RESEARCH CENTRE

Grenoble - Rhône-Alpes

2020
ACTIVITY REPORT

Project-Team

PERCEPTION

Interpretation and Modelling of Images and Videos

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Contents

Project-Team PERCEPTION

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

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A5.1. – Human-Computer Interaction

A5.3. – Image processing and analysis

A5.4. – Computer vision

A5.7. – Audio modeling and processing

A5.10.2. – Perception

A5.10.5. – Robot interaction (with the environment, humans, other robots)

A9.2. – Machine learning

A9.5. – Robotics

Other research topics and application domains

B5.6. – Robotic systems

1 Team members, visitors, external collaborators

Research Scientists

- Radu Horaud [Team leader, Inria, Senior Researcher, HDR]
- Xavier Alameda-Pineda [Inria, Researcher, HDR]
- Chris Reinke [Inria, Starting Research Position, from Mar 2020]
- Mostafa Sadeghi [Inria, Starting Research Position, from Aug 2020 until Oct 2020]
- Timothee Wintz [Inria, Starting Research Position, from Apr 2020]

Post-Doctoral Fellow

- Mostafa Sadeghi [Inria, until Jul 2020]

PhD Students

- Anand Ballou [Univ Grenoble Alpes]
- Xiaoyu Bie [Univ Grenoble Alpes]
- Guillaume Delorme [Inria]
- Wen Guo [Univ Grenoble Alpes]
- Gaetan Lepage [Inria, from Oct 2020]
- Xiaoyu Lin [Inria, from Nov 2020]
- Yihong Xu [Inria]

Technical Staff

- Soraya Arias [Inria, Engineer]
- Alex Auternaud [Inria, Engineer]
- Luis Gomez Camara [Inria, Engineer, from Sep 2020]
- Zhiqi Kang [Inria, Engineer, from Sep 2020]
- Matthieu Py [Inria, Engineer, from Feb 2020]

Interns and Apprentices

- Alvaro Gonzalez Jimenez [ENSIMAG, from Feb 2020 until Jul 2020]
- Zhiqi Kang [Inria, from Feb 2020 until Aug 2020]
- Viet Nhat Nguyen [Inria, from Feb 2020 until Aug 2020]
- Predrag Pilipovic [Inria, from Feb 2020 until Jul 2020]

Administrative Assistant

- Nathalie Gillot [Inria]

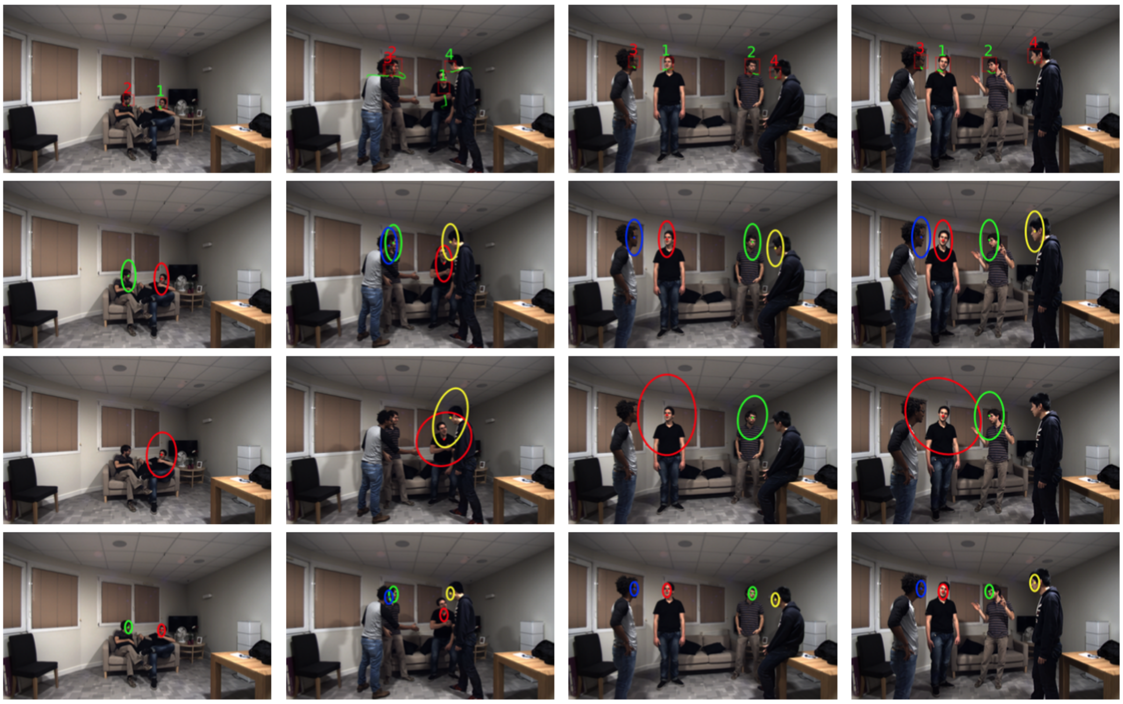


Figure 1: This figure illustrates the audio-visual multiple-person tracking that has been developed by the team [43, 30]. The tracker is based on variational inference [5] and on supervised sound-source localization [11, 34]. Each person is identified with a digit. Green digits denote speaking persons, while red digits denote silent ones. The next rows show the covariances (uncertainties) associated with the visual (second row), audio (third row) and dynamic (fourth row) contributions for tracking a varying number of persons. Notice the large uncertainty associated with audio and the small uncertainty associated with the dynamics of the tracker. In the light of this example, one may notice the complementary roles played by vision and audio: vision data are more accurate while audio data provide speech information (who speaks when). These developments have been supported by the European Union via the FP7 STREP project *"Embodied Audition for Robots"* (EARS), the ERC advanced grant *"Vision and Hearing in Action"* (VHIA) and ERC Proof of Concept grand VHIALab.

External Collaborators

- Yutong Ban [MIT, Boston Ma, US, until May 2020]
- Laurent Girin [Institut polytechnique de Grenoble, HDR]
- Simon Leglaive [CNRS, until Aug 2020]
- Xiaofei Li [Université WestLake Shangaï - Chine]

2 Overall objectives

2.1 Audio-Visual Machine Perception

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

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

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

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

3 Research program

3.1 Audio-Visual Scene Analysis

From 2006 to 2009, R. Horaud was the scientific coordinator of the collaborative European project POP (Perception on Purpose), an interdisciplinary effort to understand visual and auditory perception at the crossroads of several disciplines (computational and biological vision, computational auditory analysis, robotics, and psychophysics). This allowed the PERCEPTION team to launch an interdisciplinary research agenda that has been very active for the last five years. There are very few teams in the world that gather scientific competences spanning computer vision, audio signal processing, machine learning and human-robot interaction. The fusion of several sensorial modalities resides at the heart of the most recent biological theories of perception. Nevertheless, multi-sensor processing is still poorly understood from a computational point of view. In particular and so far, audio-visual fusion has been investigated in the framework of speech processing using close-distance cameras and microphones. The vast majority of these approaches attempt to model the temporal correlation between the auditory signals and the dynamics of lip and facial movements. Our original contribution has been to consider that audio-visual localization and recognition are equally important. We have proposed to take into account the fact that the audio-visual objects of interest live in a three-dimensional physical space and hence we contributed to the emergence of *audio-visual scene analysis* as a scientific topic in its own right. We proposed several novel statistical approaches based on supervised and unsupervised mixture models. The *conjugate mixture model* (CMM) is an unsupervised probabilistic model that allows to cluster observations from different modalities (e.g., vision and audio) living in different mathematical spaces [26], [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 [10]. 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 [9]. 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 [20, 38]. 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 [18]. 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 [13].

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 [9] and audio-visual learning [11].

We also addressed the difficult problem of audio signal processing in reverberant environments. We thoroughly studied the *convolutive transfer function* (CTF) and developed several methods and associated algorithms for the localization, separation and tracking of audio sources [34, 32, 33, 35, 31]

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 [41]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [40]. 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 [39]. 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 [22], [21],[17]. 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 [13].

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 [27]. 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 [25, 24]. 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 [8]. This turned out to be a very original approach that has been followed by many other researchers in computer vision and computer graphics.

4 Application domains

The research topics of Perception have strong social impact with great economic value. In more detail, Perception has put strong emphasis on the development of *socially assistive robots* SARs. In particular we apply social robot methodologies to gerontological healthcare. Several recently released demographic studies have emphasized that the number of very old people (over 80) will considerably increase, reaching 13% (66 million people) of the EU population by 2080. For the same period of time the working-age population will shrink, thus raising the critical issue of how the elderly will be taken care of in the coming century. This is a major concern for our civilisation that has been largely under-estimated. There is a risk that the elderly are being left behind the digitization of the society. Consequently, the development of robotic technologies addressing their needs and their expectations is mandatory. Team members will work hand-in-hand with healthcare professionals to apprehend the current limitations of SARs, to evaluate their potential, and to establish several use-cases.

The deployment of SARs, from academic laboratories to hospitals and to retirement homes, is far from being a trivial issue and it should not be done in isolation. The team is committed to work in collaboration with industrial partners, able to develop, commercialize, maintain and update SAR technologies. The Perception team has a great deal of collaborative experience with the robotics industry. In particular we have directly collaborated with Softbank Robotics Europe (formerly Aldebaran Robotics) and conducted several technology transfer actions. Currently, we work in collaboration with PAL Robotics (Spain), a leading R&D robot manufacturer, and with ERM Automatismes (France), an SME specialized in the development of service robots. These two companies are part of the EU project SPRING. It is worth to be mentioned that the Broca Hospital (Assistance Publique Hôpitaux de Paris) is a member of SPRING as well, and is highly committed to provide ethical guidance, recommendations for robot-technology acceptance by very old people, and in-depth experimental validation.

5 Highlights of the year

5.1 Awards

Xavier Alameda-Pineda and his collaborators from the University of Trento received the ACM-TOMM 2020 Nicolas D. Georganas Best Paper Award, that recognizes the most significant work in ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM) in a given calendar year: Increasing image memorability with neural style transfer, vol. 15 Issue 2, January 2019 by A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci, N. Sebe.

5.2 New Projects and Collaborations

5.2.1 Collaboration with Facebook Reality Research Lab, USA

In this project, we investigate *visually assisted speech processing*. In particular we plan to go beyond the current paradigm that systematically combines a noisy speech signal with clean lip images and which delivers a clean speech signal. The rationale of this paradigm is based on the fact that lip images are free of any type of noise. This hypothesis is merely verified in practice. Indeed, speech production is often accompanied by head motions that considerably modify the patterns of the observed lip movements. As a consequence, currently available audio-visual speech processing technologies are not usable in practice. In this project we develop a methodology that separates non-rigid face- and lip movements from rigid head movements, and we build a deep generative architecture that combines audio and visual features based on their relative merits, rather than making systematic recourse to their concatenation. The core methodology is based on robust mixture modeling and on variational auto-encoders, two methodologies that have been thoroughly investigated by the team.

5.2.2 H2020 Project SPRING

Started on Januray 1st, 2020 and finalising on May 31st, 2024, SPRING is a research and innovation action (RIA) with eight partners: Inria Grenoble (coordinator), Università degli Studi di Trento, Czech Technical University Prague, Heriot-Watt University Edinburgh, Bar-Ilan University Tel Aviv, ERM Automatismes Industriels Carpentras, PAL Robotics Barcelona, and Hôpital Broca Paris.. The main objective of SPRING (Socially Pertinent Robots in Gerontological Healthcare) is the development of socially assistive robots with the capacity of performing multimodal multiple-person interaction and open-domain dialogue. In more detail:

- The scientific objective of SPRING is to develop a novel paradigm and novel concept of socially-aware robots, and to conceive innovative methods and algorithms for computer vision, audio processing, sensor-based control, and spoken dialog systems based on modern statistical- and deep-learning to ground the required social robot skills.
- The technological objective of SPRING is to create and launch a brand new generation of robots that are flexible enough to adapt to the needs of the users, and not the other way around.
- The experimental objective of SPRING is twofold: to validate the technology based on HRI experiments in a gerontology hospital, and to assess its acceptability by patients and medical staff.

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

5.2.3 ANR JCJC Project ML3RI

Starting on March 1st 2020 and finalising on February 28th 2024, ML3RI is an ANR JCJC that has been awarded to Xavier Alameda-Pineda. Multi-person robot interaction in the wild (i.e. unconstrained and using only the robot's resources) is nowadays unachievable because of the lack of suitable machine perception and decision-taking models. *Multi-Modal Multi-person Low-Level Learning models for Robot Interaction* (ML3RI) has the ambition to develop the capacity to understand and react to low-level behavioral cues, which is crucial for autonomous robot communication. The main scientific impact of

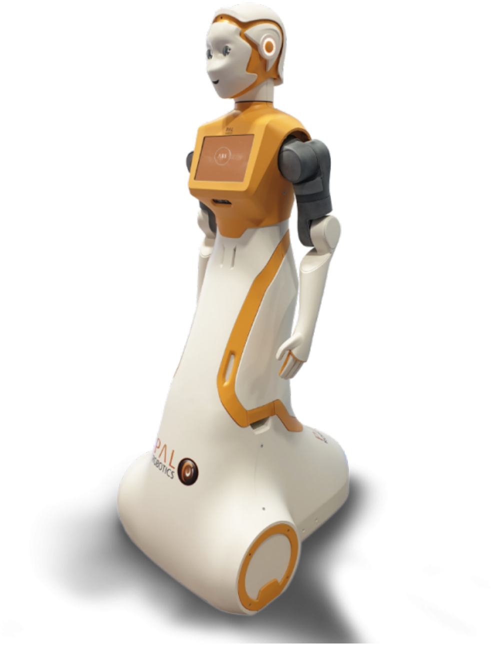


Figure 2: ARI is a robot prototype (not yet a commercially available product) designed and manufactured by PAL Robotics, located in Barcelona, and a member of the SPRING consortium.

ML3RI is to develop new learning methods and algorithms, thus opening the door to study multi-party conversations with robots. In addition, the project supports open and reproducible research.

Website: https://project.inria.fr/ml3ri/

5.3 The Social Robot ARI

The team participated to the specifications of a social-robot prototype manufactured by PAL Robotics, an industrial partner of the SPRING project. ARI is a non-holonomic differential-drive wheeled robot equipped with a pan and tilt head, with both color and depth cameras and with a microphone array that embeds the latest audio signal processing technologies. The challenge is to devolop a software suite, from low-level control to high-level planning, such that the robot has a socially-aware behaviour while it safely navigates in an ever changing environment.

6 New software and platforms

6.1 New software

6.1.1 deepMot

Name: A Differentiable Framework for Training Multi-Object Trackers

Keywords: Deep learning, Computer vision, Multi-Object Tracking

Scientific Description: The recent trend in vision-based multi-object tracking (MOT) is heading towards leveraging the representational power of deep learning to jointly learn to detect and track objects. However, existing methods train only certain sub-modules using loss functions that often do not correlate with established tracking evaluation measures such as Multi-Object Tracking Accuracy (MOTA) and Precision (MOTP). As these measures are not differentiable, the choice of appropriate loss functions for end-to-end training of multi-object tracking methods is still an open research problem. We bridge this gap by proposing a differentiable proxy of MOTA and MOTP, which we combine in a loss function suitable for end-to-end training of deep multi-object trackers. As a key ingredient, we propose a Deep Hungarian Net (DHN) module that approximates the Hungarian matching algorithm. DHN allows to estimate the correspondence between object tracks and ground truth objects to compute differentiable proxies of MOTA and MOTP, which are in turn used to optimize deep trackers directly. We experimentally demonstrate that the proposed differentiable framework improves the performance of existing multi-object trackers, and we establish a new state-of-the-art on the MOTChallenge benchmark.

Functional Description: We develop a differentiable proxy of MOTA and MOTP (Multi-Object Tracking Accuracy -MOTA and Precision-MOTP), which we combine in a loss function suitable for end-to-end training of deep multi-object trackers. As a key ingredient, we propose a Deep Hungarian Net (DHN) module that approximates the Hungarian matching algorithm. DHN allows to estimate the correspondence between object tracks and ground truth objects to compute differentiable proxies of MOTA and MOTP, which are in turn used to optimize deep trackers directly. We experimentally demonstrate that the proposed differentiable framework improves the performance of existing multi-object trackers, and we establish a new state-of-the-art on the MOTChallenge benchmark.

URL: https://team.inria.fr/perception/research/deepmot/

Publication: hal-02534894

Contact: Xavier Alameda-pineda

Participants: Yihong Xu, Xavier Alameda-pineda

Partner: Technical University of Munich (TUM)

6.1.2 dvae-speech

Name: dynamic variational auto-encoder for speech re-synthesis

Keywords: Variational Autoencoder, Deep learning, Pytorch, Speech Synthesis

Functional Description: It can be considered an library for speech community, to use different dynamic VAE models for speech re-synthesis (potentially for other speech application)

Authors: Xiaoyu Bie, Xavier Alameda-pineda, Laurent Girin

Contact: Xavier Alameda-pineda

6.1.3 AVSE-VAE

Name: Audio-visual speech enhancement based on variational autoencoder speech modeling

Keywords: Variational Autoencoder, Speech, Deep learning

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

Functional Description: This library contains PyTorch implementations of the audio-visual speech enhancement methods based on variational autoencoder (VAE), presented in the following publications : - M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoder,” August 2019. - M. Sadeghi and X. Alameda-Pineda, “Robust Unsupervised Audio-visual Speech Enhancement Using a Mixture of Variational Autoencoders,” in IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Barcelona, Spain, May 2020. - M. Sadeghi and X. Alameda-Pineda, “Mixture of Inference Networks for VAE-based Audio-visual Speech Enhancement,” December 2019.

Publications: hal-02364900, hal-02534911

Contact: Xavier Alameda-pineda

Participants: Mostafa Sadeghi, Xavier Alameda-pineda, Laurent Girin, Simon Leglaive, Radu Horaud

6.1.4 upa3dfa

Name: Unsupervised Performance Analysis of 3D Face Alignment

Keywords: Computer vision, Deep learning, 3D Face alignment, Pattern matching

Scientific Description: We address the problem of analyzing the performance of 3D face alignment (3DFA) algorithms. Traditionally, performance analysis relies on carefully annotated datasets. Here, these annotations correspond to the 3D coordinates of a set of pre-defined facial landmarks. However, this annotation process, be it manual or automatic, is rarely error-free, which strongly biases the analysis. In contrast, we propose a fully unsupervised methodology based on robust statistics and a parametric confidence test. We revisit the problem of robust estimation of the rigid transformation between two point sets and we describe two algorithms, one based on a mixture between a Gaussian and a uniform distribution, and another one based on the generalized Student’s t-distribution. We show that these methods are robust to up to 50% outliers, which makes them suitable for mapping a face, from an unknown pose to a frontal pose, in the presence of facial expressions and occlusions. Using these methods in conjunction with large datasets of face images, we build a statistical frontal facial model and an associated parametric confidence metric, eventually used for performance analysis. We empirically show that the proposed pipeline is neither method-biased nor data-biased, and that it can be used to assess both the performance of 3DFA algorithms and the accuracy of annotations of face datasets

Functional Description: This library contains codes and data for unsupervised performance analysis of 3D face alignment algorithms explained in the following publication : "Sadeghi M, Guy S, Raison A, Alameda-Pineda X, Horaud R (2020) Unsupervised Performance Analysis of 3D Face Alignment. Submitted to International Journal of Computer Vision"

Publication: hal-02543069

Contact: Xavier Alameda-pineda

Participants: Mostafa Sadeghi, Sylvain Guy, Xavier Alameda-pineda

6.1.5 CANUReID

Name: CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification

Keywords: Computer vision, Deep learning, Identification, Unsupervised learning

Scientific Description: Unsupervised person re-ID is the task of identifying people on a target data set for which the ID labels are unavailable during training. In this paper, we propose to unify two trends in unsupervised person re-ID: clustering \& fine-tuning and adversarial learning. On one side, clustering groups training images into pseudo-ID labels, and uses them to fine-tune the feature extractor. On the other side, adversarial learning is used, inspired by domain adaptation, to match distributions from different domains. Since target data is distributed across different camera viewpoints, we propose to model each camera as an independent domain, and aim to learn domain-independent features. Straightforward adversarial learning yields negative transfer, we thus introduce a conditioning vector to mitigate this undesirable effect. In our framework, the centroid of the cluster to which the visual sample belongs is used as conditioning vector of our conditional adversarial network, where the vector is permutation invariant (clusters ordering does not matter) and its size is independent of the number of clusters. To our knowledge, we are the first to propose the use of conditional adversarial networks for unsupervised person re-ID. We evaluate the proposed architecture on top of two state-of-the-art clustering-based unsupervised person re-identification (re-ID) methods on four different experimental settings with three different data sets and set the new state-of-the-art performance on all four of them.

Functional Description: We propose to unify two trends in unsupervised person re-ID: clustering & fine-tuning and adversarial learning. On one side, clustering groups training images into pseudo-ID labels, and uses them to fine-tune the feature extractor. On the other side, adversarial learning is used, inspired by domain adaptation, to match distributions from different domains. Since target data is distributed across different camera viewpoints, we propose to model each camera as an independent domain, and aim to learn domain-independent features. Straightforward adversarial learning yields negative transfer, we thus introduce a conditioning vector to mitigate this undesirable effect. In our framework, the centroid of the cluster to which the visual sample belongs is used as conditioning vector of our conditional adversarial network, where the vector is permutation invariant (clusters ordering does not matter) and its size is independent of the number of clusters. To our knowledge, we are the first to propose the use of conditional adversarial networks for unsupervised person re-ID. We evaluate the proposed architecture on top of two state-of-the-art clustering-based unsupervised person re-identification (re-ID) methods on four different experimental settings with three different data sets and set the new state-of-the-art performance on all four of them.

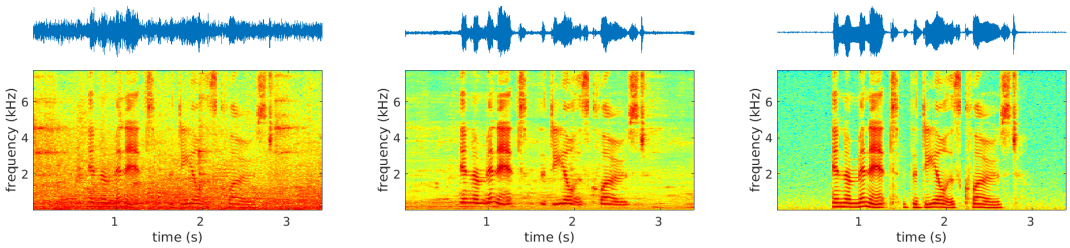
Contact: Xavier Alameda-pineda

Participants: Guillaume Delorme, Yihong Xu, Xavier Alameda-pineda

7 New results

7.1 Speech Denoising and Enhancement with LTSMs

We address the problems of single- and multichannel speech denoising [36] and enhancement [63, 53, 51] in the short-time Fourier transform (STFT) domain and in the framework of sequence-to-sequence deep learning. In the case of denoising, the magnitude of noisy speech is mapped onto the noise power spectral density. In the case of speech enhancement, the noisy speech is mapped onto clean speech. A long short-time memory (LSTM) network takes as input a sequence of STFT coefficients associated with a frequency bin of multichannel noisy-speech signals. The network's output is a sequence of single-channel clean speech at the same frequency bin. We propose several clean-speech network targets, namely, the magnitude ratio mask, the complex ideal ratio mask, the STFT coefficients and spatial filtering [63]. A prominent feature of the proposed model is that the same LSTM architecture, with identical parameters, is trained across frequency bins. The proposed method is referred to as narrow-band deep filtering. This choice stays in contrast with traditional wide-band speech enhancement methods. The proposed deep filter is able to discriminate between speech and noise by exploiting their different temporal and spatial characteristics: speech is non-stationary and spatially coherent while noise is relatively stationary and weakly correlated across channels. This is similar in spirit with unsupervised techniques, such as spectral subtraction and beamforming. We describe extensive experiments with both mixed signals (noise is added to clean speech) and real signals (live recordings). We empirically evaluate the proposed architecture variants using speech enhancement and speech recognition metrics, and we compare our results with the results obtained with several state of the art methods. In the light of these experiments we conclude that narrow-band deep filtering has very good performance, and excellent generalization capabilities in terms of speaker variability and noise type, e.g. Figure 3.



Noisy speech input

Filtered output

Ground-truth speech signal

Figure 3: An example of narrow-band deep filtering for speech enhancement [63]. Waveforms and spectrograms of the noisy (unprocessed) input, the filtered output and the ground-truth clean-speech. Four microphones were used in this example. The signal-to-noise ratio in this example is 0 dB.

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

7.2 Speech Enhancement with a Recurrent Variational Auto-Encoder

We investigate a generative approach to speech enhancement based on a recurrent variational autoencoder (RVAE). The deep generative speech model is trained using clean speech signals only, and it is combined with a nonnegative matrix factorization (NMF) noise model for speech enhancement. We propose a variational expectation-maximization algorithm where the encoder of the RVAE is fine-tuned at test time, to approximate the distribution of the latent variables given the noisy speech observations. Compared with previous approaches based on feed-forward fully-connected architectures, the proposed recurrent deep generative speech model induces a posterior temporal dynamic over the latent variables, which is shown to improve the speech enhancement results. [52].

Website: https://team.inria.fr/perception/research/icassp-2019-mvae/

7.3 Audio-visual Speech Enhancement with Conditional Variational Auto-Encoder

Variational auto-encoders (VAEs) are deep generative latent variable models that can be used for learning the distribution of complex data. VAEs have been successfully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. One advantage of this generative approach is that it does not require pairs of clean and noisy speech signals at training. In this work, we propose audio-visual variants of VAEs for single-channel and speaker-independent speech enhancement. We developed a conditional VAE (CVAE) where the audio speech generative process is conditioned on visual information of the lip region, e.g. Figure 4. At test time, the audio-visual speech generative model is combined with a noise model, based on nonnegative matrix factorization, and speech enhancement relies on a Monte Carlo expectation-maximization algorithm. Experiments were conducted with the recently published NTCD-TIMIT dataset. The results confirm that the proposed audio-visual CVAE effectively fuse audio and visual information, and it improves the speech enhancement performance compared with the audio-only VAE model, especially when the speech signal is highly corrupted by noise. We also showed that the proposed unsupervised audio-visual speech enhancement approach outperforms a state-of-the-art supervised deep learning method [46].

Website: https://team.inria.fr/perception/research/av-vae-se/

7.4 Mixture of Inference Networks for VAE-based Audio-visual Speech Enhancement

We are interested in unsupervised (unknown noise) audio-visual speech enhancement based on variational autoencoders (VAEs), where the probability distribution of clean speech spectra is simulated via an encoder-decoder architecture. The trained generative model (decoder) is then combined with a noise model at test time to estimate the clean speech. In the speech enhancement phase (test time), the initialization of the latent variables, which describe the generative process of clean speech via decoder, is crucial, as the overall inference problem is non-convex. This is usually done by using the output of the

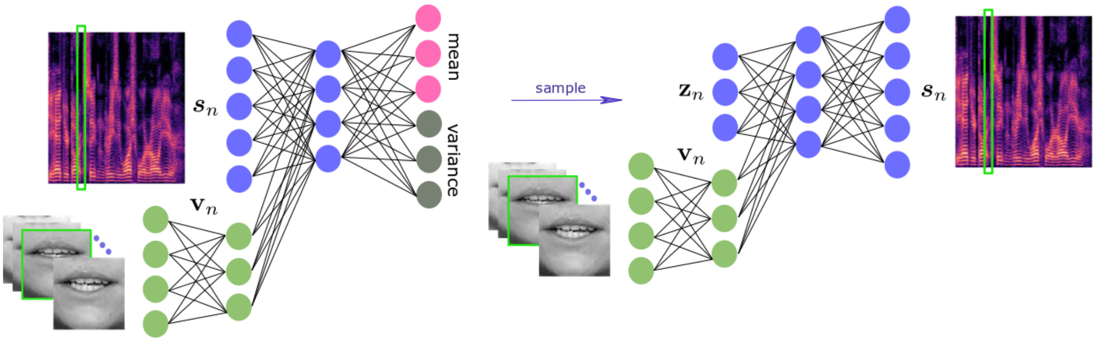


Figure 4: We proposed a conditional variational auto-encoder architecture for fusing audio and visual data for speech enhancement [46].

trained encoder where the noisy audio and clean video data are given as input. Current audio-visual models do not provide an effective initialization because the two modalities are tightly coupled (concatenated) in the associated architectures [46], Figure 4. To overcome this issue, we inspire from mixture models, and introduce the mixture of inference networks variational autoencoder (MIN-VAE), e.g. Figure 5. Two encoder networks input, respectively, audio and visual data, and the posterior of the latent variables is modeled as a mixture of two Gaussian distributions output from each encoder network. The mixture variable is also latent, and therefore the inference of learning the optimal balance between the audio and visual inference networks is unsupervised as well. By training a shared decoder, the overall network learns to adaptively fuse the two modalities. Moreover, at test time, the visual encoder, which takes (clean) visual data, is used for initialization. A variational inference approach is derived to train the proposed generative model. Thanks to the novel inference procedure and the robust initialization, the proposed audio-visual VAE exhibits superior performance on speech enhancement than using the standard audio-only as well as audio-visual counterparts [45, 57, 56].

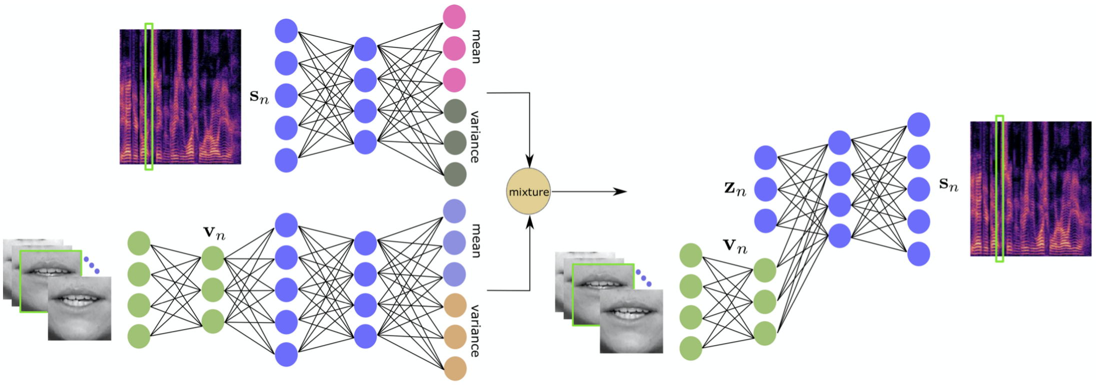


Figure 5: Architecture of the proposed mixture of inference networks VAE (MIN-VAE). A mixture of an audio- and a visual-based encoder is used to approximate the intractable posterior distribution of the latent variables.

7.5 A Comprehensive Analysis of Deep Regression

Deep learning revolutionized data science, and recently its popularity has grown exponentially, as did the amount of papers employing deep networks. Vision tasks, such as human pose estimation, did not escape from this trend. There is a large number of deep models, where small changes in the network architecture, or in the data pre-processing, together with the stochastic nature of the optimization procedures, produce notably different results, making extremely difficult to sift methods that significantly outperform others. This situation motivates the current study, in which we perform a systematic evaluation and statistical analysis of vanilla deep regression, i.e. convolutional neural networks with a linear regression top layer. This is the first comprehensive analysis of deep regression techniques. We perform experiments on four vision problems, and report confidence intervals for the median performance as well as the statistical significance of the results, if any. Surprisingly, the variability due to different data pre-processing procedures generally eclipses the variability due to modifications in the network architecture. Our results reinforce the hypothesis according to which, in general, a general-purpose network (e.g. VGG-16 or ResNet-50) adequately tuned can yield results close to the state-of-the-art without having to resort to more complex and ad-hoc regression models, [44].

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

7.6 Variational Inference and Learning of Piecewise-linear Dynamical Systems

Modeling the temporal behavior of data is of primordial importance in many scientific and engineering fields. Baseline methods assume that both the dynamic and observation equations follow linear-Gaussian models. However, there are many real-world processes that cannot be characterized by a single linear behavior. Alternatively, it is possible to consider a piecewise-linear model which, combined with a switching mechanism, is well suited when several modes of behavior are needed. Nevertheless, switching dynamical systems are intractable because their computational complexity increases exponentially with time. In this paper, we propose a variational approximation of piecewise linear dynamical systems. We provide full details of the derivation of two variational expectation-maximization algorithms, a filter and a smoother. We show that the model parameters can be split into two sets, static and dynamic parameters, and that the former parameters can be estimated off-line together with the number of linear modes, or the number of states of the switching variable. We apply the proposed method to the head-pose tracking, e.g. Figure 6 , and we thoroughly compare our algorithms with several state of the art trackers, [42].

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

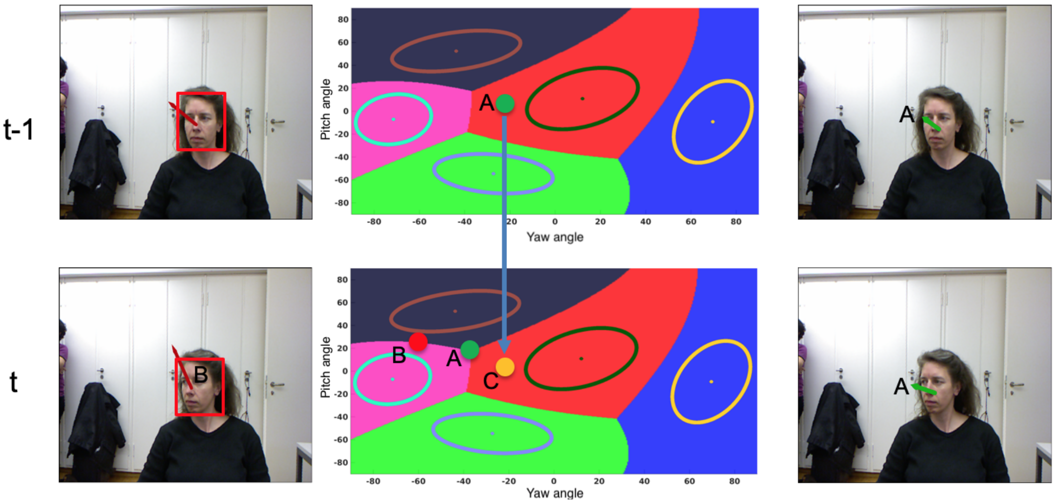
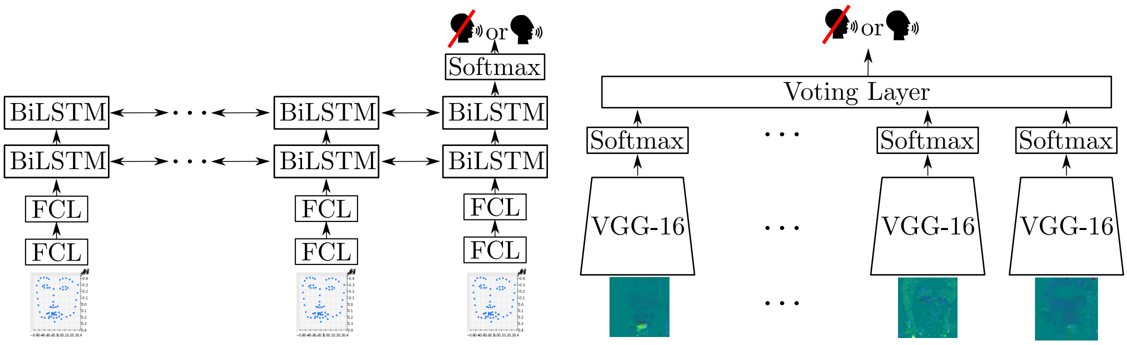


Figure 6: The proposed variational piecewise linear dynamical system (P-LDS) algorithm applied to the problem of head pose tracking (HPT). The central column shows the Gaussian mixture that models the latent space. The parameters of this mixture don't vary over time and they are learnt from a training set of input-output instances of the observed and latent variables. In this example we show the likelihood function associated with the latent variables of head pose, namely the yaw and pitch angles. The observed pose at t (red dot denoted B) is estimated from a high-dimensional feature vector that describes a face (left column). The variational means (green dots denoted A and shown with green arrows onto the right column) are inferred by the E-X step of the algorithm based on the current dynamic prediction (orange dot denoted C) and the current observation (red dot denoted B).



(a) Land-LSTM

(b) OF-ConvNet

Figure 7: Architectures of the two proposed V-VAD models, based on facial landmarks (Land-LSTM) and based on optical flow (OF-ConvNet). Both networks take as input a sequence of frames and predict as output an activity label, *speaking* or *silent*.

7.7 Visual Voice Activity Detection

Visual voice activity detection (V-VAD) uses visual features to predict whether a person is speaking or not. V-VAD is useful whenever audio VAD (A-VAD) is inefficient either because the acoustic signal is difficult to analyze or because it is simply missing. We propose two deep architectures for V-VAD, one based on facial landmarks and one based on optical flow, Figure 7. Moreover, available datasets, used for learning and for testing V-VAD, lack content variability. We introduce a novel methodology to automatically create and annotate very large datasets *in-the-wild* – WildVVAD – based on combining A-VAD with face detection and tracking. A thorough empirical evaluation shows the advantage of training the proposed deep V-VAD models with this dataset, [50].

Website: `https://team.inria.fr/perception/research/vvad/`

7.8 Deep Multi-Object tracking

The recent trend in vision-based multi-object tracking (MOT) is heading towards leveraging the representational power of deep learning to jointly learn to detect and track objects. However, existing methods train only certain sub-modules using loss functions that often do not correlate with established tracking evaluation measures such as Multi-Object Tracking Accuracy (MOTA) and Precision (MOTP). As these measures are not differentiable, the choice of appropriate loss functions for end-to-end training of multi-object tracking methods is still an open research problem. In this paper, we bridge this gap by proposing a differentiable proxy of MOTA and MOTP, which we combine in a loss function suitable for end-to-end training of deep multi-object trackers. As a key ingredient, we propose a Deep Hungarian Net (DHN) module that approximates the Hungarian matching algorithm, Figure 8. DHN allows estimating the correspondence between object tracks and ground truth objects to compute differentiable proxies of MOTA and MOTP, which are in turn used to optimize deep trackers directly. We experimentally demonstrate that the proposed differentiable framework improves the performance of existing multi-object trackers, and we establish a new state of the art on the MOTChallenge benchmark, [58].

Website: `https://github.com/yihongXU/deepMOT`

7.9 Multi-Person Monocular 3D Pose Estimation

Recent literature addressed the monocular 3D pose estimation task very satisfactorily. In these studies, different persons are usually treated as independent pose instances to estimate. However, in many every-day situations, people are interacting, and the pose of an individual depends on the pose of his/her interactees. In this paper, we investigate how to exploit this dependency to enhance current – and possibly future – deep networks for 3D monocular pose estimation. Our pose interacting network, or PI-Net, Figure 9, inputs the initial pose estimates of a variable number of interacting persons into a recurrent

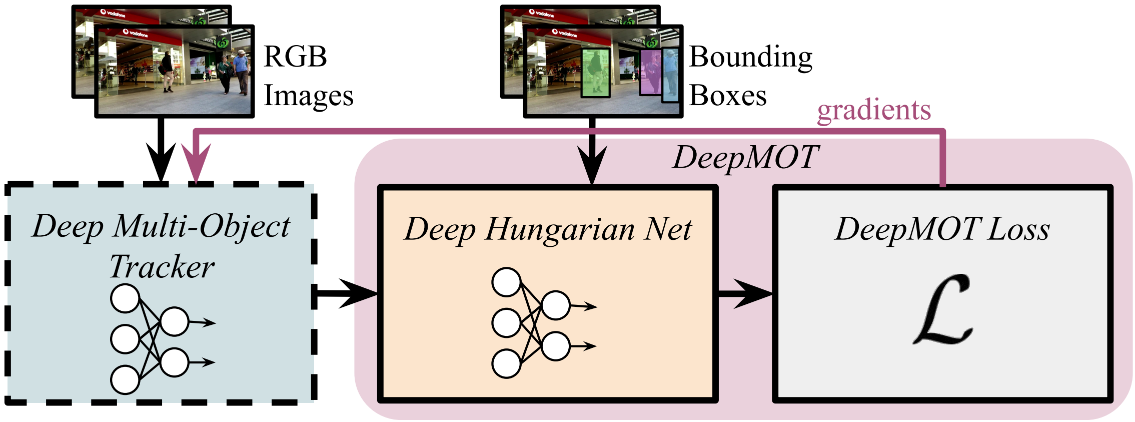


Figure 8: We propose DeepMOT, a general framework for training deep multi-object trackers including the DeepMOT loss that directly correlates with established tracking evaluation measures. The key component in our method is the Deep Hungarian Net (DHN) that provides a soft approximation of the optimal prediction-to-ground-truth assignment, and allows to deliver the gradient, back-propagated from the approximated tracking performance measures, needed to update the tracker weights.

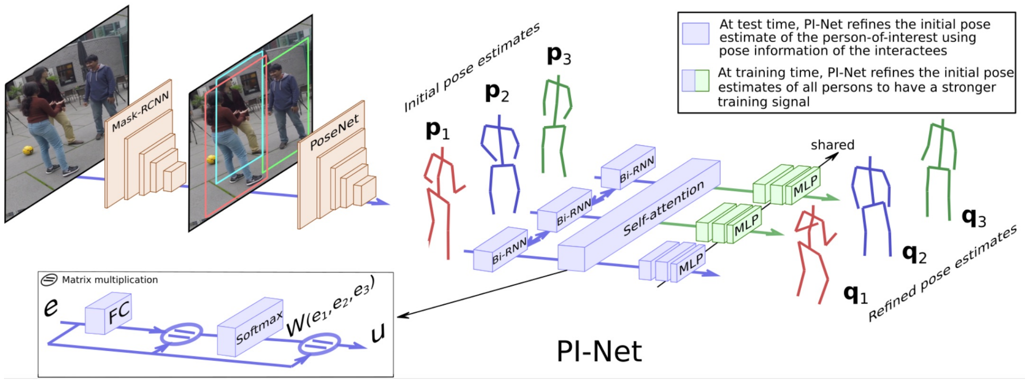


Figure 9: **PI-Net Architecture**. Mask-RCNN and PoseNet methods are used to extract the initial pose estimates. These estimates are fed into PI-Net, composed of three main blocks: Bi-RNN, Self-attention and the shared fully-connected layers. The output of PI-Net refines the initial pose estimates by exploiting the pose of the interacting persons.

architecture used to refine the pose of the person-of-interest. Evaluating such a method is challenging due to the limited availability of public annotated multi-person 3D human pose datasets. We demonstrate the effectiveness of our method in the MuPoTS dataset, setting the new state-of-the-art on it. Qualitative results on other multi-person datasets (for which 3D pose ground-truth is not available) showcase the proposed PI-Net, [49].

7.10 Expression-preserving face frontalization

Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. The main contribution of this paper is a robust frontalization method that preserves non-rigid facial deformations in order to boost the performance of expression analysis from videos of faces, e.g. lip reading. The method iteratively estimates the rigid transformation (scale, rotation, and translation) and the non-rigid deformation between 3D landmarks extracted from an arbitrarily-viewed face, and 3D vertices parameterized by a deformable shape model. An important merit of the method is its ability to deal with non-Gaussian errors in the data. For that purpose, we use the generalized Student-t distribution. The associated EM algorithm estimates a set of weights assigned to the observed landmarks, the higher the

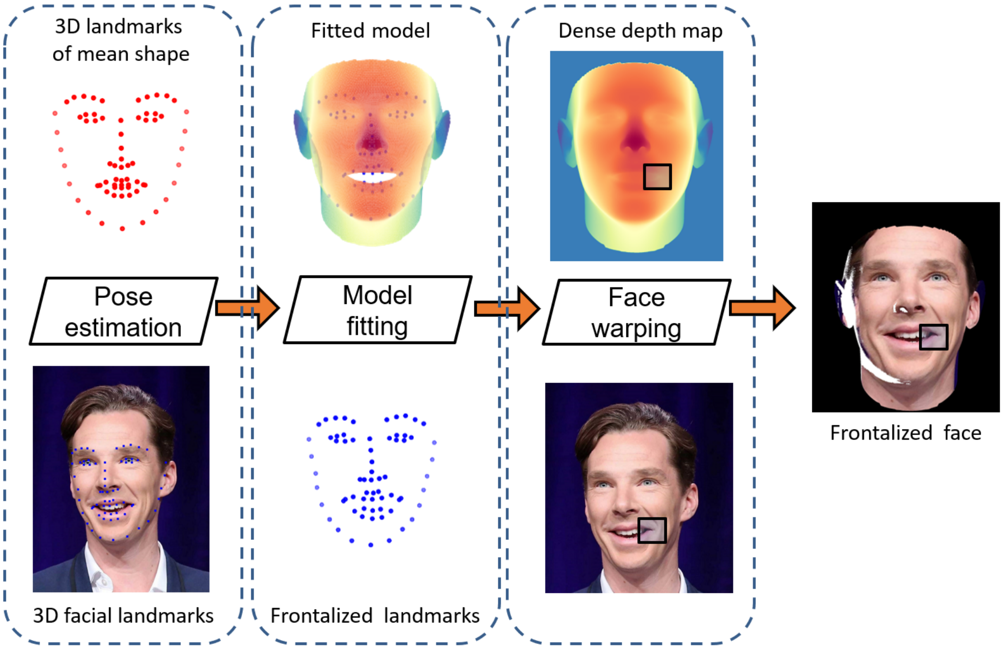


Figure 10: Overview of the proposed method. 3D landmarks extracted from a face (bottom-left) are aligned with 3D vertices associated with a frontal model (top-left). This deformable model is fitted to the frontalized landmarks (bottom-middle), yielding a deformed model aligned with the landmarks (top-middle). A dense depth map is computed by interpolating the 3D vertices of the triangulated mesh of the deformed model (top-right). This depth map is combined with the input face which is warped onto the frontal view (bottom-right).

weight the more important the landmark, thus favoring landmarks that are only affected by rigid head movements. We propose to use the zero-mean normalized cross-correlation (ZNCC) score to evaluate the ability to preserve facial expressions. Moreover, we show that the method, when incorporated into a lip reading pipeline, considerably improves the word recognition score on an in-the-wild benchmark, [62].

https://team.inria.fr/perception/research/rff/

7.11 Conditional Adversarial Network for Person Re-id

Unsupervised person re-ID is the task of identifying people on a target data set for which the ID labels are unavailable during training. In this paper, we propose to unify two trends in unsupervised person re-ID: clustering & fine-tuning and adversarial learning. On one side, clustering groups training images into pseudo-ID labels, and uses them to fine-tune the feature extractor. On the other side, adversarial learning is used, inspired by domain adaptation, to match distributions from different domains. Since target data is distributed across different camera viewpoints, we propose to model each camera as an independent domain, and aim to learn domain-independent features. Straightforward adversarial learning yields negative transfer, we thus introduce a conditioning vector to mitigate this undesirable effect. In our framework, the centroid of the cluster to which the visual sample belongs is used as conditioning vector of our conditional adversarial network, where the vector is permutation invariant (clusters ordering does not matter) and its size is independent of the number of clusters. To our knowledge, we are the first to propose the use of conditional adversarial networks for unsupervised person re-ID. We evaluate the proposed architecture on top of two state-of-the-art clustering-based unsupervised person re-identification (re-ID) methods on four different experimental settings with three different data sets and set the new state-of-the-art performance on all four of them, [48].

Website: https://team.inria.fr/perception/research/canu-reid/

7.12 Online Deep Appearance Network for Identity-Consistent Multi-Person Tracking

The analysis of effective states through time in multi-person scenarii is very challenging, because it requires to *consistently* track all persons over time. This requires a robust visual appearance model capable of re-identifying people already tracked in the past, as well as spotting newcomers. In real-world applications, the appearance of the persons to be tracked is unknown in advance, and therefore on must devise methods that are both discriminative and flexible. Previous work in the literature proposed different tracking methods with fixed appearance models. These models allowed, up to a certain extent, to discriminate between appearance samples of two different people. We propose an online deep appearance network (ODANet), a method able to simultaneously track people and update the appearance model with the newly gathered annotation-less images. Since this task is specially relevant for autonomous systems, we also describe a platform-independent robotic implementation of ODANet. Our experiments show the superiority of the proposed method with respect to the state of the art, and demonstrate the ability of ODANet to adapt to sudden changes in appearance, to integrate new appearances in the tracking system and to provide more identity-consistent tracks [47].

7.13 Socially Aware Robot Navigation

While deep learning has signifantly advanced the state-of-the art in a number of domains, it has also been successful to solve more complex tasks that involve online decision taking and control, e.g. end-to-end learning of self-driving cars. This raises the question of applying such end-to-end frameworks for robot navigation taking into account social constraints. However, compared to self-driving cars, socially-aware navigation does not have such well defined tasks, and the behavior of the other entities (people in our case) present in the scene follows much more complex patterns than in the self-driving car scenarios. Furthermore, data on social interactions is hard to acquire, both for ethical and for practical reasons. Therefore we turn our attention towards more tranditional robot control techniques to address this problem. More precisely, we exploit the well established framework of Model Predictive Control (MPC), and combine it with a social cost map to take into account conversational groups (called F-formations). For example, a group organized in an F-formation shares a private space, called the o-space, reserved to the group that should not be occupied (by a social robot). Another example of social constraints would be to respect privacy spaces of people the robot is interacting with and depending on the level of interaction. These constraints should be integrated in the control of the robot, for example in the form of a velocity map or a social cost map. Our first simulated experiments demonstrate the capabilities of the proposed MPC-based framework to address a variety of generic scenarios (joining a group, guiding or following a person) meaningful for social robotics in general.

7.14 Meta Reinforcement Learning for Robust Action Policies

We have also started investigating the use of deep reinforcement learning for socially acceptable robot action policies. A problematic point of deep reinforcement learning is the amount of data that is required to learn appropriate policies. The agent needs to explore a lot of the state space and to observe the outcomes of different actions to identify the best action per state. In the context of robotics, this learning process takes a long time. Furthermore, the learned behavior depends on the reward function that has to be defined by the user. However, it is often not foreseeable what behavior will result from a reward function. For example, in a navigation task where the robot has to approach a human, the reward function could have a component which punishes strong movements. This should ensure that the robot is not learning a policy that behaves erratic or approaches a human with a too high velocity. How strong this component influences the whole reward function can lead to vastly different behaviors. If the punishment is too small, then the robot might approach a human too fast and is perceived as threatening. If the punishment is too large, then the robot might approach too slow or does not move around obstacles to reach the human. Often, the reward function needs to be adapted to learn an appropriate behavior. For classical reinforcement learning the task would have to be learned from scratch for each new reward function costing a lot of time. As a solution to these problems the SPRING project will utilize transfer learning and meta learning techniques. Our first series of experiments demonstrate the interest of exploiting meta reinforcement learning strategies when combining tasks such as facing the prominent speaker, looking at the people involved in the social interaction and limiting the robot movements (to have a more natural beahvior).

The studies on socially aware robot navigation and meta-reinforcement learning must be confirmed beyond simulation, via a real robotic platform in a real multi-person environment. The pandemic put these experiments on hold, and we are looking forward to evaluating our developments with the physical robotic platform which will soon be available in the team's laboratory, Figure 2.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

8.1.1 VASP

Title: Visually-assisted speech processing

Duration: *1 October 2020 - 30 September 2021*

Principal investigator: *Radu Horaud*

Partner: *Facebook Reality Labs Research, Redmond WA, USA*

Summary: *We investigate audio-visual speech processing. In particular we plan to go beyond the current paradigm that systematically combines a noisy speech signal with clean lip images and which delivers a clean speech signal. The rationale of this paradigm is based on the fact that lip images are free of any type of noise. This hypothesis is merely verified in practice. Indeed, speech production is often accompanied by head motions that considerably modify the patterns of the observed lip movements. As a consequence, currently available audio-visual speech processing technologies are not usable in practice. In this project we develop a methodology that separates non-rigid face- and lip movements from rigid head movements, and we build a deep generative architecture that combines audio and visual features based on their relative merits, rather than making systematic recourse to their concatenation. It is also planned to record and annotate an audio-visual dataset that contains realistic face-to-face and multiparty conversations. The core methodology is based on robust mixture modeling and on variational auto-encoders.*

9 Partnerships and cooperations

9.1 European initiatives

9.1.1 FP7 & H2020 Projects

SPRING

Title: Socially Pertinent Robots in Gerontological Healthcare

Duration: *1 January 2020 - 31 December 2023*

Coordinator: *Xavier Alameda-Pineda, Inria*

Partners:

- BAR ILAN UNIVERSITY (Israel)
- CESKE VYSOKE UCENI TECHNICKE V PRAZE (Czech Republic)
- ERM AUTOMATISMES INDUSTRIELS (France)
- HERIOT-WATT UNIVERSITY (*(missing:COUNTRY)*)
- PAL ROBOTICS SL (Spain)

- UNIVERSITA DEGLI STUDI DI TRENTO (Italy)

Inria contact: *Xavier Alameda-Pineda*

Summary: *SPRING is an EU H2020-ICT research and innovation action (RIA) whose main objective is the development of socially assistive robots with the capacity of performing multimodal multiple-person interaction and open-domain dialogue. SPRING explores new methods at the crossroads of machine learning, computer vision, audio signal processing, spoken dialog and robotics for enhancing the interaction and communication capabilities of companion robots. The paramount application of SPRING is the use of robots in gerontological healthcare.*

9.2 National initiatives

9.2.1 ANR JCJC MLRI

Title: Multi-modal multi-person low-level learning for robot interactions

Duration: *1 March 2020 - 29 February 2024*

Principal investigator: *Xavier Alameda-Pineda, Inria*

Summary: *Robots with autonomous communication capabilities interacting with multiple persons at the same time in the wild are both a societal mirage and a scientific Ithaca. Indeed, despite the presence of various companion robots on the market, their social skills are derived from machine learning techniques functioning mostly under laboratory conditions. Moreover, current robotic platforms operate in confined environments, where on one side, qualified personnel received detailed instructions on how to interact with the robot as part of their technical training, and on the other side, external sensors and actuators may be available to ease the interaction between the robot and the environment. Trespassing these two constraints would allow a robotic platform to freely interact with multiple humans in a wide variety of every-day situations, e.g. as an office assistant, a health-care helper, a janitor or a waiter/waitress. To our understanding, interacting in the wild means that the environment is natural (unscripted conversation, regular lighting and acoustic conditions, people freely moving, etc.) and the robot is self-sufficient (uses only its sensing, acting and computing resources).*

9.2.2 Multidisciplinary Institute of Artificial Intelligence (MIAI)

Title: MIAI chair: Audio-visual machine perception and interaction for companion robots

Duration: *1 October 2019 - 30 September 2023*

Principal investigators: *Xavier Alameda-Pined and Radu Horaud, Inria*

Participants:

- Florence Forbes, Inria
- Jean-Charles Quinton, UGA
- Laurent Girin, Grenoble INP

Summary: *We are particularly interested in the development of a robot able to achieve such tasks as exploring a populated space, understanding human behavior, and engaging multimodal dialog with one or several users. These tasks require audio and visual features (e.g. clean speech, prosody, eye-gaze, head-gaze, facial expressions, lip movements, head movements, and hand gestures) to be robustly retrieved from the raw sensor data. These features cannot be reliably extracted with a static robot that listens, looks and communicates with people from a distance, because of acoustic noise and reverberation, overlapping audio sources, bad lighting, limited image resolution, limited camera field of view, non-frontal views of people, visual occlusions, etc. Audio and visual perception and communication must therefore be performed actively: given a particular task, such as face-to-face* *dialog, the robot should be able to learn how to collect clean data (e.g. frontal videos of faces and audio signals with high speech-to-noise ratios) and how to react appropriately to human verbal and non-verbal solicitations (e.g. taking speech turns in a multi-party conversation). We plan to achieve a fine coupling between scientific findings and technological developments and to demonstrate this with a companion robot that assists and entertains the elderly in healthcare facilities.*

9.2.3 ANR project MUDialbot

Title: MUlti-party perceptually-active situated DIALog for human-roBOT interaction

Duration:

Coordinator: *Fabrice Lefevre, Avignon Unibersity*

Partners:

- Avignon University
- Inria
- Hubert Curien laboratory
- Broca Hospital

Inria contact: *Radu Horaud*

Summary: The overall goal is to actively incorporate human-behavior cues in spoken human-robot communication. We intend to reach a new level in the exploitation of the rich information available with audio and visual data flowing from humans when interacting with robots. In particular, extracting highly informative verbal and non-verbal perceptive features will enhance the robot's decision-making ability such that it can take speech turns more naturally and switch between multi-party/group interactions and face-to-face dialogues where required. Recently there has been an increasing interest in companion robots that are able to assist people in their everyday life and to communicate with them. These robots are perceived as social entities and their utility for healthcare and psychological well being for the elderly has been acknowledged by several recent studies. Patients, their families and medical professionals appreciate the potential of robots, provided that several technological barriers would be overcome in the near future, most notably the ability to move, see and hear in order to naturally communicate with people, well beyond touch screens and voice commands. The scientific and technological results of the project will be implemented onto a commercially available social robot and they will be tested and validated with several use cases in a day-care hospital unit. Large-scale data collection will complement in-situ tests to fuel further researches.

9.2.4 IDEX-UGA PIMPE

Physical complex Interactions and Multi-person Pose Estimation (PIMPE) is an *International Strategic Partnerships* (ISP) three-year project between our team and Universitat Politècnica de Catalunya (UPC). The scientific challenges of PIMPE are the followings: (i) Modeling multi-person interactions in full-body pose estimation, (ii) Estimating human poses in complex multi-person physical interactions, and (iii) Generating controlled and realistic multi-person complex pose images.

Participants: Xavier Alameda-Pineda (PI), Francesc Moreno-Noguer (UPC, Co-PI).

9.2.5 IDEX-UGA MIDGen

Multimodal Interaction Data Generation (MIDGen) is an *Initiatives de Recherche Stratégiques* (IRS) three-year project between our team and Pervasive Interaction team. The scientific challenges of MIDGen is the development of multimodal perception algorithms capable of understanding social signals emitted by humans with a high degree of precision.

Participants: Dominique Vaufreydaz (PI, UGA/LIG), Xavier Alameda-Pineda (co-PI).

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: selection

Xavier Alameda-Pineda was the main co-organiser of the Fairness Accountability Transparency and Ethics in Multimedia workshop, co-located with ACM International Conference on Multimedia 2020.

Member of the conference program committees Xavier Alameda-Pineda was Area Chair for the following conferences:

- ACM International Conference on Multimedia 2020
- IAPR International Conference on Pattern Recognition 2020
- IEEE Winter Conference on Applications of Computer Vision 2021

10.1.2 Journal

Member of the editorial boards Xavier Alameda-Pineda is Associated Editor of the ACM Transactions on Multimedia Tools and Applications.

Reviewer - reviewing activities Xavier Alameda-Pineda reviewed for IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Audio, Language and Signal Processing and for IEEE Transactions on Multimedia.

10.1.3 Invited talks

Xavier Alameda-Pineda was invited to give the following talks:

- Towards audio-visual speech enhancement in robotic platforms (Dec'20) at Journée "perception et interaction homme-robot" du Groupe de Travail GT5 Interactions Personnes / Systèmes Robotiques du GDR Robotique
- Audio-visual variational speech enhancement (Sep'20) at Intelligent Sensing Summer School
- Choosing wisely your deep training loss (March'20) at Universidade NOVA de Lisboa
- Artificial Intelligence for Social Robots in Gerontological Healthcare (March'20) at European Robotics Forum

10.1.4 Teaching

- Master : Xavier Alameda-Pineda, Fundamentals of Probabilistic Data Mining, 19.5h, M2, UGA, France.
- Master : Xavier Alameda-Pineda, Machine Learning for Computer Vision and Audio Processing, 12h, M2, UGA, France.

10.1.5 Supervision

- PhD in progress: Guillaume Delorme, Deep Person Re-identification, October 2017, Xavier Alameda-Pineda and Radu Horaud,
- PhD in progress: Yihong Xu, Deep Multiple-person Tracking, October 2018, Xavier Alameda-Pineda and Radu Horaud,
- PhD in progress: Wen Guo, Deep Human Pose, October 2019, Xavier Alameda-Pineda and Radu Horaud,

- PhD in progress: Anand Ballou, Deep Reinforcement Learning for Robot Control, November 2019, Xavier Alameda-Pineda and Radu Horaud,
- PhD in progress: Louis Airale, Data Generation for Deep Multimodal Interaction Algorithms, October 2019, Xavier Alameda-Pineda and Dominique Vaufreydaz,
- PhD in progress: Xiaoyu Bie, Deep Generative Methods for Audio and Vision, December 2019, Xavier Alameda-Pineda and Laurent Girin.
- PhD in progress: Gaetan Lepage, Deep Reinforcement Learning for Robot Perception Enhancement, October 2020, Xavier Alameda-Pineda and Laurent Girin.
- PhD in progress: Xiaoyu Lin, Deep Generative Methods for Multi-Person Multi-Modal Tracking, November 2020, Xavier Alameda-Pineda and Laurent Girin.

10.1.6 Juries

Xavier Alameda-Pineda participated to the following PhD Juries as “rapporteur”:

- Daniel Michelsanti, University of Aalborg. Supervisors: Zheng-Hua Tan and Jesper Jensen.

11 Scientific production

11.1 Major publications

[1] X. Alameda-Pineda and R. Horaud. ‘A Geometric Approach to Sound Source Localization from Time-Delay Estimates’. In: *IEEE Transactions on Audio, Speech and Language Processing* 22.6 (June 2014), pp. 1082–1095. DOI: 10.1109/TASLP.2014.2317989. URL: https://hal.inria.fr/hal-00975293.

[2] X. Alameda-Pineda and R. Horaud. ‘Vision-Guided Robot Hearing’. In: *International Journal of Robotics Research* 34.4–5 (Apr. 2015), pp. 437–456. DOI: 10.1177/0278364914548050. URL: https://hal.inria.fr/hal-00990766.

[3] X. Alameda-Pineda, E. Ricci and N. Sebe. *Multimodal behavior analysis in the wild: Advances and challenges*. Academic Press (Elsevier), Dec. 2018. URL: https://hal.inria.fr/hal-01858395.

[4] N. Andreff, B. Espiau and R. Horaud. ‘Visual Servoing from Lines’. In: *International Journal of Robotics Research* 21.8 (2002), pp. 679–700. URL: http://hal.inria.fr/hal-00520167.

[5] S. Ba, X. Alameda-Pineda, A. Xompero and R. Horaud. ‘An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes’. In: *Computer Vision and Image Understanding* 153 (Dec. 2016), pp. 64–76. DOI: 10.1016/j.cviu.2016.07.006. URL: https://hal.inria.fr/hal-01349763.

[6] Y. Ban, X. Alameda-Pineda, F. Badeig, S. Ba and R. Horaud. ‘Tracking a Varying Number of People with a Visually-Controlled Robotic Head’. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vancouver, Canada, Sept. 2017. URL: https://hal.inria.fr/hal-01542987.

[7] Y. Ban, X. Alameda-Pineda, C. Evers and R. Horaud. ‘Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM’. In: *IEEE Signal Processing Letters* 26.6 (June 2019), pp. 798–802. DOI: 10.1109/LSP.2019.2908376. URL: https://hal.inria.fr/hal-01969050.

[8] F. Cuzzolin, D. Mateus and R. Horaud. ‘Robust Temporally Coherent Laplacian Protrusion Segmentation of 3D Articulated Bodies’. In: *International Journal of Computer Vision* 112.1 (Mar. 2015), pp. 43–70. DOI: 10.1007/s11263-014-0754-0. URL: https://hal.archives-ouvertes.fr/hal-01053737.

[9] A. Deleforge, F. Forbes and R. Horaud. ‘Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds’. In: *International Journal of Neural Systems* 25.1 (Feb. 2015), p. 21. DOI: 10.1142/S0129065714400036. URL: https://hal.inria.fr/hal-00960796.

[10] A. Deleforge, F. Forbes and R. Horaud. 'High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables'. In: *Statistics and Computing* 25.5 (Sept. 2015), pp. 893–911. DOI: 10.1007/s11222-014-9461-5. URL: https://hal.inria.fr/hal-00863468.

[11] A. Deleforge, R. Horaud, Y. Y. Schechner and 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* 23.4 (Apr. 2015), pp. 718–731. DOI: 10.1109/TASLP.2015.2405475. URL: https://hal.inria.fr/hal-01112834.

[12] V. Drouard, R. Horaud, A. Deleforge, S. Ba and G. Evangelidis. 'Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions'. In: *IEEE Transactions on Image Processing* 26.3 (Mar. 2017), pp. 1428–1440. DOI: 10.1109/TIP.2017.2654165. URL: https://hal.inria.fr/hal-01413406.

[13] G. Evangelidis, M. Hansard and R. Horaud. 'Fusion of Range and Stereo Data for High-Resolution Scene-Modeling'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.11 (Nov. 2015), pp. 2178–2192. DOI: 10.1109/TPAMI.2015.2400465. URL: https://hal.archives-ouvertes.fr/hal-01110031.

[14] G. Evangelidis and R. Horaud. 'Joint Alignment of Multiple Point Sets with Batch and Incremental Expectation-Maximization'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (June 2018). https://arxiv.org/abs/1609.01466, pp. 1397–1410. DOI: 10.1109/TPAMI.2017.2717829. URL: https://hal.inria.fr/hal-01413414.

[15] I. Gebru, S. Ba, X. Li and R. Horaud. 'Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.5 (July 2018). https://arxiv.org/abs/1603.09725, pp. 1086–1099. DOI: 10.1109/TPAMI.2017.2648793. URL: https://hal.inria.fr/hal-01413403.

[16] I. D. Gebru, X. Alameda-Pineda, F. Forbes and R. Horaud. 'EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.12 (Dec. 2016), pp. 2402–2415. DOI: 10.1109/TPAMI.2016.2522425. URL: https://hal.inria.fr/hal-01261374.

[17] M. Hansard, G. Evangelidis, Q. Pelorson and R. Horaud. 'Cross-Calibration of Time-of-flight and Colour Cameras'. In: *Computer Vision and Image Understanding* 134 (Apr. 2015), pp. 105–115. DOI: 10.1016/j.cviu.2014.09.001. URL: https://hal.inria.fr/hal-01059891.

[18] M. Hansard and R. Horaud. 'A Differential Model of the Complex Cell'. In: *Neural Computation* 23.9 (Sept. 2011), pp. 2324–2357. DOI: 10.1162/NECO_a_00163. URL: http://hal.inria.fr/inria-00590266.

[19] M. Hansard and R. Horaud. 'Cyclopean geometry of binocular vision'. In: *Journal of the Optical Society of America A* 25.9 (Sept. 2008), pp. 2357–2369. DOI: 10.1364/JOSAA.25.002357. URL: http://hal.inria.fr/inria-00435548.

[20] M. Hansard and R. Horaud. 'Cyclorotation Models for Eyes and Cameras'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 40.1 (Mar. 2010), pp. 151–161. DOI: 10.1109/TSMCB.2009.2024211. URL: http://hal.inria.fr/inria-00435549.

[21] M. Hansard, R. Horaud, M. Amat and G. Evangelidis. 'Automatic Detection of Calibration Grids in Time-of-Flight Images'. In: *Computer Vision and Image Understanding* 121 (Apr. 2014), pp. 108–118. DOI: 10.1016/j.cviu.2014.01.007. URL: https://hal.inria.fr/hal-00936333.

[22] M. Hansard, S. Lee, O. Choi and R. Horaud. *Time of Flight Cameras: Principles, Methods, and Applications*. Springer Briefs in Computer Science. Springer, Oct. 2012, p. 95. URL: http://hal.inria.fr/hal-00725654.

[23] R. Horaud, G. Csurka and D. Demirdjian. 'Stereo Calibration from Rigid Motions'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.12 (Dec. 2000), pp. 1446–1452. DOI: 10.1109/34.895977. URL: http://hal.inria.fr/inria-00590127.

[24] R. Horaud, F. Forbes, M. Yguel, G. Dewaele and J. Zhang. 'Rigid and Articulated Point Registration with Expectation Conditional Maximization'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.3 (Mar. 2011), pp. 587–602. DOI: 10.1109/TPAMI.2010.94. URL: http://hal.inria.fr/inria-00590265.

[25] R. Horaud, M. Niskanen, G. Dewaele and 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* 31.1 (Jan. 2009), pp. 158–163. DOI: 10.1109/TPAMI.2008.108. URL: http://hal.inria.fr/inria-00446898.

[26] V. Khalidov, F. Forbes and R. Horaud. 'Conjugate Mixture Models for Clustering Multimodal Data'. In: *Neural Computation* 23.2 (Feb. 2011), pp. 517–557. DOI: 10.1162/NECO_a_00074. URL: http://hal.inria.fr/inria-00590267.

[27] D. Knossow, R. Ronfard and R. Horaud. 'Human Motion Tracking with a Kinematic Parameterization of Extremal Contours'. In: *International Journal of Computer Vision* 79.3 (Sept. 2008), pp. 247–269. DOI: 10.1007/s11263-007-0116-2. URL: http://hal.inria.fr/inria-00590247.

[28] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot and 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* 24.8 (Aug. 2016), pp. 1408–1423. DOI: 10.1109/TASLP.2016.2554286. URL: https://hal.inria.fr/hal-01301762.

[29] S. Lathuilière, B. Massé, P. Mesejo and R. Horaud. 'Neural Network Based Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction'. In: *Pattern Recognition Letters* 118 (Feb. 2019), pp. 61–71. DOI: 10.1016/j.patrec.2018.05.023. URL: https://hal.inria.fr/hal-01643775.

[30] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud. 'Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments'. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (Mar. 2019), pp. 88–103. DOI: 10.1109/JSTSP.2019.2903472. URL: https://hal.inria.fr/hal-01851985.

[31] X. Li, L. Girin, F. Badeig and 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*. IEEE. Daejeon, South Korea: IEEE, Oct. 2016, pp. 2819–2826. DOI: 10.1109/IROS.2016.7759437. URL: https://hal.inria.fr/hal-01349771.

[32] X. Li, L. Girin, S. Gannot and R. Horaud. 'Multichannel Online Dereverberation based on Spectral Magnitude Inverse Filtering'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 27.9 (May 2019), pp. 1365–1377. DOI: 10.1109/TASLP.2019.2919183. URL: https://hal.inria.fr/hal-01969041.

[33] X. Li, L. Girin, S. Gannot and R. Horaud. 'Multichannel Speech Separation and Enhancement Using the Convolutive Transfer Function'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 27.3 (Mar. 2019), pp. 645–659. DOI: 10.1109/TASLP.2019.2892412. URL: https://hal.inria.fr/hal-01799809.

[34] X. Li, L. Girin, R. Horaud and 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* 24.11 (Nov. 2016), pp. 2171–2186. DOI: 10.1109/TASLP.2016.2598319. URL: https://hal.inria.fr/hal-01349691.

[35] X. Li, L. Girin, R. Horaud and S. Gannot. 'Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25.10 (Oct. 2017). 16 pages, 4 figures, 4 tables, pp. 1997–2012. DOI: 10.1109/TASLP.2017.2740001. URL: https://hal.inria.fr/hal-01413417.

[36] X. Li, S. Leglaive, L. Girin and R. Horaud. 'Audio-noise Power Spectral Density Estimation Using Long Short-term Memory'. In: *IEEE Signal Processing Letters* 26.6 (June 2019), pp. 918–922. DOI: 10.1109/LSP.2019.2911879. URL: https://hal.inria.fr/hal-02100059.

[37] B. Massé, S. Ba and R. Horaud. 'Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.11 (Nov. 2018). https://arxiv.org/abs/1703.04727, pp. 2711–2724. DOI: 10.1109/TPAMI.2017.2782819. URL: https://hal.inria.fr/hal-01511414.

[38] M. Sapienza, M. Hansard and R. Horaud. 'Real-time Visuomotor Update of an Active Binocular Head'. In: *Autonomous Robots* 34.1 (Jan. 2013), pp. 33–45. DOI: 10.1007/s10514-012-9311-2. URL: http://hal.inria.fr/hal-00768615.

[39] A. Zaharescu, E. Boyer and R. Horaud. 'Keypoints and Local Descriptors of Scalar Functions on 2D Manifolds'. In: *International Journal of Computer Vision* 100.1 (Oct. 2012), pp. 78–98. DOI: 10.1007/s11263-012-0528-5. URL: http://hal.inria.fr/hal-00699620.

[40] A. Zaharescu, E. Boyer and R. Horaud. 'Topology-Adaptive Mesh Deformation for Surface Evolution, Morphing, and Multi-View Reconstruction'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.4 (Apr. 2011), pp. 823–837. DOI: 10.1109/TPAMI.2010.116. URL: http://hal.inria.fr/inria-00590271.

[41] A. Zaharescu and R. Horaud. 'Robust Factorization Methods Using A Gaussian/Uniform Mixture Model'. In: *International Journal of Computer Vision* 81.3 (Mar. 2009), pp. 240–258. DOI: 10.1007/s11263-008-0169-x. URL: http://hal.inria.fr/inria-00446987.

11.2 Publications of the year

International journals

[42] X. Alameda-Pineda, V. Drouard and R. Horaud. 'Variational Inference and Learning of Piecewise-linear Dynamical Systems'. In: *IEEE Transactions on Neural Networks and Learning Systems* (21st Jan. 2021). DOI: 10.1109/TNNLS.2021.3054407. URL: https://hal.archives-ouvertes.fr/hal-02745527.

[43] Y. Ban, X. Alameda-Pineda, L. Girin and R. Horaud. 'Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.5 (3rd May 2021), pp. 1761–1776. DOI: 10.1109/TPAMI.2019.2953020. URL: https://hal.inria.fr/hal-01950866.

[44] S. Lathuilière, P. Mesejo, X. Alameda-Pineda and R. Horaud. 'A Comprehensive Analysis of Deep Regression'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (1st Sept. 2020), pp. 2065–2081. DOI: 10.1109/TPAMI.2019.2910523. URL: https://hal.inria.fr/hal-01754839.

[45] M. Sadeghi and X. Alameda-Pineda. 'Mixture of Inference Networks for VAE-based Audio-visual Speech Enhancement'. In: *IEEE Transactions on Signal Processing* (9th Mar. 2021). URL: https://hal.inria.fr/hal-02926172.

[46] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin and R. Horaud. 'Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 28 (30th May 2020), pp. 1788–1800. DOI: 10.1109/TASLP.2020.3000593. URL: https://hal.inria.fr/hal-02364900.

International peer-reviewed conferences

[47] G. Delorme, Y. Ban, G. Sarrazin and X. Alameda-Pineda. 'ODANet: Online Deep Appearance Network for Identity-Consistent Multi-Person Tracking'. In: ICPR 2021 - 25th International Conference on Pattern Recognition / Workshops. Milano / Virtual, Italy: https://www.micc.unifi.it/icpr2020/, 10th Jan. 2021. URL: https://hal.inria.fr/hal-03188744.

[48] G. Delorme, Y. Xu, S. Lathuilière, R. Horaud and X. Alameda-Pineda. 'CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification'. In: International Conference on Pattern Recognition. Milano, Italy, 10th Jan. 2021. URL: https://hal.inria.fr/hal-02882285.

[49] W. Guo, E. Corona, F. Moreno-Noguer and X. Alameda-Pineda. 'PI-Net: Pose Interacting Network for Multi-Person Monocular 3D Pose Estimation'. In: WACV 2021 - IEEE Winter Conference on Applications of Computer vision. Waikoloa, United States, 5th Jan. 2021, pp. 1–11. URL: https://hal.inria.fr/hal-02971754.

[50] S. Guy, S. Lathuilière, P. Mesejo and R. Horaud. 'Learning Visual Voice Activity Detection with an Automatically Annotated Dataset'. In: International Conference on Pattern Recognition. Milano, Italy, 10th Jan. 2021. URL: https://hal.inria.fr/hal-02882229.

[51] X. Hao, X. Su, R. Horaud and X. Li. 'FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement'. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Toronto, Canada, 6th June 2021, pp. 1–5. URL: https://hal.archives-ouvertes.fr/hal-03135727.

[52] S. Leglaive, X. Alameda-Pineda, L. Girin and R. Horaud. 'A Recurrent Variational Autoencoder for Speech Enhancement'. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelone, Spain: https://2020.ieeeicassp.org/, 4th May 2020, pp. 1–7. DOI: 10.1109/ICASSP40776.2020.9053164. URL: https://hal.archives-ouvertes.fr/hal-02329000.

[53] X. Li and R. Horaud. 'Online Monaural Speech Enhancement Using Delayed Subband LSTM'. In: Interspeech 2020. Shangai, China, 27th July 2020, pp. 2462–2466. DOI: 10.21437/Interspeech.2020-2091. URL: https://hal.inria.fr/hal-02907455.

[54] Y. Liu, M. De Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe and B. Lepri. 'Describe What to Change: A Text-guided Unsupervised Image-to-Image Translation Approach'. In: 28th ACM International Conference on Multimedia, MM'20. Seatle, United States, 12th Oct. 2020, pp. 1357–1365. DOI: 10.1145/3394171.3413505. URL: https://hal.inria.fr/hal-02930687.

[55] J. Parsa, M. Sadeghi, M. Babaie-Zadeh and C. Jutten. 'Low Mutual and Average Coherence Dictionary Learning Using Convex Approximation'. In: ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelone (virtual), Spain: https://2020.ieeeicassp.org/, 4th May 2020, pp. 3417–3421. DOI: 10.1109/ICASSP40776.2020.9052901. URL: https://hal.inria.fr/hal-02560161.

[56] M. Sadeghi and X. Alameda-Pineda. 'Robust Unsupervised Audio-visual Speech Enhancement Using a Mixture of Variational Autoencoders'. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 4th May 2020. DOI: 10.1109/ICASSP40776.2020.9053730. URL: https://hal.archives-ouvertes.fr/hal-02534911.

[57] M. Sadeghi and X. Alameda-Pineda. 'Switching Variational Auto-Encoders for Noise-Agnostic Audio-visual Speech Enhancement'. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: https://www.2021.ieeeicassp.org/, 6th June 2021, pp. 1–5. URL: https://hal.inria.fr/hal-03155445.

[58] Y. Xu, A. Osep, Y. Ban, R. Horaud, L. Leal-Taixé and X. Alameda-Pineda. 'How To Train Your Deep Multi-Object Tracker'. In: IEEE Conference on Computer Vision and Pattern Recognition. Seattle WA, United States, 14th June 2020, pp. 1–14. DOI: 10.1109/CVPR42600.2020.00682. URL: https://hal.archives-ouvertes.fr/hal-02534894.

Doctoral dissertations and habilitation theses

[59] X. Alameda-Pineda. 'Towards Probabilistic Generative Models for Socially Intelligent Robots'. Université Grenoble - Alpes, 14th Dec. 2020. URL: https://hal.inria.fr/tel-03192456.

Reports & preprints

[60] L. Airale, D. Vaufreydaz and X. Alameda-Pineda. *SocialInteractionGAN: Multi-person Interaction Sequence Generation*. 1st Mar. 2021. URL: https://hal.inria.fr/hal-03163467.

[61] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber and X. Alameda-Pineda. *Dynamical Variational Autoencoders: A Comprehensive Review*. 31st Aug. 2020. URL: https://hal.inria.fr/hal-02926215.

[62] Z. Kang, M. Sadeghi and R. Horaud. *Face Frontalization Based on Robustly Fitting a Deformable Shape Model to 3D Landmarks*. 27th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02980346.

[63] X. Li and R. Horaud. *Narrow-band Deep Filtering for Multichannel Speech Enhancement*. 23rd Sept. 2020. URL: https://hal.inria.fr/hal-02378413.

[64] V.-N. Nguyen, M. Sadeghi, E. Ricci and X. Alameda-Pineda. *Deep Variational Generative Models for Audio-visual Speech Separation*. 4th Sept. 2020. URL: https://hal.inria.fr/hal-02930662.

[65] M. Sadeghi, S. Guy, A. Raison, X. Alameda-Pineda and R. Horaud. *Unsupervised Performance Analysis of 3D Face Alignment*. 16th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02543069.