RESEARCH CENTRE

Grenoble - Rhône-Alpes

IN PARTNERSHIP WITH:

Université de Grenoble Alpes

2021

ACTIVITY REPORT

Project-Team

ROBOTLEARN

Learning, perception and control for social robots

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Contents

Project-Team ROBOTLEARN

Creation of the Project-Team: 2021 July 01

Keywords

Computer sciences and digital sciences

A5.4.2. – Activity recognition

A5.4.5. – Object tracking and motion analysis

A5.7.3. – Speech

A5.7.4. – Analysis

A5.10.2. – Perception

A5.10.4. – Robot control

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

A5.10.7. – Learning

A9.2. – Machine learning

A9.3. – Signal analysis

A9.5. – Robotics

Other research topics and application domains

B2. – Health

B5.6. – Robotic systems

1 Team members, visitors, external collaborators

Research Scientists

- Xavier Alameda-Pineda [Team leader, Inria, Researcher, from Jul 2021, HDR]
- Patrice Horaud [Inria, Senior Researcher, from Jul 2021, HDR]
- Chris Reinke [Inria, Starting Research Position, from Jul 2021]
- Timothee Wintz [Inria, Starting Research Position, from Jul 2021]

PhD Students

- Louis Airale [Univ Grenoble Alpes, from Jul 2021]
- Anand Ballou [Univ Grenoble Alpes, from Jul 2021]
- Xiaoyu Bie [Univ Grenoble Alpes, from Jul 2021]
- Guillaume Delorme [Inria, from Jul 2021]
- Wen Guo [Université polytechnique de Catalogne - Espagne, from Jul 2021]
- Gaetan Lepage [Inria, from Jul 2021]
- Xiaoyu Lin [Inria, from Jul 2021]
- Yihong Xu [Inria, from Jul 2021]

Technical Staff

- Soraya Arias [Inria, Engineer, from Jul 2021]
- Alex Auternaud [Inria, Engineer, from Jul 2021]
- Luis Gomez Camara [Inria, Engineer, from Jul 2021]
- Zhiqi Kang [Inria, Engineer, from Jul 2021 until Sep 2021]
- Matthieu Py [Inria, Engineer, from Jul 2021]
- Nicolas Turro [Inria, Engineer, July 2021]

Interns and Apprentices

- Natanael Dubois–Quilici [Inria, from Jul 2021 until Sep 2021]
- David Emukpere [Inria, from Jul 2021 until Aug 2021]

Administrative Assistant

- Nathalie Gillot [Inria, from Jul 2021]

Visiting Scientists

- Timothee Dhaussy [Institut supérieur d'électronique de Paris, Sep 2021]
- Hanyu Xuan [Université normale de la Chine de l'Est (ECNU) Shanghai, from Jul 2021]

External Collaborator

- Laurent Girin [Institut polytechnique de Grenoble, from Jul 2021, HDR]

2 Overall objectives

In recent years, social robots have been introduced into public spaces, such as museums, airports, commercial malls, banks, show-rooms, schools, universities, hospitals, and retirement homes, to mention a few examples. In addition to classical robotic skills such as navigating in complex environments, grasping and manipulating objects, i.e. *physical interactions*, social robots must be able to communicate with people and to adopt appropriate behavior. Welcoming newcomers, providing various pieces of information, and entertaining groups of people are typical services that social robots are expected to provide in the near future.

Prominent examples of this type of robots, with great scientific, technological, economical and social impact, are *Socially Assistive Robots* (SARs). SARs are likely to play an important role in healthcare and psychological well-being, in particular during non-medical phases inherent to any hospital process [97, 82, 87, 96]. It is well established that properly handling patients during these phases is of paramount importance, as crucial as the medical phases. It is worth to be noticed that non-medical phases represent a large portion of the total hospitalization time. It has been acknowledged that SARs could be well suited for explaining complex medical concepts to patients with limited health literacy [38]. They can coordinate with medical staff and potentially reduce the amount of human resources required for instructing each individual patient [58]. There is a consensus among physicians and psychotherapists that the use of robots in group settings has a positive impact on health, such as decreased stress and loneliness, and improved mood and sociability [39, 29]. Therefore, one can confidently assert that social-robot research is likely to have a great potential for healthcare and that robot companionship is likely to improve both psychological well-being and the relationship between patients and hospital professionals. Beyond healthcare, socially intelligent robots will have applications in education, retail, public relationship and communication, etc. Thanks to the collaboration with industrial partners we can expect direct impact in tourism (PAL Robotics) and education (ERM Automatismes Industriels).

Nevertheless, today's state-of-the-art in robotics is not well-suited to fulfill these needs. Indeed, social-robot platforms that are currently available, whether laboratory prototypes or commercial systems, are based on interface technologies borrowed from smartphones, namely touch-screens and voice commands. This creates two bottlenecks: (i) it limits the use of robots to a handful of simple scenarios which leads to (ii) social robots not being well accepted by a large percentage of users such as the elderly. While there are research programs and projects which have tackled some of these challenges, existing commercially available robots cannot (or only to a very limited extent) recognize individual behaviors (e.g. facial expressions, hand- and body-gestures, head- and eye-gaze) or group behaviors (e.g. who looks at whom, who speaks to whom, who needs robot assistance, etc.). They cannot distinguish between patients, family members, and carers in order to adopt proper attitudes and to exchange adequate pieces of information. They do not have the ability to take social (or non-verbal) signals into account while they are engaged in spoken dialogue and they cannot connect the dialogue with the persons and objects that are physically present in their surroundings. These limitations are largely due to the fact that human-robot interaction technologies are based on algorithms that have been designed for reactive single-user dialog, mostly based on keyword spotting where the robot waits to be instructed what to do based on a limited set of scripted actions. In some cases, the user even has to resort to a handheld microphone or smartphone to overcome the limitations of the built-in microphones and speech recognition systems. We would like to develop robots that are responsible for their perception, and act to enhance the quality of the signals they receive, instead of asking the users to adapt their behavior to the robotic platform.

The scientific ambition of ROBOTLEARN *is to train robots to acquire the capacity to* ***look****,* ***listen****,* ***learn****,* ***move*** *and* ***speak*** *in a socially acceptable manner.*

The scientific ambition of ROBOTLEARN, outlined above, may be broken down into the following three objectives:

1. Develop deep probabilistic models and methods that allow the fusion of audio and visual data, possibly sequential, recorded with cameras and microphones, and in particular with sensors onboard of robots.

2. Increase the performance of human behaviour understanding using deep probabilistic models and jointly exploiting auditory and visual information.

3. Learn robot-action policies that are socially acceptable and that enable robots to better perceive humans and the physical environment.

This will require several new scientific and technological developments. The scientific objectives of ROBOTLEARN stand at the cross-roads of several topics: computer vision, audio signal processing, speech technology, statistical learning, deep learning, and robotics. In partnership with several companies (e.g. PAL Robotics and ERM Automatismes Industriels), the technological objective is to 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 is to validate the scientific and technological progress in the real world. Furthermore, we believe that ROBOTLEARN will contribute with tools and methods able to process robotic data (perception and action signals) in such a way that connections with more abstract representations (semantics, knowledge) are possible. The developments needed to discover and use such connections could be addressed through collaborations. Similarly, aspects related to robot deployment in the consumer world, such as ethics and acceptability will be addressed in collaboration, for instance, with the Broca day-care hospital in Paris.

ROBOTLEARN will build on the scientific expertise that has been developed over the past years by the Perception team. The main emphasis of the Perception team has been the development of audio-visual machine perception, from fundamental principles to the implementation of human-robot interaction algorithms and of software based on these principles.

Particular emphasis has been put on statistical learning and inference principles, and their implementation in terms of practical solvers. For the past five years, the following problems were addressed: separation and localization of multiple (static or moving) audio sources, speech enhancement and separation4, multiple person tracking using visual, audio, and audio-visual observations, head-pose and eye-gaze estimation and tracking for understanding human-human and human-robot social interactions, and visually- and audio-guided robot control.

The formulations of choice have been *latent variable mixture models*, *dynamic Bayesian networks* (DBNs), and their extensions. Robust mixture models were developed, e.g. for clustering audio-visual data [56], for modeling the acoustic-articulatory tract [57], or for registering multiple point sets [50]. DBNs may well be viewed as hybrid state-space models, i.e. models that combine continuous and discrete latent variables. DBNs often lead to intractable maximum a posteriori (MAP) problems. For this reason, approximate inference has been thoroughly investigated. In particular a number of variational expectation maximization (VEM) algorithms were developed, such as high-dimensional regression with latent output [45] and with spatial Markov dependencies [43], sound-source separation and localization [44], [63], multiple person tracking using visual observations [35], audio observations [37, 72], or audio-visual fusion [55, 15], head-pose and eye-gaze estimation and tracking [48], [83], [14]. Variational approximation has also been combined with generative deep neural networks for audio-visual speech enhancement [93]. Very recently, we have reviewed the literature on deep probabilistic sequential modeling, and proposed a model class called dynamical variational autoencoders, see [16] (preprint).

In parallel, we addressed the problems of speech localization, speech separation and speech enhancement in reverberant environments. This is an extremely important topic in the framework of robot audition. Nonetheless, the formulation that we proposed and the associated algorithms can be used in the general case of multi-channel audio signal processing in adverse acoustic conditions. Traditionally, audio signals are represented as spectrograms using the short-time Fourier transform (STFT). In the case of multiple channels, one has to combine spectrograms associated with different microphones and the *multiplicative transfer function* (MTF) is often used for this purpose. The multiplicative model is not well suited when the task consists of distinguishing between the direct-path sound, on one side, and early and late reverberations, on the other side. Instead we proposed to use the *convolutive transfer function* (CTF). The CTF model was combined with supervised localization [46] to yield a sound localization method that is immune to the presence of reverberation [78]. We used a probabilistic setting to extend

this method to multiple sound sources [79], to online localization and tracking [72], to dereverberation [73], and to speech separation and enhancement [77, 76]. We also developed a method for online speech dereverberation [75].

The use of audio signal processing in robotics – *robot audition* – has received less attention, compared to the long history of research in robot vision. We contributed to this new research topic in several ways. We thoroughly studied the geometry of multiple microphones for the purpose of sound localization from time delays [32] and for fusing audio and visual data [33]. The use of the CTF mentioned above, in conjunction with microphones embedded into robot heads, has been thoroughly investigated and implemented onto our robotic platforms [74, 71, 31]. In parallel, we investigated novel approaches to sensor-based robot control based on reinforcement learning [65, 66].

3 Research program

ROBOTLEARN will be structured in three research axes, allowing to develop socially intelligent robots, as depicted in the accompanying figure. First, on deep probabilistic models, which include the large family of deep neural network architectures, the large family of probabilistic models, and their intersection. Briefly, we will investigate how to jointly exploit the representation power of deep network together with the flexibility of probabilistic models. A well-known example of such combination are variational autoencoders. Deep probabilistic models are the methodological backbone of the proposed projet, and set the foundations of the two other research axes. Second, we will develop methods for the automatic understanding of human behavior from both auditory and visual data. To this aim we will design our algorithms to exploit the complementary nature of these two modalities, and adapt their inference and on-line update procedures to the computational resources available when operating with robotic platforms. Third, we will investigate models and tools allowing a robot to automatically learn the optimal social action policies. In other words, learn to select the best actions according to the social environment. Importantly, these action policies should also allow us to improve the robotic perception, in case this is needed to better understand the ongoing interaction. We believe that these two research axes, grounded on deep and probabilistic models, will ultimately enable us to train robots to acquire social intelligence, meaning, as discussed in the introduction, the capacity to look, listen, learn, move and speak.

3.1 Deep probabilistic models

A large number of perception and interaction processes require temporal modeling. Consider for example the task of extracting a clean speech signal from visual and audio data. Both modalities live in high-dimensional observation spaces and one challenge is to extract low-dimensional embeddings that encode information in a compact way and to update it over time. These high-dimensional to low-dimensional mappings are nonlinear in the general case. Moreover, audio and visual data are corrupted by various perturbations, e.g. by the presence of background noise which is mixed up with the speech signal uttered by a person of interest, or by head movements that overlap with lip movements. Finally, for robotics applications, the available data is scarce, and datasets captured in other settings can only serve as proxies, thus requiring either adaptation [99] or the use of unsupervised models [36]. Therefore, the problem is manyfold: to extract low-dimensional compact representations from high-dimensional inputs, to disregard useless data in order to retain information that is relevant for the task at hand, to update and maintain reliable information over time, and to do so in without (or with very few) annotated data from the robot.

This class of problems can be addressed in the framework of state-space models (SSMs). In their most general form, SSMs are stochastic nonlinear systems with latent variables. Such a system is composed of a state equation, that describes the dynamics of the latent (or state) variables, and M observation equations (an observation equation for each sensorial modality m) that predict observations from the state of the system, namely:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{v}_t \qquad \mathbf{y}_t^m = g_m(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t^m, \forall\, m \in \{1 \dots M\}, \tag{1}$$

where the latent vector $\mathbf{x} \in \mathbb{R}^L$ evolves according to a nonlinear stationary Markov dynamic model driven by the observed control variable $\mathbf{u}$ and corrupted by the noise $\mathbf{v}$. Similarly, the observed vectors $\mathbf{y}^m \in \mathbb{R}^{D_m}$

are modeled with nonlinear stationary functions of the current state and current input, affected by noise $\mathbf{w}^m$. Models of this kind have been examined for decades and their complexity increases from linear-Gaussian models to nonlinear and non-Gaussian ones. Interestingly, they can also be viewed in the framework of probabilistic graphical models to represent the conditional dependencies between the variables. The objective of an SSM is to infer the sequence of latent variables by computing the posterior distribution of the latent variable, conditioned by the sequence of observations, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

When both f and g are linear and when the noise processes $\mathbf{v}$ and $\mathbf{w}$ are both Gaussian, this becomes a linear dynamical system (LDS), also well known as the Kalman filter (KF), which is usually solved in the framework of probabilistic latent variable models. Things become more complex when both f and g are nonlinear, as the integrals required by the evaluation of the posterior become intractable. Several methods were proposed to deal with nonlinear SSMs, e.g. Bayesian tracking with particle filters, the extended Kalman filter (EKF), and the unscented Kalman filter (UKF).

Outcomes of nonlinear and non-Gaussian Bayesian trackers based on sampling were reviewed and discussed [34], most notably the problems of degeneracy, choice of importance density, and resampling. The basic idea of EKF is to linearize the equations using a first-order Taylor expansion and to apply the standard KF to the linearized model. The additional error due to linearization is not taken into account which may lead to sub-optimal performance. Rather than approximating a nonlinear dynamical system with a linear one, UKF specifies the state distribution using a minimal set of deterministically selected sample points. The sample points, when propagated through the true nonlinear system, capture the posterior state distribution accurately up to the third-order Taylor expansion. An expectation-maximization (EM) algorithm was proposed in [88] that alternates between an extended Kalman smoother which estimates an approximate posterior distribution (E-step), and nonlinear regression using a Gaussian radial basis function network to approximate f and g (M-step).

An alternative to nonlinear SSMs is to consider K different linear dynamic regimes and to introduce an additional discrete variable, a *switch*, that can take one out of K values – the switching Kalman filter (SKF). The drawback of SKFs is the exponential increase of the number of mixture components of the posterior distribution over time, namely K^t, hence an approximate posterior must be evaluated at each time step, e.g. the generalized pseudo-Bayes of order 2 (GPB2) algorithm [83].

A similar type of intractability (exponential increase of the number of mixture components of the posterior distribution) appears in the case when SSMs are used to track several objects and when there are several possible observations that are likely to be associated with each object. In such cases, additional discrete hidden variables are necessary, namely a variable that associates the i-th observation $\mathbf{y}_{it}$ with the j-th object $\mathbf{x}_{jt}$ at time t. Let these variables be denoted with $Z \in \mathbb{N}$, e.g. $Z_{it} = j$ means that observation i at t is assigned to object j. The number of mixture components of the posterior distribution after t time steps is N^{Mt}, where N is the number of state variables (objects to be tracked) and M is the number of observed variables. Problems like these can be solved in the framework of Bayesian variational inference. We developed a general framework for variational multiple object tracking and proposed several tractable variational expectation-maximization algorithms (VEM) for visual, audio, and audio-visual multiple-object tracking, [35, 72, 37, 15].

Very recently, there has been strong interest into building SSMs in the framework of deep neural networks (DNNs). This is a very promising topic of research for several reasons. It allows the representation of arbitrary nonlinear state and observation functions, f and g, using a plethora of feedforward and recurrent neural network architectures and hence to develop practical discriminative and generative deep filters, without the limitations of linear-Gaussian models that have been the state-of-the-art for several decades. In its general form, an RNN replaces eq. (1) with (for simplicity, we consider a single modality and hence we omit the modality index m):

$$\mathbf{x}_{t+1} = f(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{u}_t + \mathbf{b}), \qquad \mathbf{y}_t = g(\mathbf{V}\mathbf{x}_t + \mathbf{c}), \tag{2}$$

where $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$ are hidden-to-hidden, input-to-hidden and hidden-to-output weight matrices, $\mathbf{b}$, $\mathbf{c}$ are bias vectors, and f, g are activation functions. The discriminative formulation allows end-to-end learning using a loss and simple and scalable stochastic gradient descent methods, thus exploiting the power of deep neural networks to represent data. It opens the door to devising SSMs with high-dimensional observation spaces. Nevertheless, while discriminative recurrent neural network (RNN) learning is well understood and efficient training methods are available, they are strongly dependent on the availability of large corpora of annotated data. In some cases, data annotation can be done relatively easily, e.g. adding

various noise types to speech signals, in many other cases, augmenting the data with the corresponding ground-truth annotations is cumbersome.

One can build on the analogy between SSMs, i.e. (1) and RNNs, i.e. (2). Roughly speaking, $\mathbf{u}_t$ is the network input, $\mathbf{x}_t$ is the internal (or hidden) state and $\mathbf{y}_t$ is the output. A large number of combinations of feedforward and recurrent network architectures are possible in order to build the two functions. These combinations must be carefully studied as there is no universal solution to solve all the problems raised by processing complex audio and visual data. For example, the back-propagation Kalman filter (BKF) [59] combines a feedforward convolutional neural network (CNN) that transforms the input into into a low-dimensional vector which serves then as input for an RNN, improperly named Kalman filter.

In addition to require large amounts of annotated data for training their parameters, RNNs suffer from another main drawback: they are deterministic. Therefore, it is not possible to learn, exploit and track over time the uncertainty associated with the underlying temporal processes. Moreover, it is unclear how to use such models in unsupervised settings where the test data might be scarce and from a slighly different statistical distribution (typical case in robotic applications). Recently, there has been a burgeoning literature that addresses these issues, at the cross-roads of deep recurrent neural networks and probabilistic models. We recently released an extensive and comprehensive review of these model and methods and proposed several promising research avenues [16]. In more detail, we proposed a novel class of models that may well be viewed as an umbrella for several methodologies that were recently proposed in the literature, we unified the notations, and we identified a number of promising research lines. We termed this class of models Dynamic Variational Autoencoders (DVAE). In one sentence, this means that we aim at modeling recurrent processes and the associated uncertainty by means of deep neural networks and probabilistic models. We name the larger family of all these methods as Deep Probabilistic Models (DPMs), which form a backbone among the methodological foundations of ROBOTLEARN.

Learning DPMs is challenging from the theoretical, methodological and computational points of view. Indeed, the problem of learning, for instance, deep generative Bayesian filters in the framework of nonlinear and non-Gaussian SSMs remains intractable and approximate solutions, that are both optimal from a theoretical point of view and efficient from a computational point of view, remain to be proposed. We plan to investigate both discriminative and generative deep recurrent Bayesian networks and to apply them to audio, visual and audio-visual processing tasks.

Exemplar application: audio-visual speech enhancement

Speech enhancement is the task of filtering a noisy speech signal, e.g. speech corrupted by the ambient acoustics. In the recent past we have developed a handful of methods to address this task in challenging scenarios (e.g. high reverberation or very low signal-to-noise ratios).

We first proposed an architecture based on LSTMs to perform spectral-noise estimation [81] and speech enhancement [80]. The idea of the latter is to map speech signals into the spectral domain using the short-time Fourier transform (STFT) and hence to represent audio signals in a time-frequency space. The input of the proposed LSTM-based narrow-band filter is a noisy signal while the target used for network training is a noise-free signal. This *discriminative deep filter* formulation yields excellent results when applied to speech enhancement. Since the filter processes the STFT input frequency-wise (hence the name narrow band) it is generalizable to other types of temporal data. For example we can use this same concept to process human gestures and facial expressions over time.

In order to capture and exploit the uncertainty, we also exploited variational auto-encoders (VAEs) [62] which are feed-forward encoder-decoder latent variable networks, that have recently gained an immense popularity. We developed a VAE-based speech enhancement method which learns a speech model. At test time, this pre-trained speech model is combined with a nonnegative matrix factorization (NMF) noise model whose parameters are estimated from an observed noise-corrupted speech signal [69, 70]. This formulation has two distinctive features: (i) there is no need to learn in the presence of various noise types, since the VAE network learns a clean-speech model, and (ii) pairs of noisy- and clean-speech signals are not necessary for training, as it is the case with discriminative approaches. Currently the use of NMF techniques limits the representation power of the noise signal. More powerful models, such as DVAEs could also be used within the same general-purpose formulation.

We have also started to investigate the extension of unimodal (audio) VAE-based speech enhancement

method to multimodal (audio-visual) speech enhancement. It is well established that audio and visual data convey complementary information for the processing of speech. In particular the two modalities are affected by completely different sources of noise. Indeed, audio-speech is contaminated by additive noise due to the presence of other audio sources, while visual-speech is contaminated by occlusions and by head movements. Currently, audio-visual processing methods assume clean visual information and it is absolutely not clear how to deal with noisy visual data in the framework of speech processing.

Along this line of research, we have proposed an audio-visual VAE that is trained using synchronized audio-speech and visual-speech data, thus yielding an audio-visual prior model for speech. At test time, the approach follows the same idea as in the case of audio speech enhancement: NMF for audio-noise estimation and speech reconstruction [93]. Very recently we have started to develop the concept of *mixture of variational auto-encoders* (MVAE) which is an attempt to put the two modalities on an equal footing [89, 90], as well as their temporal extension [91]. The central idea is to consider an audio encoder and a visual encoder that are jointly trained with a shared decoder. The general architecture of proposed MVAE formulation is shown on Figure 1. As is the case with VAEs, this leads to an intractable posterior distribution and we resort to variational inference to devise a tractable solver.

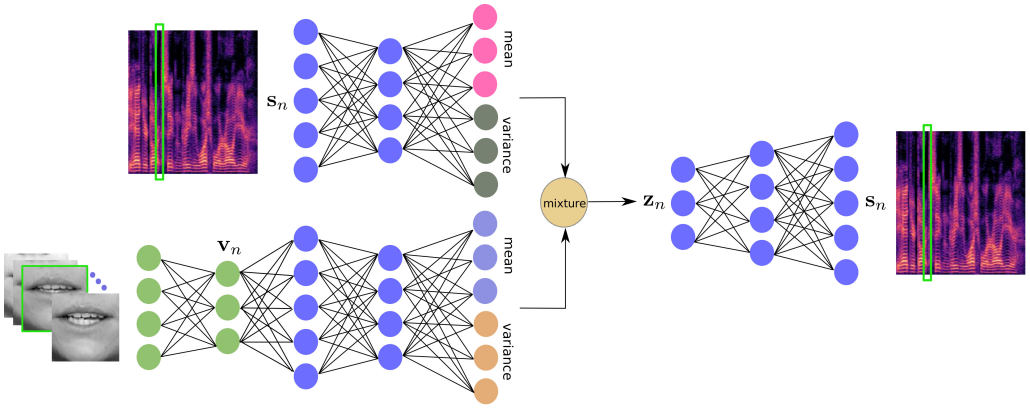

Figure 1: The proposed mixture VAE architecture for learning a speech generative model using audio and visual information (from [89]).

Research directions

We will investigate the following topics on deep probabilistic models:

- *Discriminative deep filters.* We plan to build on our recent work on discriminative deep filtering for speech enhancement [80], in order to address challenging problems associated with the temporal modeling and data fusion for robot perception and action. In particular we plan to devise novel algorithms that enable the robotic platform to, for instance, robustly track the visual focus of attention, or appropriately react to its changes. Such tasks require end-to-end learning, from the detection of facial and body landmarks to the prediction of their trajectories and activity recognition. In particular, we will address the task of characterizing temporal patterns of behavior in flexible settings, e.g. users not facing the camera. For example, lip reading for speech enhancement and speech recognition must be performed in unconstrained settings, e.g. in the presence of rigid head motions or when the user's face is partially occluded. Discriminative deep filters will also be investigated, within the framework of reinforcement learning, to devise optimal action policies exploiting sequential multi-modal data.

- *Generative deep recurrent neural networks.* Most of the VAE-based methods in the literature are tailored to use uni-modal data. VAE models for multimodal data are merely available and we are among the first to propose an audio-visual VAE model for speech enhancement [93]. Nevertheless, the proposed framework treats the two modalities unevenly. We started to investigate the use of mixture models in an attempt to put the two modalities on an equal footing [89, 90, 91]. However,

this is a long term endeavor since it raises many difficult questions from both theoretical and algorithmic points of view. Indeed, while the concept of noisy speech is well formalized in the audio signal processing domain, it is not understood in the computer vision domain. We plan to thoroughly address the combination of generative deep networks with robust mixture modeling, using for instance heavy-tailed Student-t distributions, and coping with the added complexity by means of variational approximations. Eventually, we will consider combinations of VAEs with sequential models such as for instance RNNs, and with attention-based architectures such as transformers [51]. Ideally, we will work towards devising generic methodologies spaning a wide variety of temporal models. As already mentioned, we started to investigate this problem in the framework of our work on speech enhancement [68], which may be viewed either as a recurrent VAE or, more generally, as a non-linear DNN-based formulation of SSMs. We will apply this kind of deep generative/recurrent architectures to other problems that are encountered in audio-visual perception and we will propose case-by-case tractable and efficient solvers.

3.2 Human behavior understanding

Interactions between a robot and a group of people require human behavior understanding (HBU) methods. Consider for example the tasks of detecting eye-gaze and head-gaze and of tracking the gaze directions associated with a group of participants. This means that, in addition to gaze detection and gaze tracking, it is important to detect persons and to track them as well. Additionally, it is important to extract segments of speech, to associate these segments with persons and hence to be able to determine over time who looks to whom and who is the speaker and who are the listeners. The temporal and spatial fusion of visual and audio cues stands at the basis of understanding social roles and of building a multimodal conversational model.

Performing HBU tasks in complex, cluttered and noisy environments is challenging for several reasons: participants come in an out of the camera field of view, their photometric features, e.g. facial texture, clothing, orientation with respect to the camera, etc., vary drastically, even over short periods of time, people look at an object of interest (a person entering the room, a speaking person, a TV/computer screen, a wall painting, etc.) by turning their heads away from the camera, hence facial image analysis is difficult, small head movements are often associated with speech which perturbs both lip reading and head-gaze tracking, etc. Clearly, understanding multi-person human-robot interaction is complex because the person-to-person and person-to-object, in addition to person-to-robot, interactions must explicitly be taken into account.

We propose to perform audio-visual HBU by taking explicitly into account the complementary nature of these two modalities. Differently from one current trend in AV learning [30, 42, 54], we opt for unsupervised probabilitic methods that can (i) assign observations to persons without supervision, (ii) be combined with various probabilistic noise models and (iii) and fuse various cues depending on their availability in time (i.e. handle missing data). Indeed, in face-to-face communication, the robot must choose with who it should engage dialog, e.g. based on proximity, eye gaze, head movements, lip movements, facial expressions, etc., in addition to speech. Unlike in the single-user human-robot interaction case, it is crucial to associate temporal segments of speech to participants, referred to as speech diarization. Under such scenarios, speech signals are perturbed by noise, reverberation and competing audio sources, hence speech localization and speech enhancement methods must be used in conjunction with speech recognition. The relationship with natural language understanding and spoken dialog, while very relevant, falls outside the team's expertise. This relationship will be investigated in collaboration with the Interaction Lab at Heriot-Watt University (lead by Prof. Oliver Lemon), a partner of H2020 SPRING project and with the Laboratoire d'Intelligence Artificielle at Université d'Avignon (professor Fabrice Lefèvre), partner of ANR μDialbot project.

As already explained (see Section 3.1) we have recently investigated various aspects of *dynamic* HBU, namely multiple-person tracking based on visual [35], audio [72, 37], or audio-visual information [15], head-pose estimation [49], eye-gaze tracking [83], e.g. Fig. 2, and audio-visual diarization [55]. Our recent work has relied on Gaussian mixture regression [45], on dynamic Bayesian networks [85] and on their variational approximations, e.g. [15]. Such probabilistic and statistical formulations provide robust, powerful and flexible unsupervised learning techniques for HBU. In parallel, there has been strong interest in using deep learning techniques for HBU, e.g. person detection, person tracking,

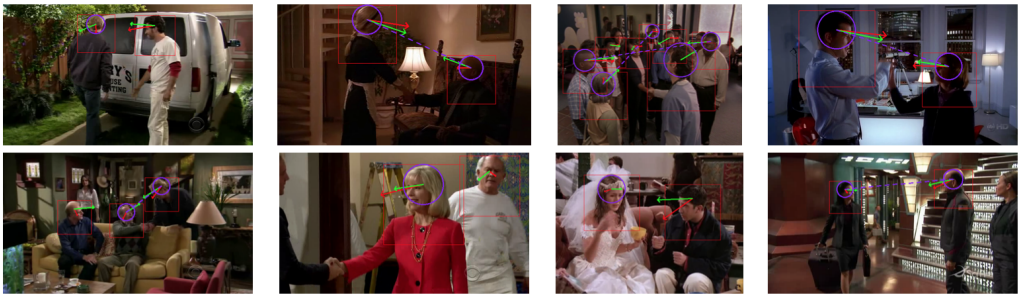

Figure 2: This figure shows some examples of eye-gaze detection and tracking obtained with the method proposed in [83]. The algorithm infers both eye-gaze (green arrows) and visual focus of attention (blue circles) from head-gaze [49] (red arrows). A side-effect of this inference is the detection of people looking at each other (dashed blue line).

facial expression recognition, etc. Nevertheless, deep neural networks still have difficulties in capturing motion information directly from image sequences. For example, human activity detection, tracking and recognition use pre-computed optical flow to compute motion information. Most of the work has focused on HBU at a single person level and less effort has been devoted into developing deep learning methods for studying group activities and behavior, in particular in the context of interaction.

A comprehensive analysis of groups of people should rely on combining Bayesian modeling with deep neural networks. Indeed, this enables us to sum up the flexibility of the former with the representative and discriminative power of the latter. We plan to combine deep generative networks (see Section 3.1) trained for person tracking with person descriptors based on deep discriminative learning. Fully generative strategies will also be investigated, possibly exploiting features pre-trained in discriminative settings, thus exploiting large-scale annotated datasets available for certain tasks. Indeed, current state of the art provides DNN architectures well suited for learning embeddings of images and of image primitives. However, these embeddings are learned off-line using very large training datasets to guarantee data variability and generality. It is however necessary to perform some kind of adaptation to the distribution of the particular data at hand, e.g. collected with robot sensors. If these data are available in advance, off-line adaptation can be done, otherwise the adaptation needs to be performed on-line or at run time. Such strategies will be useful given the particular experimental conditions of practical human-robot interaction scenarios.

On-line learning based on deep neural networks is far from being well understood. We plan to thoroughly study the incorporation of on-line learning into both Bayesian and discriminative deep networks. In the practical case of interaction, real-time processing is crucial. Therefore, a compromise must be found between the size of the network, its discriminative power and the computational cost of the learning and prediction algorithms. Clearly, there is no single solution given the large variety of problems and scenarios that are encountered in practice.

Exemplar application: multi-person facial landmark tracking

The problem of facial landmark tracking of mulitple persons can be formulated as a two-stage problem, namely, first we track each face, and second we extract facial landmarks from each tracked face, e.g. [92], and this in a robust manner. We recently proposed solutions to perform on-line multi-person tracking and started to explore how to robustly extract landmarks.

We proposed an on-line method to concurrently track a variable number of people and update the appearance model of each person [47], in order to make it more robust to changes in illumination, pose, etc. Such appearance models must yield extremely discriminative descriptors, such that two observed faces are unambiguously assigned to two different persons being tracked. This means that observation-to-person assignments, computed during the tracking itself, must be used to fine tune the (possibly deep) representation. This fine tuning needs to be carefully investigated and properly incorporated into the probabilistic tracker. Indeed, if enough data is available, the network could be fine-tuned, as in [18]. Otherwise, the representation could be updated without fine-tuning the network [47].

From these tracks, one could attempt to analyse facial expressions based on, e.g. facial landmarks. Even if several facial landmark extraction algorithms exist in the literature, how to properly separate rigid head motions (such as turning the head or simply moving) from non-rigid face movements, i.e. expressions, within the current models is unclear. Indeed, the analysis of facial expressions is a difficult task by its own, even if rigid head movements have been subtracted out. We started investigating how to assess the quality of the extracted landmarks and we plan to use these assessments to design and train architectures implementing SSMs for robust facial landmark extraction and tracking. Such architectures could be easily used for other tasks such as human gesture recognition, robust body landmark estimation, facial expression recognition or speech activity estimation.

More generally, these examples are instances of the problem of on-line discriminative learning. Generally, discriminative learning uses deterministic targets/labels for learning, such as the ones produced by manual annotation of large amounts of data. In the on-line case we do not have the luxury of manual annotation. We must therefore rely on less reliable labels. In other words, we must compute and maintain over time a measure of label reliability.

Consider again the problem of tracking N persons over time and let M be the number of observations at each time step. Let a latent discrete variable Z denote the association between an observation and a person. At each time step t, each observation i is assigned to each person j with probability $p(Z_{it} = j)$. Therefore, one can replace deterministic labels with their probability distribution function. Case-by-case analysis must be carefully carried out in order to choose the proper network and learning strategy.

Research directions

Our research plan on human behavior understanding is summarized as follows:

- *Deep visual descriptors.* One of the most important ingredients of HBU is to learn visual representations of humans using deep discriminative networks. This process comprises detecting people and body parts in images and then extracting 2D or 3D landmarks. We plan to combine body landmark detectors and facial landmark detectors, based on feedforward architectures, with landmark tracking based on recurrent neural networks. The advantage is twofold: to eliminate noise, outliers and artefacts, which are inherent to any imaging process, and to build spatio-temporal representations for higher-level processes such as action and gesture recognition. While the task of noise filtering can be carried out using existing techniques, the task of removing outliers and artefacts is more difficult. Based on our recent work on robust deep regression, we plan to develop robust deep learning methods to extract body and facial landmarks. In addition to the Gaussian-uniform mixture used in [67], we plan to investigate the Student t-distribution and its variants as it has interesting statistical properties, such as robustness due to their so-called heavy tail. Moreover, we plan to combine deep learning methods with robust rigid registration methods in order to distinguish between rigid and non-rigid motion and to separate them. This research will combine robust probability distributions with deep learning and hence will lead to novel algorithms for robustly detecting landmarks and tracking them over time. Simultaneously, we will address the problem of assessing the quality of the landmarks without systematic recourse to annotated datasets.

- *Deep audio descriptors.* We will also investigate methods for extracting descriptors from audio signals. These descriptors must be free of noise and reverberation. While there are many noise filtering and dereverberation methods available, they are not necessarily well adapted to the tasks involved in live interaction between a robot and a group of people. In particular, they often treat the case of a static acoustic scene: both the sources and the microphones remain fixed. This represents a strong limitation and the existing methods must be extended to deal with dynamic acoustic scenes, e.g. [63]. Based on our recent work [73], we plan to develop deep audio descriptors that are robust against noise and reverberation. We will train these descriptors to help the tasks of speech enhancement and speech dereverberation in order to facilitate down-stream tasks such as speech-source localization and speech recognition. Moreover, we plan to develop a speaker recognition method that can operate in a complex acoustic environment. As done in computer vision for person re-identification [53], recent works adapt the embedding network to an unknown domain. Adversarial strategies to further increase the performance have also been proposed [60],

and we have contributed for person re-identification [18]. How to exploit these strategies with a continuous flow of observations acquired by a robotic platform remains to be investigated.

3.3 Learning and control for social robots

Traditionally, research on human-robot interaction focused on single-person scenarios also called dyadic interactions. However, over the past decade several studies were devoted to various aspects of *multi-party* interactions, meaning situations in which a robot interacts with a group of two or more people [94]. This line of research is much more challenging because of two main reasons. First, the behavioral cues of each individual and of the group need to be faithfully extracted (and assigned to each individual). Second, the behavioral dynamics of groups of people can be pushed by the presence of the robot towards competition [41] or even bullying [40]. This is why some studies restrict the experimental conditions to very controlled collaborative scenarios, often lead by the robot, such as quiz-like game playing [98] or very specific robot roles [52]. Intuitively, constraining the scenario also reduces the gesture variabilty and the overall interaction dynamics, leading to methods and algorithms with questionable generalisation to free and natural social multi-party interactions.

Whenever a robot participates in such multi-party interactions, it must perform *social actions*. Such robot social actions are typically associated with the need to perceive a person or a group of persons in an optimal way as well as to take appropriate decisions such as to safely move towards a selected group, to pop into a conversation or to answer a question. Therefore, one can distinguish between two types of robot social actions: (i) *physical actions* which correspond to synthesizing appropriate motions using the robot actuators (motors), possibly within a sensorimotor loop, so as to enhance perception and maintain a natural interaction and (ii) *spoken actions* which correspond to synthesizing appropriate speech utterances by a spoken dialog system. In ROBOTLEARN we will focus on the former, and integrate the latter via collaborations with research groups having with established expertise in speech technologies.

For example, robust speech communication requires clean speech signals. Nevertheless, clean speech could be retrieved by the robot in several ways and based on different strategies. The first strategy is that the robot stays still and performs audio signal processing in order to reconstruct clean speech signals from noisy ones, e.g. in the presence of reverberation and of competing audio sources. The second strategy consists of moving towards a speaking person in order to face her/him directly and to optimize the quality of the audio signals gathered with the onboard microphones. Therefore, apparently simple speech communication tasks between a robot and a person involve a complex analysis in order to take appropriate decisions: Is the room noisy? Are there many people in the robot's field of view? How far are they? Are they looking at the robot? Is speech enhancement sufficient, or should the robot move towards a person in order to reduce the effects of room reverberation and of ambient noise? Clearly, robot perception and robot action are intimately interleaved, and the robot actions should be selected on the premise that social behavior counts.

In this regard we face three problems. First, given the complexity of the environment and the inherent limitations of the robot's perception capabilities, e.g. limited camera field of view, cluttered spaces, complex acoustic conditions, etc., the robot will only have access to a partial representation of the environment, and up to a certain degree of accuracy. Second, for learning purposes, there is no easy way to annotate which are the best actions the robot must choose given a situation: supervised methods are therefore not an option. Third, since the robot cannot learn from scratch by random exploration in a new environment, standard model-free RL approaches cannot be used. Some sort of previous knowledge on the environment or a similar one should be exploited. Finally, given that the robot moves within a populated environment, it is desirable to have the capability to enforce certain constrains, thus limiting the range of possible robot actions.

Building algorithms to endow robots with autonomous decision taking is not straightforward. Two relatively distinct paradigms are available the literature. First, one can devise customized strategies based on techniques such as *robot motion planning* combined with *sensor-based robot control*. These techniques lack generalization, in particular when the robot acts in complex, dynamic and unconstrained environments. Second, one can let the robot devise its own strategies based on *reinforcement learning* (RL) – a machine learning paradigm in which "agents" learn by themselves by trial and error to achieve successful strategies[95]. It is very difficult, however, to enforce any kind of soft- or hard-constraint within

this framework. We will showcase these two scientific streams with one group of techniques for each one: *model predictive control* (MPC) and Q-learning, *deep Q-networks* (DQNs), more precisely. These two techniques are promising. Moreover, they are well documented in the robotics and machine learning. Nevertheless, combining them is extremely challenging.

MPC is a generic framework which allows the incorporation of constraints in the process of robot decision-taking. More formally MPC requires (i) a transition function $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t)$, i.e. generalization of (1), (ii) a correction function $e(\mathbf{x}_t, \mathbf{u}_t) \in \mathbb{R}$ and (iii) an optional constraint function h. The MPC problem is formally stated as an optimisation problem [64]:

$$\min_{\mathbf{u}_0,\dots,\mathbf{u}_{T-1}} \sum_{t=0}^{T-1} e(\mathbf{x}_t, \mathbf{u}_t) \qquad \text{s.t.} \quad h(\mathbf{u}_0,\dots,\mathbf{u}_{T-1},\mathbf{x}_0,\dots,\mathbf{x}_{T-1}) \leq 0, \tag{3}$$

where T is the time horizon considered in the optimisation problem.

Often, one can devise efficient solvers to find the optimal control sequence $\mathbf{u}_0^*,\dots,\mathbf{u}_{T-1}^*$. As discussed before, the advantage of MPC is the possibility to include constrains, modeled through h. Such constraints can be used to enforce safety or other must-comply rules, the scenario at hand may require. Even if it is technically possible to learn the transition function f, this has high computational cost. Therefore, one limitation of MPC is the common assumption that the transition function f is completely known. In *purely geometric* tasks, this makes sense, since one can have a fairly accurate model of how the perception of the objects present in the evolves with the robot actions. However, it is much more complex to model how the behavior of people (from their body pose to their high-level global behavior) will change due to the robot actions. One may then rather *learn* the transition function.

Alternatively, an appealing framework for learning robot behavior is DQN.[84] As any RL method, DQN is based on rewards, evaluated at each time step t and after taking an action $\mathbf{u}_t$ at state $\mathbf{x}_t$, $r_t = r(\mathbf{x}_t, \mathbf{u}_t)$. The aim is to learn the optimal *action policy* π, i.e. the one that maximises the expected accumulated reward: $\overline{r}_t = \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau}$, where $0 \leq \gamma < 1$ is a discount factor. To do so, DQN uses the so-called Q function, which is defined for a certain action policy π at a state-action $(\mathbf{x}, \mathbf{u})$ pair, as the expected accumulated reward when following policy π:

$$Q^\pi(\mathbf{x}, \mathbf{u}) = \mathrm{E}_{f,\pi}[\overline{r}_t | \mathbf{x}_t = \mathbf{x}, \mathbf{u}_t = \mathbf{u}], \tag{4}$$

where the expectation is taken over the future state distribution, using f, therefore the latter becomes a stochastic mapping rather than a deterministic one, and the action distribution, using π. Implicitly, this means that the function Q models jointly the effect of the transition function f and of the policy action π. Thus, once Q is learned, the effects of f and π cannot be disentangled.

It can be shown that the optimal Q function satisfies the following Bellman equation:

$$Q^*(\mathbf{x}, \mathbf{u}) = \mathrm{E}_{\mathbf{x}' \sim f(\mathbf{x},\mathbf{u})}[r(\mathbf{x}, \mathbf{u}) + \gamma \max_{\mathbf{u}'} Q(\mathbf{x}', \mathbf{u}')]. \tag{5}$$

In DQN, the Q function is approximated by a deep neural network, which is learned by stochastic gradient descent based on the Bellman equation. While DQN has been successfully applied to various control problems, in particular computer games, it suffers from various drawbacks. First, DQN exhibits high performance when the set of actions is discrete, as opposed to continuous actions much more suitable in robotics. Second, and this is common to the majority of RL approaches, DQN requires lots of trajectories (sequences of state-action pairs) for training. These are usually obtained through computer simulations, raising a question that remains widely open: how to efficiently simulate social interactions that follow a data distribution that the agent will face in the real world? Third, by design, RL (and hence DQN) cannot be trained in the presence of constraints. Certainly, one can discourage certain robot behaviors by designing large negative rewards of some state-action pairs, but this does not guarantee that the robot will never execute such state-action pairs.

Summarizing, on the one hand we have sensor-based robot control techniques, such as MPC, that require a faithful representation of the transition function f so as to compute the optimal action trajectory, and do not allow learning. On the other hand we have learning-based techniques that allow to learn the transition function f (together with the optimal policy function), but they cannot be coupled with hard-constraints. Our scenario is complex enough to require learning (part of) the transition function, and at the same time we would like to enforce constraints when controling the robot.

Exemplar application: audio-visual robot gaze control

Recently, we applied DQN to the problem of controlling the gaze of a robotic head using audio and visual information [65, 66]. In summary, the robot learns by itself how to turn its head towards a group of speaking faces. The DQN-like architecture is based on a long short-time memory (LSTM) network that takes as input a sequence of states $\mathbf{x}_{t-k}, \ldots \mathbf{x}_t$ (namely motor positions, person detection and sound-source detection and localization) and which predicts a Q-value for each possible action (stay still, look up, look down, look left and look right).

In order to speed up training in real time, we proposed to simulate the pose of people in the scene using standard pose-estimation datasets that contain ground-truth pose. We combined the poses of different people thanks to a set of hand-crafted rules. Additionally we emulated the output of a sound localisation algorithm that would provide the direction of the most prominent active sources. The reward given to the agent would be the number of faces found in the field of view, plus an extra reward if the speaking face was within the field of view. In this way the robot learned actions that maximise the number of people within the field of view. In addition, the robot satisfactorily learned to look at a speaking person when found that person belonged to a group of people, and to look around (explore) when none of the participants were within the field of view.

While this application may seem very simple, one must understand that simulating such data in a realistic manner is not straightforward. In addition, lots of simulations were required before fine-tuning the DQN with real-world data: the pre-training phase was very intense for such a simple task. Thus, scaling up such a simulation to more complex scenarios, e.g. where one has to take into account conversational and group dynamics, remains an open question. Other strategies allowing better generalization, such as meta RL, would be highly desirable.

Research directions

- *Constrained RL.* Naturally one may be tempted to combine MPC and DQN, but this is unfortunately not possible. Indeed, DQN cannot disentangle the policy π from the environment f, and MPC requires an explicit expression for f to solve the associated optimisation problem, their direct combination is not possible. We will investigate two directions. First, to devise methodologies able to efficiently learn the transition function f, to later on use it within the MPC framework. Second, to design learning methodologies that are combined with MPC, so that the actions taken within the learning process satisfy the required constraints. A few combinations of RL and MPC for robot navigation in human-free scenarios [86, 61], as well as MPC variants driven by datahave recently appeared in the literature. How to adapt this recent trend to dynamic complex environments such as a multi-party conversational situation is still to be investigated. Additionally, the use of audio-visual fusion in this context needs to be explored deeply, and this also holds for the second research line.

- *Meta RL.* An additional challenge, independent from the learning and control combination foreseen, is the data distribution gap between the simulations and the real-world. Meta-learning, or the ability to learn how to learn, can provide partial answers to this problem. Indeed, developing machine learning methods able to understand how the learning is achieved can be used to extend this learning to a new task and speed up the learning process on the new task. Recent developments proposed meta-learning strategies specifically conceived for reinforcement learning, leading to Meta-RL methods. One promising trend in Meta-RL is to have a probabilistic formulation involving SSMs and VAEs, i.e. hence sharing the methodology based on dynamical variational autoencoders described before. Very importantly, we are not aware of any studies able to combine Meta-RL with MPC to handle the constraints, and within a unified formulation. From a methodological perspective, this is an important challenge we face in the next few years.

4 Application domains

For the last decades, there has been an increasing interest in robots that cooperate and communicate with people. As already mentioned, we are interested *Socially Assistive Robots* (SARs) that can communicate

with people and that are perceived as social entities. So far, the humanoid robots developed to fill this role are mainly used as research platforms for human-robot collaboration and interaction and their prices, if at all commercially available, are in the 6-digit-euro category, e.g. 250,000€ for the iCub robot and Romeo humanoid robots, developed by the Italian Institute of Technology and SoftBank Robotics Europe, respectively, as well as the REEM-C and TALOS robots from PAL Robotics. A notable exception being the NAO robot which is a humanoid (legged) robot, available at an affordable price. Apart from humanoid robots, there are also several companion robots manufactured in Europe and available at a much lower price (in the range 10,000–30,000 €) that address the SAR market. For example, the Kompaï, the TIAGo, and the Pepper robots are wheeled indoor robotic platforms. The user interacts with these robots via touch screen and voice commands. The robots manage shopping lists, remember appointments, play music, and respond to simple requests. These affordable robots (Kompaï, TIAGo, NAO, and Pepper) rapidly became the platforms of choice for many researchers in cognitive robotics and in HRI, and they have been used by many EU projects, e.g. HUMAVIPS, EARS, VHIA, and ENRICHEME.

When interacting, these robots rely on a few selected modalities. The voice interface of this category of robots, e.g. Kompaï, NAO, and Pepper, is based on speech recognition similar to speech technologies used by smart phones and table-top devices, e.g. Google Home. *Their audio hardware architecture and software packages are designed to handle single-user face-to-face spoken dialogue based on keyword spotting, but they can neither perform multiple sound-source analysis, fuse audio and visual information for more advanced multi-modal/multi-party interactions, nor hold a conversation that exceeds a couple of turns and that is out of very narrow predefined domain.*

To the best of our knowledge, the only notable efforts to overcome some of the limitations mentioned above are the FP7 EARS and H2020 MuMMER projects. The EARS project's aim was to redesign the microphone-array architecture of the commercially available humanoid robot NAO, and to build a robot head prototype that can support software based on advanced multi-channel audio signal processing. The EARS partners were able to successfully demonstrate the usefulness of this microphone array for speech-signal noise reduction, dereverberation, and multiple-speaker localisation. Moreover, the recent IEEE-AASP Challenge on Acoustic Source Localisation and Tracking (LOCATA) comprises a dataset that uses this microphone array. The design of NAO imposed severe constraints on the physical integration of the microphones and associated hardware. Consequently and in spite of the scientific and practical promises of this design, SoftBank Robotics has not integrated this technology into their commercially available robots NAO and Pepper. In order to overcome problems arising from human-robot interaction in unconstrained environments and open-domain dialogue on the Pepper robot, the H2020 MuMMER project aimed to deploy an entertaining and helpful robot assistant to a shopping mall. While they had initial success with short deployments of the robot to the mall, they were not specifically addressing the issues arising from multi-party interaction: Pepper's audio hardware/software design cannot locate and separate several simultaneously emitting speech sources.

To conclude, *current robotic platforms available in the consumer market, i.e. with large-scale deployment potential, are neither equipped with the adequate hardware nor endowed with the appropriate software required for multi-party social interactions in real-world environments.*

In the light of the above discussion, the partners of the H2020 SPRING project decided to build a robot prototype well suited for socially assistive tasks and shared by the SPRING partners as well as by other EU projects. We participated to the specifications of the ARI robot prototype (shown on the right), designed, developed and manufactured by PAL Robotics, an industrial partner of the SPRING project. ARI is a ROS-enabled, 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. Seven ARI robot units were delivered to the SPRING partners in April 2021.

We are committed to implement our algorithms and associated software packages onto this advanced robotic platform, from low-level control to high-level perception, interaction and planning tasks, such that the robot has a socially-aware behaviour while it safely navigates in an ever changing environment. We will experiment in environments of increasing complexity, e.g. our robotic lab, the Amiqual4Home facility, the Inria Grenoble cafeteria and Login exhibition, as well as the Broca hospital in Paris. The expertise that the team's engineers and researchers have acquired for the last decade would be crucial for present and future robotic developments and experiments.

Figure 3: The ARI robot from PAL Robotics.

5 Highlights of the year

Over the past year, we have many scientific contributions that we would like to quickly summarise. More details will be provided later on.

We have developed a transformer-based architecture for multiple object tracking is now under review at TPAMI Beyond tracking, we have also contributed to multi-person body pose estimation, see our WACV paper on the topic. In this line, we have collected, curated and exploited the Extreme Pose Interaction (ExPI) dataset, where we investigate the prediction of human motion in complex actions such as aerial/acrobatic dancing steps.

We have also worked towards exploiting facial landmarks to frontalise the face, i.e. remove rigid movements, while keeping the lip movements and use them for visual speech recognition (see the associated ICCV-W publication). This is naturally related to our previous work on robust 3D face alignment. Naturally related, our contributions on speech enhancement/separation include the switching VAE for AV speech enhancement (ICASSP 2021), and the mixture VAE for speech enhancement (TSP) and for separation (MLSP 2021). These models merge the VAE methodology with other probabilistic models, including some sort of temporal dependency. We published an extensive review of models including the temporal dependency within the deep generative model, or Dynamical Variational Autoencoders, in Foundations and Trends on Machine Learning and Interspeech'21. Going back to the use of facial landmarks, we have investigated how to learn to generate inter-action sequences (submitted to TAFFC).

This past year we have also investigated how to learn robot action policies in various contexts. First, by developeding a navigation module based on the model predictive control (MPC) methodology. This is now working on ARI. We have worked on meta/transfer reinforcement learning (RL), generalising successor features to non-linear reward functions or xi-learning. We have also contributed to the use of neural episodic control in combination with linear successor features (presented at NeurIPS-W).

6 New software and platforms

6.1 New software

6.1.1 TransCenter

Name: TransCenter: Transformers with Dense Queries for Multiple-Object Tracking

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

Scientific Description: Transformer networks have proven extremely powerful for a wide variety of tasks since they were introduced. Computer vision is not an exception, as the use of transformers has become very popular in the vision community in recent years. Despite this wave, multiple-object tracking (MOT) exhibits for now some sort of incompatibility with transformers. We argue that the standard representation — bounding boxes with insufficient sparse queries — is not optimal to learning transformers for MOT. Inspired by recent research, we propose TransCenter, the first transformer-based MOT architecture for dense heatmap predictions. Methodologically, we propose the use of dense pixel-level multi-scale queries in a transformer dual-decoder network, to be able to globally and robustly infer the heatmap of targets' centers and associate them through time. TransCenter outperforms the current state-of-the-art in standard benchmarks both in MOT17 [2] and MOT20 [1]. Our ablation study demonstrates the advantage in the proposed architecture compared to more naive alternatives.

Functional Description: TransCenter is a software for multiple-object tracking using deep neural networks. It allows tracking multiple people in a very crowded scenes.

URL: https://team.inria.fr/robotlearn/transcenter-transformers-with-dense-queriesfor-multiple-object-tracking/

Publication: hal-03295680

Contact: Soraya Arias

Participants: Yihong Xu, Guillaume Delorme, Xavier Alameda Pineda, Daniela Rus, Yutong Ban, Chuang Gan

6.1.2 xi_learning

Name: Xi Learning

Keywords: Reinforcement learning, Transfer Learning

Functional Description: Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor features (SF) are a prominent transfer mechanism in domains where the reward function changes between tasks. They reevaluate the expected return of previously learned policies in a new target task and to transfer their knowledge. A limiting factor of the SF framework is its assumption that rewards linearly decompose into successor features and a reward weight vector. We propose a novel SF mechanism, ξ-learning, based on learning the cumulative discounted probability of successor features. Crucially, ξ-learning allows to reevaluate the expected return of policies for general reward functions. We introduce two ξ-learning variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on ξ-learning with function approximation demonstrate the prominent advantage of ξ-learning over available mechanisms not only for general reward functions, but also in the case of linearly decomposable reward functions.

URL: https://gitlab.inria.fr/robotlearn/xi_learning

Authors: Chris Reinke, Xavier Alameda Pineda

Contact: Chris Reinke

6.1.3 Social MPC

Keyword: Robotics

Functional Description: A library for controlling a social robot. This library allows a non-holonomic robot to navigate in a crowded environment using model predictive control and social force models. This library has been developed for the SPRING project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871245.

The main components of this library are: - A module to determine optimal positioning of a robot in a group, using methods from the litterature. - A navigation component to compute optimal paths - The main module, implementing a model predictive controller using the Jax library to determine optimal commands to steer the robot

Authors: Alex Auternaud, Timothee Wintz, Chris Reinke

Contact: Alex Auternaud

6.1.4 2D Social Simulator

Keywords: Simulator, Robotics

Functional Description: A python based simulator using Box2D allowing a robot to interact with people. This software enables: - The configuration of a scene with physical obstacles and people populating a room - The simulation of the motion of a robot in this space - Social force models for the behaviour of people, groups between themselves and in reaction to the motion of the robot

Rendering is done using PyGame and is optional (headless mode is possible).

A gym environment is provided for reinforcement learning.

URL: https://gitlab.inria.fr/spring/wp6_robot_behavior/2D_Simulator

Authors: Alex Auternaud, Timothee Wintz, Chris Reinke

Contact: Alex Auternaud

6.1.5 PI-NET

Name: Pose Interacting Network for Multi-Person Monocular 3D Pose Estimation

Keywords: Pytorch, Pose estimation, Deep learning, -

Scientific Description: Monocular 3D multi-person human pose estimation aims at estimating the 3D joints of several people from a single RGB image. PI-Net, inputs the initial pose estimates of a variable number of interactees into a recurrent architecture used to refine the pose of the person-of-interest. 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. PI-Net is implemented in PyTorch.

Functional Description: Monocular 3D multi-person human pose estimation aims at estimating the 3D joints of several people from a single RGB image. PI-Net, inputs the initial pose estimates of a variable number of interactees into a recurrent architecture used to refine the pose of the person-of-interest. 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. PI-Net is implemented in PyTorch.

URL: https://github.com/GUO-W/PI-Net

Publication: hal-02971754

Contact: Xavier Alameda Pineda

Participants: Wen Guo, Xavier Alameda Pineda

6.1.6 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 a library for speech community, to use different dynamic VAE models for speech re-synthesis (potentially for other speech application)

URL: https://github.com/XiaoyuBIE1994/DVAE-speech

Publication: hal-02926215

Authors: Xiaoyu Bie, Xavier Alameda Pineda, Laurent Girin

Contact: Xavier Alameda Pineda

6.2 New platforms

Participants: Alex Auternaud, Timothée Wintz, Chris Reinke, Luis Camara, Nicolas Turro, Soraya Arias, Radu Horaud, Xavier Alameda-Pineda.

This year we have received the ARI robot (see Figure 3) from PAL Robotics, in the framework of the H2020 SPRING project. ARI is a high-performance robotic platform designed for a wide range of multimodal expressive gestures and behaviors, making it the ideal social robot and suitable for Human-Robot-Interaction. We have customised the platrofms to the needs of the H2020 SPRING project, adding microphones and cameras to adapt its sensing capabilities to the needs of the project. Since a few months now, we are operating ARI and obtaining the first results with it.

7 New results

7.1 Transformed-based multiple object tracking

Participants: Yihong Xu, Radu Horaud, Xavier Alameda-Pineda.

Transformer networks have proven extremely powerful for a wide variety of tasks since they were introduced. Computer vision is not an exception, as the use of transformers has become very popular in the vision community in recent years. Despite this wave, multiple-object tracking (MOT) exhibits for now some sort of incompatibility with transformers. We argue that the standard representation - bounding boxes with insufficient sparse queries - is not optimal to learning transformers for MOT. Inspired by recent research, we propose TransCenter, the first transformer-based MOT architecture for dense heatmap predictions. Methodologically, we propose the use of dense pixel-level multi-scale queries in a transformer dual-decoder network, to be able to globally and robustly infer the heatmap of targets' centers and associate them through time. TransCenter outperforms the current state-of-the-art in standard benchmarks both in MOT17 and MOT20. Our ablation study demonstrates the advantage in the proposed architecture compared to more naive alternatives. See 6.1.1.

7.2 Multiperson body pose estimation in interactive environments

Participants: Wen Guo, Xavier Alameda-Pineda.

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 work, 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, inputs the initial pose estimates of a variable number of interactees into a recurrent 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.

7.3 Extreme Pose Interaction (ExPI) Dataset

Participants: Wen Guo, Xavier Alameda-Pineda.

Human motion prediction aims to forecast future poses given a sequence of past 3D skeletons. While this problem has recently received increasing attention, it has mostly been tackled for single humans in isolation. In this work, we explore this problem when dealing with humans performing collaborative tasks, we seek to predict the future motion of two interacted persons given two sequences of their past skeletons. We propose a novel cross interaction attention mechanism that exploits historical information of both persons, and learns to predict cross dependencies between the two pose sequences. Since no dataset to train such interactive situations is available, we collected ExPI (Extreme Pose Interaction), a new lab-based person interaction dataset of professional dancers performing Lindy-hop dancing actions, which contains 115 sequences with 30K frames annotated with 3D body poses and shapes. We thoroughly evaluate our cross interaction network on ExPI and show that both in short- and long-term predictions, it consistently outperforms state-of-the-art methods for single-person motion prediction. See the dedicated webpage.

7.4 Robust Face Frontalization For Visual Speech Recognition

Participants: Zhiqi Kang, Mostafa Sadeghi, Radu Horaud, Xavier Alameda-Pineda.

Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. The main contribution is a robust method that preserves non-rigid facial deformations, i.e. expressions. The method iteratively estimates the rigid transformation and the non-rigid deformation between 3D landmarks extracted from an arbitrarily-viewed face, and 3D vertices parameterized by a deformable shape model. The one merit of the method is its ability to deal with large Gaussian and non-Gaussian errors in the data. For that purpose, we use the generalized Student-t distribution. The associated EM algorithm assigns a weight to each observed landmark, the higher the weight the more important the landmark, thus favouring landmarks that are only affected by rigid head movements. We propose to use the zero-mean normalized cross-correlation score to evaluate the ability to preserve facial expressions. We show that the method, when incorporated into a deep lip-reading pipeline, considerably improves the word classification score on an in-the-wild benchmark. See the decicated webpage.

7.5 Switching Variational Autoencoders

Participants: Mostafa Sadeghi, Xavier Alameda-Pineda.

Recently, audio-visual speech enhancement has been tackled in the unsupervised settings based on variational auto-encoders (VAEs), where during training only clean data is used to train a generative model for speech, which at test time is combined with a noise model, e.g. nonnegative matrix factorization (NMF), whose parameters are learned without supervision. Consequently, the proposed model is agnostic to the noise type. When visual data is clean, audio-visual VAE-based architectures usually outperform the audio-only counterpart. The opposite happens when the visual data is corrupted by clutter, e.g. the speaker not facing the camera. In this work, we propose to find the optimal combination of these two architectures through time. More precisely, we introduce the use of a latent sequential variable with Markovian dependencies to switch between different VAE architectures through time in an unsupervised manner: leading to switching variational auto-encoder (SwVAE). We propose a variational factorization to approximate the computationally intractable posterior distribution. We also derive the corresponding variational expectation-maximization algorithm to estimate the parameters of the model and enhance the speech signal. Our experiments exhibit the performance of SwVAE.

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

Participants: Mostafa Sadeghi, Xavier Alameda-Pineda.

In this work, we are interested in unsupervised (unknown noise) speech enhancement, where the probability distribution of clean speech spectrogram is simulated via a latent variable generative model, also called the decoder. Recently, variational autoencoders (VAEs) have gained much popularity as probabilistic generative models. In VAEs, the posterior of the latent variables is computationally intractable, and it is approximated by a so-called encoder network. Motivated by the fact that visual data, i.e. lip images of the speaker, provide helpful and complementary information about speech, some audio-visual architectures have been recently proposed. The initialization of the latent variables at test time is crucial as the overall inference problem is non-convex. This is usually done by using the output of the 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. To overcome this issue, we inspire from mixture models, and introduce the mixture of inference networks variational autoencoder (MIN-VAE). 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 network is unsupervised as well. By training a shared decoder, the overall network learns to adaptively fuse the two modalities. Moreover, at test time, the video 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.

7.7 Deep Variational Generative Models for Audio-visual Speech Separation

Participants: Mostafa Sadeghi, Xavier Alameda-Pineda.

In this work, we are interested in audio-visual speech separation given a single-channel audio recording as well as visual information (lips movements) associated with each speaker. We propose an unsupervised technique based on audio-visual generative modeling of clean speech. More specifically, during training, a latent variable generative model is learned from clean speech spectrograms using a variational auto-encoder (VAE). To better utilize the visual information, the posteriors of the latent variables are inferred from mixed speech (instead of clean speech) as well as the visual data. The visual modality also serves as a prior for latent variables, through a visual network. At test time, the learned generative model (both for speaker-independent and speaker-dependent scenarios) is combined with an unsupervised non-negative matrix factorization (NMF) variance model for background noise. All the latent variables and noise parameters are then estimated by a Monte Carlo expectation-maximization algorithm. Our experiments show that the proposed unsupervised VAE-based method yields better separation performance than NMF-based approaches as well as a supervised deep learning-based technique.

7.8 Dynamical Variational Autoencoders

Participants: Xiaoyu Bie, Laurent Girin, Xavier Alameda-Pineda.

In this work, we are interested in audio-visual speech separation given a single-channel audio recording as well as visual information (lips movements) associated with each speaker. We propose an unsupervised technique based on audio-visual generative modeling of clean speech. More specifically, during training, a latent variable generative model is learned from clean speech spectrograms using a variational auto-encoder (VAE). To better utilize the visual information, the posteriors of the latent variables are inferred from mixed speech (instead of clean speech) as well as the visual data. The visual modality also serves as a prior for latent variables, through a visual network. At test time, the learned generative model (both for speaker-independent and speaker-dependent scenarios) is combined with an unsupervised non-negative matrix factorization (NMF) variance model for background noise. All the latent variables and noise parameters are then estimated by a Monte Carlo expectation-maximization algorithm. Our experiments show that the proposed unsupervised VAE-based method yields better separation performance than NMF-based approaches as well as a supervised deep learning-based technique.

7.9 A Benchmark of Dynamical Variational Autoencoders applied to Speech Spectrogram Modeling

Participants: Xiaoyu Bie, Laurent Girin, Xavier Alameda-Pineda.

The Variational Autoencoder (VAE) is a powerful deep generative model that is now extensively used to represent high-dimensional complex data via a low-dimensional latent space learned in an unsupervised manner. In the original VAE model, input data vectors are processed independently. In recent years, a series of papers have presented different extensions of the VAE to process sequential data, that not only model the latent space, but also model the temporal dependencies within a sequence of data vectors and corresponding latent vectors, relying on recurrent neural networks. We recently performed a comprehensive review of those models and unified them into a general class called Dynamical Variational Autoencoders (DVAEs). In the present work, we present the results of an experimental benchmark comparing six of those DVAE models on the speech analysis-resynthesis task, as an illustration of the high potential of DVAEs for speech modeling.

7.10 Variational Inference and Learning of Piecewise-linear Dynamical Systems

Participants: Xavier Alameda-Pineda, Radu Horaud.

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 work, 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, and we thoroughly compare our algorithms with several state of the art trackers.

7.11 SocialInteractionGAN: Multi-person Interaction Sequence Generation

Participants: Louis Airale, Dominique Vaufreydaz, Xavier Alameda-Pineda.

Prediction of human actions in social interactions has important applications in the design of social robots or artificial avatars. In this work, we model human interaction generation as a discrete multi-sequence generation problem and present SocialInteractionGAN, a novel adversarial architecture for conditional interaction generation. Our model builds on a recurrent encoder-decoder generator network and a dual-stream discriminator. This architecture allows the discriminator to jointly assess the realism of interactions and that of individual action sequences. Within each stream a recurrent network operating on short subsequences endows the output signal with local assessments, better guiding the forthcoming generation. Crucially, contextual information on interacting participants is shared among agents and reinjected in both the generation and the discriminator evaluation processes. We show that the proposed SocialInteractionGAN succeeds in producing high realism action sequences of interacting people, comparing favorably to a diversity of recurrent and convolutional discriminator baselines. Evaluations are conducted using modified Inception Score and Fréchet Inception Distance metrics, that we specifically design for discrete sequential generated data. The distribution of generated sequences is shown to approach closely that of real data. In particular our model properly learns the dynamics of interaction sequences, while exploiting the full range of actions.

7.12 ξ-Learning: Successor Feature Transfer Learning for General Reward Functions

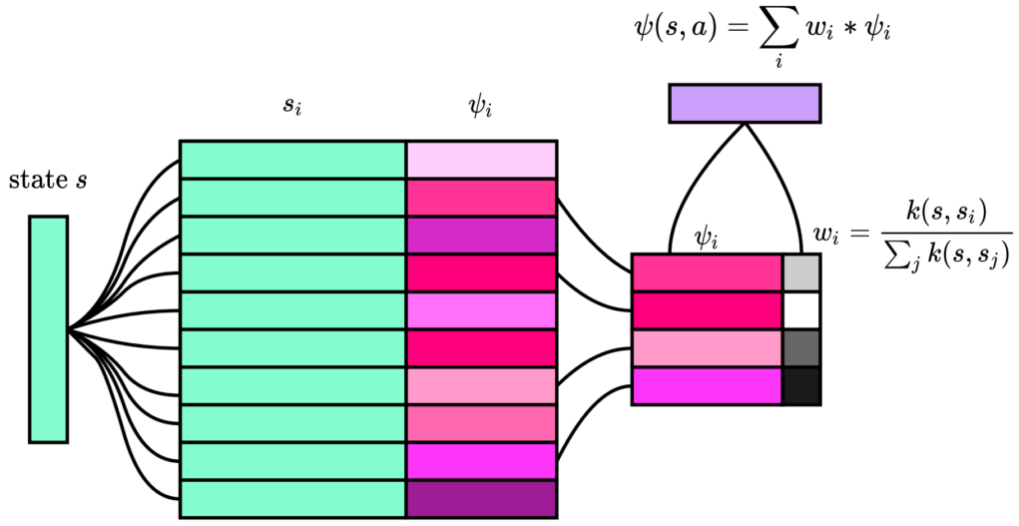

Figure 4: Neural epsodic control for successor features.

Participants: Chris Reinke, Xavier Alameda-Pineda.

Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor features (SF) are a prominent transfer mechanism in domains where the reward function changes between tasks. They reevaluate the expected return of previously learned policies in a new target task and to transfer their knowledge. A limiting factor of the SF framework is its assumption that rewards linearly decompose into successor features and a reward weight vector. We propose a novel SF mechanism, ξ-learning, based on learning the cumulative discounted probability of successor features. Crucially, ξ-learning allows to reevaluate the expected return of policies for general reward functions. We introduce two ξ-learning variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on ξ-learning with function approximation demonstrate the prominent advantage of ξ-learning over available mechanisms not only for general reward functions but also in the case of linearly decomposable reward functions.

7.13 Successor Feature Neural Episodic Control

Participants: David Emukpere, Xavier Alameda-Pineda, Chris Reinke.

A longstanding goal in reinforcement learning is to build intelligent agents that show fast learning and a flexible transfer of skills akin to humans and animals. We investigate the integration of two frameworks for tackling those goals: episodic control and successor features. Episodic control is a cognitively inspired approach relying on episodic memory, an instance-based memory model of an agent's experiences. Meanwhile, successor features and generalized policy improvement (SF&GPI) is a meta and transfer learning framework allowing to learn policies for tasks that can be efficiently reused for later tasks which have a different reward function. Individually, these two techniques have shown impressive results in vastly improving sample efficiency and the elegant reuse of previously learned policies. Thus, we outline a combination of both approaches in a single reinforcement learning framework and empirically illustrate its benefits.

8 Bilateral contracts and grants with industry

8.1 Bilateral Grants with Industry

8.1.1 VASP

Participants: Zhiqi Kang, Mostafa Sadeghi, Xavier Alameda-Pineda, Radu Horaud.

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 H2020 Project SPRING

Participants: Alex Auternaud, Timothée Wintz, Chris Reinke, Luis Camara, Gaetan Lepage, Nicolas Turro, Soraya Arias, Radu Horaud, Xavier Alameda-Pineda.

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

9.2 National initiatives

9.2.1 ANR JCJC Project ML3RI

Participants: Chris Reinke, Xiaoyu Lin, Radu Horaud, Xavier Alameda-Pineda.

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

Website

9.2.2 ANR MIAI Chair

Participants: Xiaoyu Bie, Anand Ballou, Radu Horaud, Xavier Alameda-Pineda.

The overall goal of the MIAI chair "Audio-visual machine perception & interaction for robots" it to enable socially-aware robot behavior for interactions with humans. Emphasis on unsupervised and weakly supervised learning with audio-visual data, Bayesian inference, deep learning, and reinforcement learning. Challenging proof-of-concept demonstrators. We aim to develop robots that explore populated spaces, understand human behavior, engage multimodal dialog with several users, etc. These tasks require audio and visual cues (e.g. clean speech signals, eye-gaze, head-gaze, facial expressions, lip movements, head movements, hand and body 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 reverberation and noise, overlapping audio sources, bad lighting, limited image resolution, narrow camera field of view, visual occlusions, etc. We will investigate audio and visual perception and communication, e.g. face-to-face dialog: the robot should learn how to collect clean data (e.g. frontal faces, signals with high speech-to-noise ratios) and how to react appropriately to human verbal and non-verbal solicitations. We plan to demonstrate these skills with a companion robot that assists and entertains the elderly in healthcare facilities.

Website: Website

10 Dissemination

Participants: Radu Horaud, Xavier Alameda-Pineda.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Area Chair Xavier Alameda-Pineda was Area Chair for IEEE/CVF WACV 2021, ACM Multimedia 2021, and AAAI 2022.

10.1.2 Scientific events: selection

Reviewer Xavier Alameda-Pineda was a reviewer for IEEE/CVF CVPR 2022 and for IEEE ICASSP 2022.

10.1.3 Journal

Member of the editorial boards During 2021, Xavier Alameda-Pineda was Associated Editor of three top-tier journals: Computer Vision and Image Understanding, ACM Transactions on Mutimedia Tools and Applications and IEEE Transactions on Multimedia.

10.1.4 Invited talks

During 2021, the members of the team were invited to give a series of talks:

- Deep generative modeling of sequential data with dynamical variational autoencoders (Jun'21) at IEEE ICASSP 2021.
- Unsupervised Learning for Human Robot Perception (Jun'21) at Robotics and AI Summer School 2021.
- Towards socially intelligent robots: preliminary results of the H2020 SPRING and the ANR ML3RI projects (Jun'21) at PI Stories University of Trento
- Unsupervised Audio-Visual Fusion for Upstream Human Behavior Understanding (May'21) at AI4Media Workshop on New Learning Paradigms and Distributed AI4Media
- Variational Autoencoders for Audio, Visual and Audio-Visual Learning (Feb'21) at DaSCI Webinars
- Speaker localisation and enhancement in populated environments – invited talk (Jan'21) at ICPR 2020 Workshop on Deep Learning for Human-Centric Activity Understanding
- Combining auditory and visual data to enhance the speech signal – invited talk (Jan'21) at ICPR 2020 Workshop on Multimodal pattern recognition for social signal processing in human computer interaction

10.1.5 Leadership within the scientific community

Since 2021, Xavier Alameda-Pineda is the vice-chair of the 9th Technical Committee of the International Association for Pattern Recognition with title "pattern recognition in human machine interaction."

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

In 2021, Xavier Alameda-Pineda was involved in teaching two coursers at Masters 2 level:

- Fundamentals of Probabilistic Data Mining - at Master of Science in Industrial and Applied Mathematics
- Machine Learning for Computer Vision and Audio Processing - at Master of Science in Informatics at Grenoble

10.2.2 Supervision (defences)

PhD defence: Guillaume Delorme, Adaptation de domaine non supervisée pour modèle de suivi multi-partie et identification visuelle appliquée à l'interaction homme-robot, defended on October 8th, 2021. Directors: Radu Horaud and Xavier Alameda-Pineda.

MSc defence: David Emukpere, Successor Feature Neural Episodic Control, defended on June 22nd, 2021. Directors: Xavier Alameda-Pineda and Chris Reinke.

10.2.3 Juries

In 2021, Xavier Alameda-Pineda participated to the following PhD committees as examiner:

- Julien Audibert (U. Sorbonne)
- Maria Kabtoul (University Grenoble-Alpes)

and to the following ones as a reviewer:

- Manuel Pariente (U. Lorraine)
- Marco Godi (U. Verona)

11 Scientific production

11.1 Major publications

[1] 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 (1st June 2019), pp. 798–802. DOI: 10.1109/LSP.2019.2908376. URL: https://hal.inria.fr/hal-01969050.

[2] 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 (1st June 2018), pp. 1397–1410. DOI: 10.1109/TPAMI.2017.2717829. URL: https://hal.inria.fr/hal-01413414.

[3] 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 (2nd July 2018), pp. 1086–1099. DOI: 10.1109/TPAMI.2017.2648793. URL: https://hal.inria.fr/hal-01413403.

[4] 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 (1st Feb. 2019), pp. 61–71. DOI: 10.1016/j.patrec.2018.05.023. URL: https://hal.inria.fr/hal-01643775.

[5] 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.

[6] 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 (8th Mar. 2019), pp. 88–103. DOI: 10.1109/JSTSP.2019.2903472. URL: https://hal.inria.fr/hal-01851985.

[7] X. Li, S. Gannot, L. Girin and R. Horaud. 'Multichannel Identification and Nonnegative Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26.10 (21st May 2018), pp. 1755–1768. DOI: 10.1109/TASLP.2018.2839362. URL: https://hal.inria.fr/hal-01645749.

[8] 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 (1st Mar. 2019), pp. 645–659. DOI: 10.1109/TASLP.2019.2892412. URL: https://hal.inria.fr/hal-01799809.

[9] 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.

[10] 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 (1st Nov. 2018), pp. 2711–2724. DOI: 10.1109/TPAMI.2017.2782819. URL: https://hal.inria.fr/hal-01511414.

[11] 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.

[12] A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci and N. Sebe. 'Increasing Image Memorability with Neural Style Transfer'. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 15.2 (1st June 2019). DOI: 10.1145/3311781. URL: https://hal.inria.fr/hal-01858389.

[13] D. Xu, X. Alameda-Pineda, W. Ouyang, E. Ricci, X. Wang and N. Sebe. 'Probabilistic Graph Attention Network with Conditional Kernels for Pixel-Wise Prediction'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (14th Dec. 2020). DOI: 10.1109/TPAMI.2020.3043781. URL: https://hal.inria.fr/hal-03328687.

11.2 Publications of the year

International journals

[14] 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.

[15] 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 (1st May 2021), pp. 1761–1776. DOI: 10.1109/TPAMI.2019.2953020. URL: https://hal.inria.fr/hal-01950866.

[16] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber and X. Alameda-Pineda. 'Dynamical Variational Autoencoders: A Comprehensive Review'. In: *Foundations and Trends in Machine Learning* 15.1-2 (2nd Dec. 2021), pp. 1–175. DOI: 10.1561/2200000089. URL: https://hal.inria.fr/hal-02926215.

International peer-reviewed conferences

[17] X. Bie, L. Girin, S. Leglaive, T. Hueber and X. Alameda-Pineda. 'A Benchmark of Dynamical Variational Autoencoders applied to Speech Spectrogram Modeling'. In: Interspeech 2021 - 22nd Annual Conference of the International Speech Communication Association. Brno, Czech Republic, 30th Aug. 2021, pp. 46–50. DOI: 10.21437/Interspeech.2021-256. URL: https://hal.inria.fr/hal-03295657.

[18] G. Delorme, Y. Xu, S. Lathuilière, R. Horaud and X. Alameda-Pineda. 'CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification'. In: ICPR 2020 - 25th International Conference on Pattern Recognition. Milano, Italy: IEEE, 2021, pp. 4428–4435. DOI: 10.1109/ICPR48806.2021.9412431. URL: https://hal.inria.fr/hal-02882285.

[19] D. Emukpere, X. Alameda-Pineda and C. Reinke. 'Successor Feature Neural Episodic Control'. In: NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. Virtual, Canada, 6th Dec. 2021, pp. 1–12. URL: https://hal.inria.fr/hal-03426874.

[20] S. Guy, S. Lathuilière, P. Mesejo and R. Horaud. 'Learning Visual Voice Activity Detection with an Automatically Annotated Dataset'. In: ICPR 2020 - 25th International Conference on Pattern Recognition. Milano, Italy: IEEE, 10th Jan. 2021, pp. 4851–4856. DOI: 10.1109/ICPR48806.2021.9412884. URL: https://hal.inria.fr/hal-02882229.

[21] 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: ICASSP 2021 - IEEE International Conference on Acoustics, Speech, and Signal Processing. Toronto, Canada: IEEE, 6th June 2021, pp. 1–5. URL: https://hal.archives-ouvertes.fr/hal-03135727.

[22] Z. Kang, R. Horaud and M. Sadeghi. 'Robust Face Frontalization For Visual Speech Recognition'. In: ICCVW 2021 - International Conference on Computer Vision Workshops. Montreal - Virtual, Canada: IEEE, 11th Oct. 2021, pp. 2485–2495. DOI: 10.1109/ICCVW54120.2021.00281. URL: https://hal.inria.fr/hal-03326002.

[23] V.-N. Nguyen, M. Sadeghi, E. Ricci and X. Alameda-Pineda. 'Deep Variational Generative Models for Audio-visual Speech Separation'. In: IEEE International Workshop on Machine Learning for Signal Processing. Gold Coast, Australia, Oct. 2021. URL: https://hal.inria.fr/hal-02930662.

Reports & preprints

[24] X. Bie, S. Leglaive, X. Alameda-Pineda and L. Girin. *Unsupervised Speech Enhancement using Dynamical Variational Auto-Encoders*. 22nd July 2021. URL: https://hal.inria.fr/hal-03295630.

[25] W. Guo, X. Bie, X. Alameda-Pineda and F. Moreno-Noguer. *Multi-Person Extreme Motion Prediction with Cross-Interaction Attention*. 22nd July 2021. URL: https://hal.inria.fr/hal-03295672.

[26] C. Reinke and X. Alameda-Pineda. *Xi-Learning: Successor Feature Transfer Learning for General Reward Functions*. 12th Nov. 2021. URL: https://hal.inria.fr/hal-03426870.

[27] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus and X. Alameda-Pineda. *TransCenter: Transformers with Dense Queries for Multiple-Object Tracking*. 22nd July 2021. URL: https://hal.inria.fr/hal-03295680.

[28] G. Yang, P. Rota, X. Alameda-Pineda, D. Xu, M. Ding and E. Ricci. *Variational Structured Attention Networks for Deep Visual Representation Learning*. 22nd July 2021. URL: https://hal.inria.fr/hal-03296152.

11.3 Cited publications

[29] J. Abdi, A. Al-Hindawi, T. Ng and M. P. Vizcaychipi. 'Scoping review on the use of socially assistive robot technology in elderly care'. In: *BMJ open* 8.2 (2018), e018815.

[30] T. Afouras, A. Owens, J. S. Chung and A. Zisserman. 'Self-supervised learning of audio-visual objects from video'. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020, pp. 208–224.

[31] X. Alameda-Pineda, S. Arias, Y. Ban, G. Delorme, L. Girin, R. Horaud, X. Li, B. Mourgue and G. Sarrazin. 'Audio-Visual Variational Fusion for Multi-Person Tracking with Robots'. In: *ACMMM 2019 - 27th ACM International Conference on Multimedia*. Nice, France: ACM Press, Oct. 2019, pp. 1059–1061. DOI: 10.1145/3343031.3350590. URL: https://hal.inria.fr/hal-02354514.

[32] 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-00910081.

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

[34] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp. 'A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking'. In: *IEEE Transactions on signal processing* 50.2 (2002), pp. 174–188.

[35] 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.

[36] 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: IEEE, Sept. 2017, pp. 4144–4151. DOI: 10.1109/IROS.2017.8206274. URL: https://hal.inria.fr/hal-01542987.

[37] 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.

[38] T. W. Bickmore, L. M. Pfeifer and M. K. Paasche-Orlow. 'Using computer agents to explain medical documents to patients with low health literacy'. In: *Patient education and counseling* 75.3 (2009), pp. 315–320.

[39] J. Broekens, M. Heerink, H. Rosendal et al. 'Assistive social robots in elderly care: a review'. In: *Gerontechnology* 8.2 (2009), pp. 94–103.

[40] D. Brščić, H. Kidokoro, Y. Suehiro and T. Kanda. 'Escaping from children's abuse of social robots'. In: *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. 2015, pp. 59–66.

[41] W.-L. Chang, J. P. White, J. Park, A. Holm and S. Šabanović. 'The effect of group size on people's attitudes and cooperative behaviors toward robots in interactive gameplay'. In: *RO-MAN International Symposium on Robot and Human Interactive Communication*. IEEE. 2012, pp. 845–850.

[42] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson and K. Grauman. 'Soundspaces: Audio-visual navigation in 3d environments'. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer. 2020, pp. 17–36.

[43] A. Deleforge, F. Forbes, S. Ba and R. Horaud. 'Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies'. In: *IEEE Journal of Selected Topics in Signal Processing* 9.6 (Sept. 2015), pp. 1037–1048. DOI: 10.1109/JSTSP.2015.2416677. URL: https://hal.inria.fr/hal-01136465.

[44] 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), 21p. DOI: 10.1142/S0129065714400036. URL: https://hal.inria.fr/hal-00960796.

[45] 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.

[46] 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.

[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, Jan. 2021. URL: https://hal.inria.fr/hal-03188744.

[48] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge and R. Horaud. 'Head Pose Estimation via Probabilistic High-Dimensional Regression'. In: *IEEE International Conference on Image Processing, ICIP 2015*. Proceedings of the IEEE International Conference on Image Processing. Quebec City, QC, Canada: IEEE, Sept. 2015, pp. 4624–4628. DOI: 10.1109/ICIP.2015.7351683. URL: https://hal.inria.fr/hal-01163663.

[49] 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.

[50] 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), pp. 1397–1410. DOI: 10.1109/TPAMI.2017.2717829. URL: https://hal.inria.fr/hal-01413414.

[51] L. Fang, T. Zeng, C. Liu, L. Bo, W. Dong and C. Chen. 'Transformer-based Conditional Variational Autoencoder for Controllable Story Generation'. In: *arXiv preprint arXiv:2101.00828* (2021).

[52] M. E. Foster, A. Gaschler and M. Giuliani. 'Automatically classifying user engagement for dynamic multi-party human–robot interaction'. In: *International Journal of Social Robotics* 9.5 (2017), pp. 659–674.

[53] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi and T. S. Huang. 'Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6112–6121.

[54] R. Gao and K. Grauman. 'Visualvoice: Audio-visual speech separation with cross-modal consistency'. In: *IEEE/CVF CVPR*. 2021.

[55] 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), pp. 1086–1099. DOI: 10.1109/TPAMI.2017.2648793. URL: https://hal.inria.fr/hal-01413403.

[56] 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.

[57] L. Girin, T. Hueber and X. Alameda-Pineda. 'Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25.3 (Mar. 2017), pp. 662–673. DOI: 10.1109/TASLP.2017.2651398. URL: https://hal.archives-ouvertes.fr/hal-01485540.

[58] M. Gombolay, X. J. Yang, B. Hayes, N. Seo, Z. Liu, S. Wadhwania, T. Yu, N. Shah, T. Golen and J. Shah. 'Robotic assistance in the coordination of patient care'. In: *The International Journal of Robotics Research* 37.10 (2018), pp. 1300–1316.

[59] T. Haarnoja, A. Ajay, S. Levine and P. Abbeel. 'Backprop kf: Learning discriminative deterministic state estimators'. In: *Advances in neural information processing systems*. 2016, pp. 4376–4384.

[60] J. Huh, H. S. Heo, J. Kang, S. Watanabe and J. S. Chung. 'Augmentation adversarial training for self-supervised speaker recognition'. In: *arXiv preprint arXiv:2007.12085* (2020).

[61] N. Karnchanachari, M. I. Valls, D. Hoeller and M. Hutter. 'Practical Reinforcement Learning For MPC: Learning from sparse objectives in under an hour on a real robot'. In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 211–224.

[62] D. P. Kingma and M. Welling. 'Auto-encoding variational bayes'. In: *arXiv preprint arXiv:1312.6114* (2013).

[63] 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.

[64] J. Lafaye, C. Collette and P.-B. Wieber. 'Model predictive control for tilt recovery of an omnidirectional wheeled humanoid robot'. In: *International conference on robotics and automation (ICRA)*. IEEE. 2015, pp. 5134–5139.

[65] S. Lathuilière, B. Massé, P. Mesejo and R. Horaud. 'Deep Reinforcement Learning for Audio-Visual Gaze Control'. In: *IROS 2018 - IEEE/RSJ International Conference on Intelligent Robots and Systems*. Madrid, Spain: IEEE, Oct. 2018, pp. 1555–1562. DOI: 10.1109/IROS.2018.8594327. URL: https://hal.inria.fr/hal-01851738.

[66] 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.

[67] S. Lathuilière, P. Mesejo, X. Alameda-Pineda and R. Horaud. 'DeepGUM: Learning Deep Robust Regression with a Gaussian-Uniform Mixture Model'. In: *ECCV 2018 - European Conference on Computer Vision*. Vol. 11209. Lecture Notes in Computer Science. Munich, Germany: Springer, Sept. 2018, pp. 205–221. DOI: 10.1007/978-3-030-01228-1_13. URL: https://hal.inria.fr/hal-01851511.

[68] 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*. IEEE. Barcelone, Spain, May 2020, pp. 1–7. DOI: 10.1109/ICASSP40776.2020.9053164. URL: https://hal.archives-ouvertes.fr/hal-02329000.

[69] S. Leglaive, L. Girin and R. Horaud. 'A variance modeling framework based on variational autoencoders for speech enhancement'. In: *EEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Aalborg, Denmark: IEEE, Sept. 2018, pp. 1–6. DOI: 10.1109/MLSP.2018.8516711. URL: https://hal.inria.fr/hal-01832826.

[70] S. Leglaive, L. Girin and R. Horaud. 'Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization'. In: *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*. Brighton, United Kingdom: IEEE, May 2019, pp. 101–105. DOI: 10.1109/ICASSP.2019.8683704. URL: https://hal.inria.fr/hal-02005102.

[71] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud. 'A Cascaded Multiple-Speaker Localization and Tracking System'. In: *Proceedings of the LOCATA Challenge Workshop - a satellite event of IWAENC 2018*. Tokyo, Japan, Sept. 2018, pp. 1–5. URL: https://hal.inria.fr/hal-01957137.

[72] 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.

[73] X. Li, S. Gannot, L. Girin and R. Horaud. 'Multichannel Identification and Nonnegative Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26.10 (May 2018), pp. 1755–1768. DOI: 10.1109/TASLP.2018.2839362. URL: https://hal.inria.fr/hal-01645749.

[74] 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.

[75] 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.

[76] 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.

[77] X. Li, L. Girin and R. Horaud. 'Expectation-Maximization for Speech Source Separation using Convolutive Transfer Function'. In: *CAAI Transactions on Intelligent Technologies* 4.1 (Mar. 2019), pp. 47–53. DOI: 10.1049/trit.2018.1061. URL: https://hal.inria.fr/hal-01982250.

[78] 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.

[79] 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.

[80] X. Li and R. Horaud. 'Narrow-band Deep Filtering for Multichannel Speech Enhancement'. working paper or preprint. Sept. 2020. URL: https://hal.inria.fr/hal-02378413.

[81] 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.

[82] J. A. Mann, B. A. MacDonald, I.-H. Kuo, X. Li and E. Broadbent. 'People respond better to robots than computer tablets delivering healthcare instructions'. In: *Computers in Human Behavior* 43 (2015), pp. 112–117.

[83] 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), pp. 2711–2724. DOI: 10.1109/TPAMI.2017.2782819. URL: https://hal.inria.fr/hal-01511414.

[84] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al. 'Human-level control through deep reinforcement learning'. In: *nature* 518.7540 (2015), pp. 529–533.

[85] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[86] A. Nagabandi, G. Kahn, R. S. Fearing and S. Levine. 'Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning'. In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 7559–7566.

[87] M. Pino, M. Boulay, F. Jouen and A. S. Rigaud. '"Are we ready for robots that care for us?" Attitudes and opinions of older adults toward socially assistive robots'. In: *Frontiers in aging neuroscience* 7 (2015), p. 141.

[88] S. Roweis and Z. Ghahramani. 'Learning nonlinear dynamical systems using the expectation-maximization algorithm'. In: *Kalman filtering and neural networks* 6 (2001), pp. 175–220.

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

[90] 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. DOI: 10.1109/ICASSP40776.2020.9053730. URL: https://hal.archives-ouvertes.fr/hal-02534911.

[91] 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: IEEE, June 2021, pp. 1–5. URL: https://hal.inria.fr/hal-03155445.

[92] M. Sadeghi, S. Guy, A. Raison, X. Alameda-Pineda and R. Horaud. 'Unsupervised Performance Analysis of 3D Face Alignment'. Submitted to Computer Vision and Image Understanding. Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02543069.

[93] 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 (May 2020), pp. 1788–1800. DOI: 10.1109/TASLP.2020.3000593. URL: https://hal.inria.fr/hal-02364900.

[94] S. Sebo, B. Stoll, B. Scassellati and M. F. Jung. 'Robots in groups and teams: a literature review'. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–36.

[95] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[96] Y.-H. Wu, V. Cristancho-Lacroix, C. Fassert, V. Faucounau, J. de Rotrou and A.-S. Rigaud. 'The attitudes and perceptions of older adults with mild cognitive impairment toward an assistive robot'. In: *Journal of Applied Gerontology* 35.1 (2016), pp. 3–17.

[97] Y.-H. Wu, J. Wrobel, M. Cornuet, H. Kerhervé, S. Damnée and A.-S. Rigaud. 'Acceptance of an assistive robot in older adults: a mixed-method study of human–robot interaction over a 1-month period in the Living Lab setting'. In: *Clinical interventions in aging* 9 (2014), p. 801.

[98] M. Żarkowski. 'Multi-party turn-taking in repeated human–robot interactions: an interdisciplinary evaluation'. In: *International Journal of Social Robotics* 11.5 (2019), pp. 693–707.

[99] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker and W. Burgard. 'Vr-goggles for robots: Real-to-sim domain adaptation for visual control'. In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1148–1155.