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

3. Research Program

3.1. Audio-Visual Scene Analysis

From 2006 to 2009, R. Horaud was the scientific coordinator of the collaborative European project POP (Perception on Purpose), an interdisciplinary effort to understand visual and auditory perception at the crossroads of several disciplines (computational and biological vision, computational auditory analysis, robotics, and psychophysics). This allowed the PERCEPTION team to launch an interdisciplinary research agenda that has been very active for the last five years. There are very few teams in the world that gather scientific competences spanning computer vision, audio signal processing, machine learning and human-robot interaction. The fusion of several sensorial modalities resides at the heart of the most recent biological theories of perception. Nevertheless, multi-sensor processing is still poorly understood from a computational point of view. In particular and so far, audio-visual fusion has been investigated in the framework of speech processing using close-distance cameras and microphones. The vast majority of these approaches attempt to model the temporal correlation between the auditory signals and the dynamics of lip and facial movements. Our original contribution has been to consider that audio-visual localization and recognition are equally important. We have proposed to take into account the fact that the audio-visual objects of interest live in a three-dimensional physical space and hence we contributed to the emergence of *audio-visual scene analysis* as a scientific topic in its own right. We proposed several novel statistical approaches based on supervised and unsupervised mixture models. The *conjugate mixture model* (CMM) is an unsupervised probabilistic model that allows to cluster observations from different modalities (e.g., vision and audio) living in different mathematical spaces [9], [16]. 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 [15]. 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 [19]. 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 [18]. The European project HUMAVIPS (2010-2013), coordinated by R. Horaud, applied audio-visual scene analysis to human-robot interaction.

3.2. Binocular 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 [4], [11]. 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 [5]. This addresses the fundamental issue of how local image structure is represented in the brain/computer and how this structure is used for estimating a dense disparity field. Therefore, the main originality of our work is to address both computational and biological issues within a unifying model of binocular vision. Another equally important problem that still remains to be solved is how to integrate binocular depth maps over time. Recently, we have addressed this problem and proposed a semi-global optimization framework that starts with sparse yet reliable matches and proceeds with propagating them over both space and time. The concept of seed-match propagation has then been extended to TOF-stereo fusion.

3.3. Binaural Hearing

Audio-visual fusion algorithms necessitate that the two modalities are represented in the same mathematical space. Binaural hearing 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 [16]. 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 [15]. 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 [18] and audio-visual learning [24]. Currently we generalize this approach to an arbitrary number of microphones.

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 [14]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [12]. 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 [13]. 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 [34], [21], [20]. 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.

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 [10]. 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 [8], [7]. 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 [17]. This turned out to be a very original approach that has been followed by many other researchers in computer vision and computer graphics.

4. New Software and Platforms

4.1. The MIXCAM Hardware/Software Platform

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.

Website: https://team.inria.fr/perception/mixcam-lab/

4.2. Audiovisual Robots and Heads

We have developed two audiovisual (AV) robot heads: the POPEYE head and the NAO stereo head. Both are equipped with a binocular vision system and with four microphones. The software modules comprise stereo matching and reconstruction, sound-source localization and audio-visual fusion. POPEYE has been developed within the European project POP (https://team.inria.fr/perception/pop/) in collaboration with the project-team MISTIS and with two other POP partners: the Speech and Hearing group of the University of

Figure 2. The MIXCAM laboratory is a multiple-camera multiple-PC hardware/software platform that combines high-resolution color (RGB) cameras with low-resolution time-of-flight (TOF) cameras. The cameras are arranged in "units", where each unit is composed of two RGB cameras and one TOF camera (left image). Currently the system is composed of four such units (right image), or a total of eight RGB and four TOF cameras. Over years, in collaboration with 4D View Solutions, we have developed and maintained software packages for camera, multiple-camera, and cross-modal calibration, 3D reconstruction, multiple-camera stereo, TOF-stereo fusion, and image-based rendering.

Sheffield and the Institute for Systems and Robotics of the University of Coimbra. The NAO stereo head was developed under the European project HUMAVIPS (http://humavips.inrialpes.fr) in collaboration with Aldebaran Robotics (which manufactures the humanoid robot NAO) and with the University of Bielefeld, the Czech Technical Institute, and IDIAP. The software modules that we develop are compatible with both these robot heads [33].

For more information on POPEYE and on NAO please visit https://team.inria.fr/perception/popeye/ and https://team.inria.fr/perception/nao/.

5. New Results

5.1. Highlights of the Year

- In 2014 Antoine Deleforge (team member 2009-2013) received the **Signal, Image and Vision best PhD prize** for his thesis *"Acoustic Space Mapping: A Machine Learning Approach to Sound Source Separation and Localization"*, defended in December 2013 and advised by Radu Horaud. The prize is jointly awarded by GDR ISIS, Club EEA, and GRETSI.

  Website: http://www.inria.fr/centre/grenoble/actualites/apprendre-a-rester-attentif-a-ses-locuteurs
- Radu Horaud was awarded an **ERC Advanced Grant** for his five year project VHIA *"Vision and Hearing in Action"*, grant number 340113, 2014-2019.

  Website: https://team.inria.fr/perception/vhia/.
- The PERCEPTION team was awarded an **ANR BLANC** two year project MIXCAM *"Real-Time Visual Reconstruction by Mixing Multiple Depth and Color Cameras"*, in collaboration with 4D View Solutions, 2014-2016.

  Website: https://team.inria.fr/perception/mixcam-project/

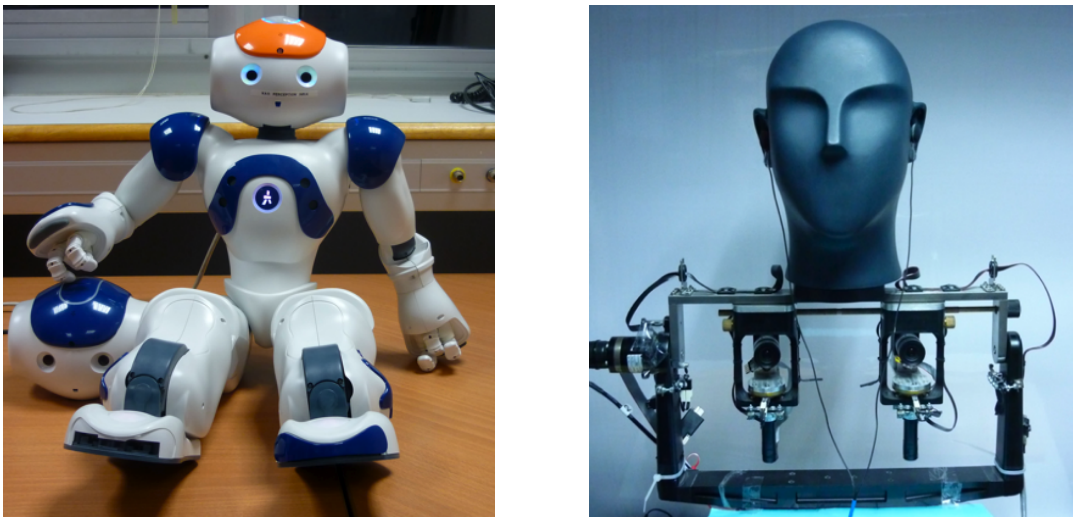

Figure 3. Left: The consumer humanoid robot NAO is equipped with a binocular-binaural head specially designed for human-humanoid interaction; Right: The binocular-binaural robot head POPEYE equipped with a four degrees of freedom stereo camera pair and with an acoustic dummy head.

- The PERCEPTION team was awarded an **FP7 STREP** three year project EARS *"Embodied Audition for Robots"*, in collaboration with Friedrich Alexander Universiteit, coordinator (Germany), Ben Gurion University (Israel), Imperial College (UK), Humboldt University Berlin (Germany) and Aldebaran Robotics (France), 2014-2017.

  Website: https://team.inria.fr/perception/ears/

5.2. Acoustic Space Learning on Binaural Manifolds

We addressed the problems of modeling the acoustic space generated by a full-spectrum sound source and of using the learned model for the localization and separation of multiple sources that simultaneously emit sparse-spectrum sounds. We lay theoretical and methodological grounds in order to introduce the binaural manifold paradigm. We perform an in-depth study of the latent low-dimensional structure of the high-dimensional interaural spectral data, based on a corpus recorded with a human-like audiomotor robot head, namely the POPEYE robot shown on Fig 3 (right). A non-linear dimensionality reduction technique is used to show that these data lie on a two-dimensional (2D) smooth manifold parameterized by the motor states of the listener, or equivalently, the sound source directions, e.g., Fig. 4. We propose a probabilistic piecewise affine mapping model (PPAM) specifically designed to deal with high-dimensional data exhibiting an intrinsic piecewise linear structure. We derive a closed-form expectation-maximization (EM) procedure for estimating the model parameters, followed by Bayes inversion for obtaining the full posterior density function of a sound source direction. We extend this solution to deal with missing data and redundancy in real world spectrograms, and hence for 2D localization of natural sound sources such as speech. We further generalize the model to the challenging case of multiple sound sources and we propose a variational EM framework. The associated algorithm, referred to as variational EM for source separation and localization (VESSL) yields a Bayesian estimation of the 2D locations and time-frequency masks of all the sources. Comparisons of the proposed approach with several existing methods reveal that the combination of acoustic-space learning with Bayesian inference enables our method to outperform state-of-the-art methods [18], [24].

Website: https://team.inria.fr/perception/research/acoustic-learning/

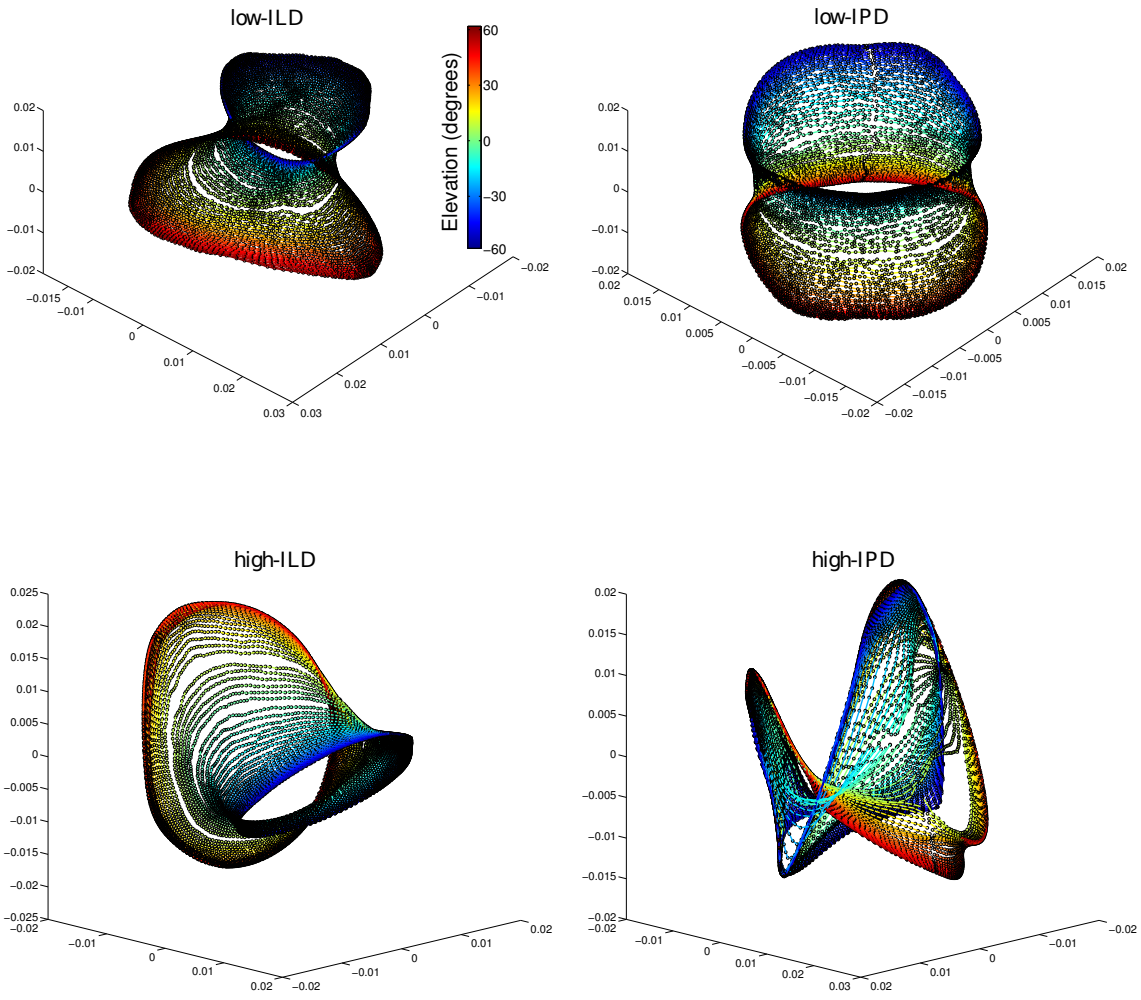

Figure 4. This figure illustrates the concept of binaural manifold. A wide-spectrum sound is recorded with a binaural acoustic dummy head and an interaural high-dimensional spectral representation of this sound is mapped onto a low-dimensional (2) space. This reveals the two-dimensional manifold of possible sound-source directions that is embedded in the interaural spectral features. Please consult [18] for more details.

5.3. Geometric Sound Source Localization

We address the problem of sound-source localization from time-delay estimates using arbitrarily-shaped non-coplanar microphone arrays. A novel geometric formulation is proposed, together with a thorough algebraic analysis and a global optimization solver [15]. The proposed model is thoroughly described and evaluated. The geometric analysis, stemming from the direct acoustic propagation model, leads to necessary and sufficient conditions for a set of time delays to correspond to a unique position in the source space. Such sets of time delays are referred to as *feasible sets*. We formally prove that every feasible set corresponds to exactly one position in the source space, whose value can be recovered using a closed-form localization mapping. Therefore we seek for the optimal feasible set of time delays given, as input, the received microphone signals. This time delay estimation problem is naturally cast into a programming task, constrained by the feasibility conditions derived from the geometric analysis. A global branch-and-bound optimization technique is proposed to solve the problem at hand, hence estimating the best set of feasible time delays and, subsequently, localizing the sound source. Extensive experiments with both simulated and real data are reported; we compare our methodology to four state-of-the-art techniques. This comparison shows that the proposed method combined with the branch-and-bound algorithm outperforms existing methods. These in-depth geometric understanding, practical algorithms, and encouraging results, open several opportunities for future work.

Website: https://team.inria.fr/perception/research/geometric-sound-source-localization/

5.4. Joint Registration of Multiple Point Sets

We developed a probabilistic generative model and its associated algorithm to jointly register multiple point sets. The vast majority of state-of-the-art registration techniques select one of the sets as the *model* and perform pairwise alignments between the other sets and this set. The main drawback of this mode of operation is that there is no guarantee that the model-set is free of noise and outliers, which contaminates the estimation of the registration parameters. Unlike previous work, the proposed method treats all the point sets on an equal footing: they are realizations of a Gaussian mixture (GMM) and the registration is cast into a clustering problem [26]. We formally derive an EM algorithm that estimates both the GMM parameters and the rotations and translations that map each individual set onto the *central* model. The mixture means play the role of the registered set of points while the variances provide rich information about the quality of the registration. We thoroughly validate the proposed method with challenging datasets, we compare it with several state-of-the-art methods, and we show its potential for fusing real depth data.

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

5.5. High-Dimensional Regression

The problem of approximating high-dimensional data with a low-dimensional representation is addressed. The article makes the following contributions. An inverse regression framework is proposed, which exchanges the roles of input and response, such that the low-dimensional variable becomes the regressor, and which is tractable. A mixture of locally-linear probabilistic mapping model is introduced, that starts with estimating the parameters of the inverse regression, and follows with inferring closed-form solutions for the forward parameters of the high-dimensional regression problem of interest. Moreover, a partially-latent paradigm is introduced, such that the vector-valued response variable is composed of both observed and latent entries, thus being able to deal with data contaminated by experimental artifacts that cannot be explained with noise models. The proposed probabilistic formulation could be viewed as a latent-variable augmentation of regression. Expectation-maximization (EM) procedures are introduced, based on a data augmentation strategy which facilitates the maximum-likelihood search over the model parameters. Two augmentation schemes are proposed and the associated EM inference procedures are described in detail; they may well be viewed as generalizations of a number of EM regression, dimension reduction, and factor analysis algorithms. The proposed framework is validated with both synthetic and real data. Experimental evidence is provided that the method outperforms several existing regression techniques [19], [25].

Website: https://team.inria.fr/perception/research/high-dim-regression/

5.6. Audiovisual Speaker Detection, Localization and Interaction with NAO

In this research we address the problem of audio-visual speaker detection. We introduce an online system working on the humanoid robot NAO. The scene is perceived with two cameras and two microphones. A *multimodal* Gaussian mixture model (GMM) fuses the information extracted from the auditory and visual sensors. The system is implemented based on a platform-independent middleware library and it is able to process the information online (17 visual frames per second). A detailed method description and the system implementation are provided, with special emphasis on the online processing issues that must be addressed, and the proposed solutions. Experimental validation is done over five different scenarios, with no special lighting, nor special acoustic conditions, leading to good results [16].

Website: https://team.inria.fr/perception/research/audiovisual-nao/

5.7. EM for Weighted-Data Clustering

Figure 5. We developed a novel multimodal clustering method that is based on expectation-maximization (EM) with weighted data. The left image shows auditory features (green), namely sound source positions mapped onto the image plane using [24] and visual features (blue, lip landmarks), as well as the active speaker (yellow square). The right image shows the results of our weighted-data EM algorithm that finds three clusters. Among these clusters, the active audio-visual cluster is marked with a transparent blue circle.

Data clustering has received a lot of attention and many methods, algorithms and software packages are currently available. Among these techniques, parametric finite-mixture models play a central role due to their interesting mathematical properties and to the existence of maximum-likelihood estimators based on expectation-maximization (EM). In this paper we propose a new mixture model that associates a weight with each observed data point. We introduce a Gaussian mixture with weighted data and we derive two EM algorithms [29]: the first one considers the weight of each observed datum to be fixed, while the second one treats each weight as a hidden variable drawn from a gamma distribution. We provide a general-purpose scheme for weight initialization and we thoroughly validate the proposed algorithms by comparing them with several parametric and non-parametric clustering techniques. We demonstrate the utility of our method for clustering heterogeneous data, namely data gathered with different sensorial modalities, e.g., audio and vision.

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

5.8. Continuous Action Recognition

Continuous action recognition is more challenging than isolated recognition because classification and segmentation must be simultaneously carried out. We build on the well known dynamic time warping (DTW) framework and devise a novel visual alignment technique, namely dynamic frame warping (DFW), which performs isolated recognition based on per-frame representation of videos, and on aligning a test sequence with a model sequence. Moreover, we propose two extensions which enable to perform recognition concomitant with segmentation, namely one-pass DFW and two-pass DFW. These two methods have their roots in the domain of continuous recognition of speech and, to the best of our knowledge, their extension to continuous visual action recognition has been overlooked. We test and illustrate the proposed techniques with a recently released dataset (RAVEL) [32] and with two public-domain datasets widely used in action recognition (Hollywood-1 and Hollywood-2). We also compare the performances of the proposed isolated and continuous recognition algorithms with several recently published methods [22].

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

5.9. Skeletal Quads

Recent advances on human motion analysis have made the extraction of human skeleton structure feasible, even from single depth images. This structure has been proven quite informative for discriminating actions in a recognition scenario. In this context, we propose a local skeleton descriptor that encodes the relative position of joint quadruples. Such a coding implies a similarity normalization transform that leads to a compact (6D or 5D) view-invariant skeletal feature, referred to as skeletal quad. In the references below, we use this descriptor in conjunction with Fisher kernel in order to encode gesture or action (sub)sequences. The short length of the descriptor compensates for the large inherent dimensionality associated to Fisher vectors. We investigate the performance in both isolated [28] and continuous [27] recognition scenarios.

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

6. Partnerships and Cooperations

6.1. National Initiatives

6.1.1. ANR

6.1.1.1. 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.

6.2. European Initiatives

6.2.1. FP7 & H2020 Projects

6.2.1.1. EARS

Type: FP7

Challenge: Cognitive Systems and Robotics

Instrument: Specific Targeted Research Project

Objectif: Robotics, Cognitive Systems and Smart Spaces, Symbiotic Interaction

Duration: January 2014 - December 2016

Coordinator: Friedrich Alexander Universiteit (Germany)

Partners: Inria (France), Ben Gurion University (Israel), Imperial College (UK), Humboldt University Berlin (Germany), and Aldebaran Robotics (France)

Inria contact: Radu Horaud

Abstract: 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 localize 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.

6.2.1.2. VHIA

Type: FP7

Instrument: ERC Advanced Grant

Duration: February 2014 - January 2019

Principal Investigator: Radu Horaud

Abstract: The objective of VHIA is to elaborate a holistic computational paradigm of perception and of perception-action loops. We propose 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 launch and 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 that are based on unimodal biological theories. These theories have hypothesized a modular view of perception, postulating that there are quasi-independent and parallel perceptual pathways in the brain. VHIA takes 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 whose intentions inherently supervise both 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, such as how to build a joint audiovisual space from heterogeneous, noisy, ambiguous and physically different visual and auditory stimuli, how to properly model seamless interaction based on perception and action, 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 robot technology will have a considerable social and economical impact and that there will be millions of humanoids, in our homes, schools and offices, which will be able to naturally communicate with us.

6.3. International Initiatives

6.3.1. Inria International Partners

6.3.1.1. Declared Inria International Partners

- The Czech Technical University in Prague (Dr. Jan Cech)
- The Technion (Prof. Yoav Schechner)
- Queen Mary University London (Dr. Miles Hansard)
- Bar Ilan University (Prof. Sharon Gannot)
- University of Cordoba (Prof. Manuel Jesus Marin Jimenez)
- University of Patras (Prof. Manolis Psarakis)
- Oxford Brookes University (Dr. Fabio Cuzzolin)

6.4. International Research Visitors

6.4.1. Visits of International Scientists

- Prof. Sharon Gannot (Bar Ilan University)
- Prof. Manuel Jesus Marin Jimenez (Cordoba University)

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. Journals

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*.

8. Bibliography

Major publications by the team in recent years

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

[2] Y. DUFOURNAUD, C. SCHMID, R. HORAUD. *Image matching with scale adjustment*, in "Computer Vision and Image Understanding", February 2004, vol. 93, n° 2, pp. 175–194 [*DOI :* 10.1016/J.CVIU.2003.07.003], http://hal.inria.fr/inria-00548555

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

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

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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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° 6, pp. 1082-1095 [*DOI :* 10.1109/TASLP.2014.2317989], https://hal.inria.fr/hal-00975293

[16] X. ALAMEDA-PINEDA, R. HORAUD. *Vision-Guided Robot Hearing*, in "The International Journal of Robotics Research", October 2014, 20 p. [*DOI :* 10.1177/0278364914548050], https://hal.inria.fr/hal-00990766

[17] F. CUZZOLIN, D. MATEUS, R. HORAUD. *Robust Temporally Coherent Laplacian Protrusion Segmentation of 3D Articulated Bodies*, in "International Journal of Computer Vision", August 2014, 28 p. [*DOI :* 10.1007/s11263-014-0754-0], https://hal.archives-ouvertes.fr/hal-01053737

[18] A. DELEFORGE, F. FORBES, R. HORAUD. *Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds*, in "International Journal of Neural Systems", March 2014 [*DOI :* 10.1142/S0129065714400036], https://hal.inria.fr/hal-00960796

[19] A. DELEFORGE, F. FORBES, R. HORAUD. *High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables*, in "Statistics and Computing", February 2014 [*DOI :* 10.1007/s11222-014-9461-5], https://hal.inria.fr/hal-00863468

[20] M. HANSARD, G. EVANGELIDIS, Q. PELORSON, R. HORAUD. *Cross-calibration of Time-of-flight and Colour Cameras*, in "Computer Vision and Image Understanding", November 2014, 11 p. [*DOI :* 10.1016/j.cviu.2014.09.001], https://hal.inria.fr/hal-01059891

[21] 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

[22] K. KULKARNI, G. EVANGELIDIS, J. CECH, R. HORAUD. *Continuous Action Recognition Based on Sequence Alignment*, in "International Journal of Computer Vision", August 2014, 26 p. [*DOI :* 10.1007/s11263-014-0758-9], https://hal.archives-ouvertes.fr/hal-01058732

[23] R. PERRIER, E. ARNAUD, P. STURM, M. ORTNER. *Estimation of an Observation Satellite's Attitude using Multimodal Pushbroom Cameras*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", September 2014, 14 p. [*DOI :* 10.1109/TPAMI.2014.2360394], https://hal.inria.fr/hal-01093238

International Conferences with Proceedings

[24] A. DELEFORGE, V. DROUARD, L. GIRIN, R. HORAUD. *Mapping Sounds on Images Using Binaural Spectrograms*, in "22nd European Signal Processing Conference (EUSIPCO-2014)", Lisbonne, Portugal, September 2014, https://hal.archives-ouvertes.fr/hal-01019287

[25] A. DELEFORGE, F. FORBES, R. HORAUD. *Hyper-spectral Image Analysis with Partially-Latent Regression*, in "22nd European Signal Processing Conference", Lisbon, Portugal, September 2014, https://hal.archives-ouvertes.fr/hal-01019360

[26] G. EVANGELIDIS, D. KOUNADES-BASTIAN, R. HORAUD, E. PSARAKIS. *A Generative Model for the Joint Registration of Multiple Point Sets*, in "European Conference on Computer Vision", Zurich, Switzerland, Springer, September 2014, pp. 109-122 [*DOI :* 10.1007/978-3-319-10584-0_8], https://hal.archives-ouvertes.fr/hal-01019661

[27] G. EVANGELIDIS, G. SINGH, R. HORAUD. *Continuous gesture recognition from articulated poses*, in "ChaLearn Looking at People Workshop in conjunction with ECCV 2014 – European Conference on Computer Vision", Zurich, Switzerland, September 2014, https://hal.archives-ouvertes.fr/hal-01082981

[28] G. EVANGELIDIS, G. SINGH, R. HORAUD. *Skeletal Quads: Human Action Recognition Using Joint Quadruples*, in "ICPR 2014 - International Conference on Pattern Recognition", Stockholm, Sweden, September 2014, https://hal.archives-ouvertes.fr/hal-00989725

[29] I.-D. GEBRU, X. ALAMEDA-PINEDA, R. HORAUD, F. FORBES. *Audio-Visual Speaker Localization via Weighted Clustering*, in "IEEE Workshop on Machine Learning for Signal Processing", Reims, France, September 2014, 6 p. , https://hal.archives-ouvertes.fr/hal-01053732

[30] M. JANVIER, X. ALAMEDA-PINEDA, L. GIRIN, R. HORAUD. *Sound Representation and Classification Benchmark for Domestic Robots*, in "2014 IEEE International Conference on Robotics and Automation (ICRA 2014)", Hong-Kong, China, May 2014, https://hal.inria.fr/hal-00952092

Other Publications

[31] M. HANSARD, G. EVANGELIDIS, Q. PELORSON, R. HORAUD. *Cross-calibration of Time-of-flight and Colour Cameras*, June 2014, 18 pages, 12 figures, 3 tables, https://hal.archives-ouvertes.fr/hal-01019668

References in notes

[32] X. ALAMEDA-PINEDA, J. SANCHEZ-RIERA, J. WIENKE, V. FRANC, J. CECH, K. KULKARNI, A. DELEFORGE, R. HORAUD. *RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities*, in "Journal on Multimodal User Interfaces", March 2013, vol. 7, n^{o} 1-2, pp. 79-91 [*DOI :* 10.1007/s12193-012-0111-y], http://hal.inria.fr/hal-00720734

[33] J. CECH, R. MITTAL, A. DELEFORGE, J. SANCHEZ-RIERA, X. ALAMEDA-PINEDA, R. HORAUD. *Active-Speaker Detection and Localization with Microphones and Cameras Embedded into a Robotic Head*, in

"Humanoids 2013 - IEEE-RAS International Conference on Humanoid Robots", Atlanta, United States, IEEE Robotics Society, September 2013, http://hal.inria.fr/hal-00861465

[34] 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