RESEARCH CENTRE

Inria Centre at Université Grenoble Alpes

IN PARTNERSHIP WITH:

Université de Grenoble Alpes

2024

ACTIVITY REPORT

Project-Team

ROBOTLEARN

Learning, perception and control for social robots

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Inria

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, until Sep 2024]
- Xavier Alameda Pineda [Team leader, INRIA, Senior Researcher, from Oct 2024]
- Patrice Horaud [INRIA, Emeritus]
- Chris Reinke [INRIA, Starting Research Position, until Aug 2024]

Post-Doctoral Fellows

- Xiaoyu Lin [UGA, Post-Doctoral Fellow, from Aug 2024]
- Samir Sadok [UGA, Post-Doctoral Fellow, from May 2024]

PhD Students

- Anand Ballou [INRIA, until Feb 2024]
- Gaetan Lepage [INRIA]
- Xiaoyu Lin [INRIA]

Technical Staff

- Alex Auternaud [INRIA, Engineer, until Apr 2024]
- Ahamed Mohamed [INRIA, Engineer, from Oct 2024]
- Kirubakaran Ramamoorthy [INRIA, Engineer, until Jun 2024]
- Victor Sanchez [INRIA, Engineer, until Apr 2024]

Interns and Apprentices

- Ghazi Shazan Ahmad [INRIA, Intern, until Jun 2024]
- Daniel Jost [INRIA, Intern, until Feb 2024]
- Estelle Long-Merle [UGA, Intern, until Jan 2024]

Administrative Assistant

- Nathalie Gillot [INRIA]

External Collaborators

- Laurent Girin [GRENOBLE INP]
- Thomas Hueber [CNRS, from Sep 2024]
- Olivier Perrotin [CNRS, from Nov 2024]

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.

Nevertheless, today's state-of-the-art in robotics is not well-suited to fulfill these needs, and there are two main bottlenecks: (i) robots are limited to a handful of simple scenarios which leads to (ii) social robots not being well accepted by a large percentage of users. 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 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. 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. We identify three main 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.

ROBOTLEARN stands at the cross-roads of several fields: 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.

From a methodological perspective, the challenge is at least three-fold. First, to reduce the amount of human intervention needed to adapt the designed learning models in a new environment. We aim to further develop strategies based on unsupervised learning and unsupervised domain adaptation, within the framework of deep probabilistic modling with latent variables [38]. Second, to successfully exploit auditory and visual data for human behavior understanding. For instance by developing mechanisms that manage to model and learn the complementarity between sounds and images [6]. Third, by developping reinforcement learning algorithms that can transfer previous knowledge to future tasks and environments. One potential way forward is to anchor the learning into key features that can be hand-crafted or learned [28].

3 Research program

ROBOTLEARN will be structured in three research axes, allowing to develop socially intelligent robots. 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 [43] or the use of unsupervised models [31]. 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 the two functions are linear, the model boils down to a linear dynamical system, that can be learned with an exact Expectation-Maximization (EM) algorithm. Beyond this simple case, non-linearity can be achieved via mixtures of K linear models or more general non-linear (e.g. deep neural) functions. Either case, learning and inference cannot be exact and must be approximated, either by using variational EM algorithms [30, 39, 32, 2], amortized variational inference [38, 27] or a combination of both techniques [17, 9].

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: deep probabilistic sequential modeling We have investigated a latent-variable generative model called mixture of dynamical variational autoencoders (MixDVAE) to model the dynamics of a system composed of multiple moving sources. A DVAE model is pre-trained on a single-source dataset to capture the source dynamics. Then, multiple instances of the pre-trained DVAE model are integrated into a multi-source mixture model with a discrete observation-to-source assignment latent variable. The posterior distributions of both the discrete observation-to-source assignment variable and the continuous DVAE variables representing the sources content/position are estimated using the variational expectation-maximization algorithm, leading to multi-source trajectories estimation. We illustrated the versatility of the proposed MixDVAE model on two tasks: a computer vision task, namely multi-object tracking, and an audio processing task, namely single-channel audio source separation. Consequently, this mixture models allows to mix different non-linear source models within the maximum likelihood umbrella and combine the model with other probabilistic models as well.

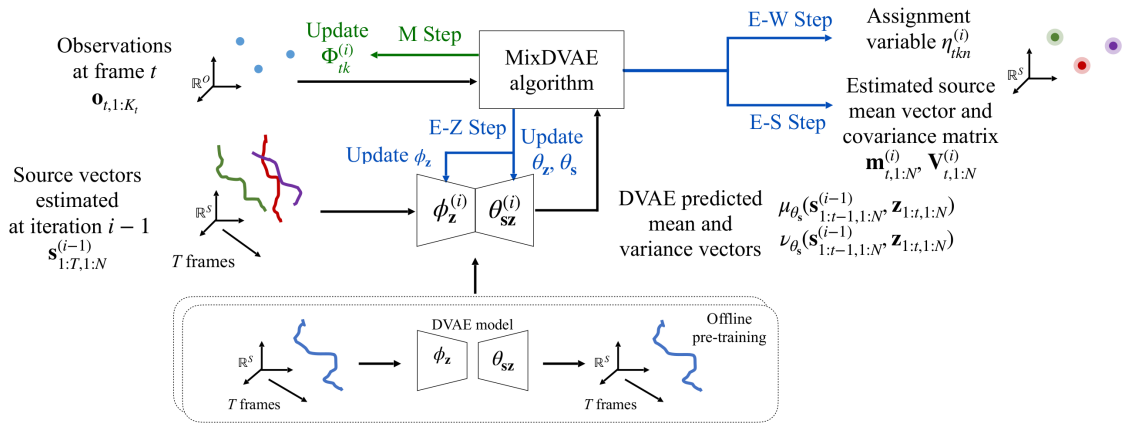

Figure 1: MixDVAE overall diagram.

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 [29, 35, 37], 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.

It is also 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. Either way we will need some sort of on-line learning to perform final adaptation. 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: expression-preserving face frontalization Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. We proposed a frontalization methodology that preserves non-rigid facial deformations in order to boost the performance of visually assisted speech communication. The method alternates between the estimation of (i) the rigid transformation (scale, rotation, and translation) and (ii) the non-rigid deformation between an arbitrarily-viewed face and a face model. The method has two important merits: it can deal with non-Gaussian errors in the data and it incorporates a dynamical face deformation model. For that purpose, we used the generalized Student t-distribution in combination with a linear dynamic system in order to account for both rigid head motions and time-varying facial deformations caused by speech production. We proposed to use the zero-mean normalized cross-correlation (ZNCC) score to evaluate the ability of the method to preserve facial expressions. We showed that the method, when incorporated into deep learning pipelines, namely lip reading and speech enhancement, improves word recognition and speech intelligibility scores by a considerable margin.

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 [40]. 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 [34] or even bullying [33]. 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 [42] or very specific robot roles [36]. 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.

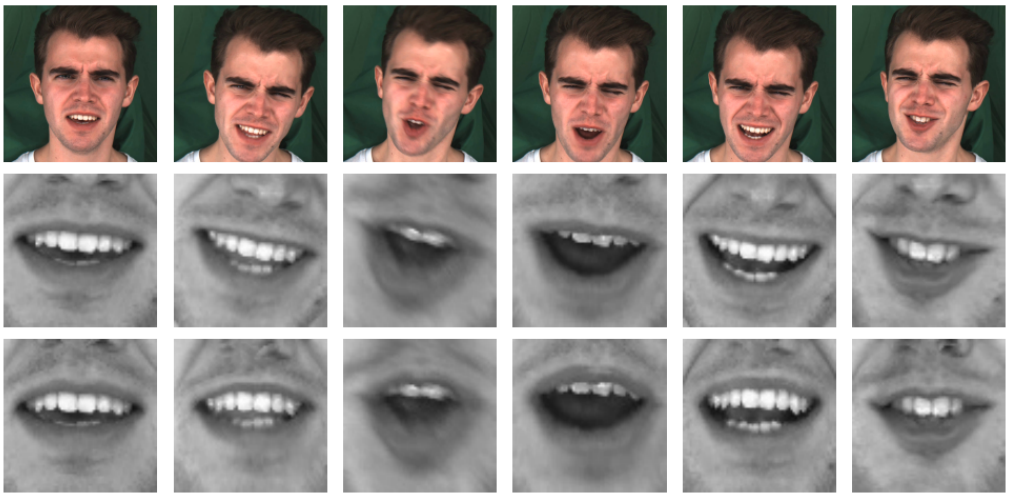

Figure 2: Some results of the proposed expression-preserving face frontalization method.

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.

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[41]. 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.

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.

Exemplar application: transfering poilicies via successor feature representations Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor Representations (SR) and their extension Successor Features (SF) are prominent transfer mechanisms in domains where reward functions change between tasks. They reevaluate the expected return of previously learned policies in a new target task to transfer their knowledge. The SF framework extended SR by linearly decomposing rewards into successor features and a reward weight vector allowing their application in high-dimensional tasks. But this came with the cost of having a linear relationship between reward functions and successor features, limiting its application to tasks where such a linear relationship exists. We proposed a novel formulation of SR based on learning the cumulative discounted probability of successor features, called Successor Feature Representations (SFR). Crucially, SFR allows to reevaluate the expected return of policies for general reward functions. We introduced different SFR variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on SFR with function approximation demonstrate its advantage over SF not only for general reward functions, but also in the case of linearly decomposable reward functions.

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

Figure 3: The ARI robot from PAL Robotics.

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 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.

5 Social and environmental responsibility

5.1 Impact of research results

Our line of research on developing unsupervised learning methods exploiting audio-visual data to understand social scenes and to learn to interact within is very interesting and challenging, and has large economical and societal impact. Economical impact since the auditory and visual sensors are the most common one, and we can find (many of) them in almost every smartphone in the market. Beyond telephones, manufacturers designing new systems meant for human use, should take into account the need for verbal interaction, and hence for audio-visual perception. A clear example of this potential is the transfer of our technology to a real robotic platform, for evaluation within a day-care hospital (DCH). This is possible thanks to the H2020 SPRING EU project, that assesses the interest of social robotics in the non-medical phases of a regular day for elder patients in a DCH. We are evaluating the performance of our methods for AV speaker tracking, AV speech enhancement, and AV sound source separation, for future technology transfer to the robot manufacturer. This is the first step toward a robot that can be part of the social environment of the DCH, helping to reduce patient and companion stress, at the same time as being a useful tool for the medical personnel. We are confident that developing robust AV perception and action capabilities for robots and autonomous systems, will make them more suitable for environments populated with humans.

6 Highlights of the year

6.1 Hiring and promotions

A new permanent team member, Dr. Stéphane Lathuilière, joined the team as Inria Starting Faculty Position. Also, Xavier Alameda-Pineda was promoted to Directeur de Recherche de 2eme classe.

6.2 Keynote Speaker at RFIAP/CAp 2024

Xavier Alameda-Pineda was invited to give a keynote at RFIAP/CAp 2024 on the topic "Learning for Companion Robots: Preparation and Adaptation."

6.3 PhD Defences

Anand Ballou defended his PhD on "Meta-reinforcement learning for social robotics." Social robots aim to assist, cooperate, take part in collaboration tasks and engage with human users. This requires the robotics system to interact within an unknown environment. The current predominant way to tackle this challenge relies on machine learning tools and algorithms. Machine learning approaches allow the robot to learn information from its environment. A popular machine-learning approach for social robotics is reinforcement learning. Reinforcement learning (RL) is a framework for solving decision-making problems in which an agent learns to interact with its environment through a trial-and-error process. By interacting with its environment, the agent will receive feedback that will either reward or penalize the taken action. The goal of reinforcement learning algorithms is to learn an action selection strategy (policy), that will maximize the reward the agent will receive. RL is well suited to solve decision problems such as robot social interaction environments, where labeled datasets are usually unavailable. However, applying reinforcement learning directly to a social environment is challenging due to several factors. One such factor is the diversity of environments a social agent will face when deployed, and the necessity for the agent to adapt to different user preferences. However, it is notoriously challenging for reinforcement learning agents to adapt to environments with different dynamics or reward functions.

To help tackle this issue, this thesis investigates two ways to improve reinforcement learning agents' adaptability. In the first part, we study the usability of meta-reinforcement learning approaches in a social robotics context. In this respect, we provide improvements to a state-of-the-art meta-reinforcement learning algorithm to generate more diverse behaviors. In the second part, we study how to directly use user feedback for fast adaptation to new unknown user preferences. To accomplish this, we propose to integrate meta-reinforcement learning in the classical preference-based learning framework to build a robust and time-saving algorithm for preference-based learning in the context of social robotics. We moreover benchmark our approach on a suite of social reinforcement learning environments, which allow us to test our algorithm on various social tasks with different settings and complex user preferences.

Xiaoyu Lin defended her PhD on "Deep latent-variable generative models for multimedia processing." Deep probabilistic generative models hold a crucial position within the realm of machine learning research. They serve as powerful tools for comprehending complex real-world data, such as image, audio, and text, by modeling their underlying distributions. This capability further enables the generation of new data samples. Moreover, these models can be utilized to discover hidden structures and the intrinsic factors of variation within data. The data representations that are learned through this process can be leveraged across a spectrum of downstream prediction tasks, thereby enhancing the decision-making process. Another research direction involves leveraging the flexibility and robust generalization ability of deep probabilistic generative models for solving intricate scientific and engineering problems. Though supervised deep learning methods applied to sophisticatedly designed neural architectures have achieved state-of-the-art performance across various domains, their practical application to real-world situations remains constrained. These limitations arise from the necessity of extensive volumes of annotated data for training and a shortfall in model interpretability. In this PhD work, we explore an alternative approach using deep probabilistic generative models within an unsupervised or weakly supervised framework to overcome these hurdles. Specifically, the proposed approach involves initially pre-training a deep probabilistic generative model with natural or synthetic signals to embed prior knowledge about the complex data patterns. Subsequently, this pre-trained model is integrated into an extended latent variable generative model (LVGM) to address the specific practical problem. Our research focuses on a specific type of deep probabilistic generative model designed for sequential data, referred to as dynamical variational auto-encoders (DVAEs). DVAEs are a family of deep latent variable models extended from the variational auto-encoder (VAE) for sequential data modeling. They leverage a sequence of latent vectors to depict the intricate temporal dependencies within the sequential observed data. By integrating DVAEs within a LVGM, we address a range of audio and visual tasks, namely multi-object tracking, single-channel audio source separation, and speech enhancement. The solutions are derived based on variational inference methods. Additionally, we also investigate a novel architecture, HiT-DVAE, which incorporates the Transformer architecture within the probabilistic framework of DVAEs. HiT-DVAE and its variant, LigHT-DVAE, both demonstrate excellent performance in speech modeling through robust sequential data handling. The findings from our experiments confirm the potential of deep probabilistic generative models to address real-world problems with limited labeled data, offering scalable and interpretable solutions. Furthermore, the introduction of HiT-DVAE represents a significant advancement in the field, combining the strengths of Transformer architectures with probabilistic modeling for enhanced sequential data analysis. These works not only contribute to the theoretical understanding of deep generative models, but also demonstrate their practical applicability across various domains, laying the groundwork for future innovations in machine learning.

7 New software, platforms, open data

7.1 New software

7.1.1 xi_learning

Name: Successor Feature Representation 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/sfr_learning

Contact: Chris Reinke

7.1.2 Social MPC

Keyword: Navigation

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

URL: https://gitlab.inria.fr/spring/wp6_robot_behavior/robot_behavior

Contact: Alex Auternaud

7.1.3 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/multiparty_interaction_simulator

Contact: Alex Auternaud

7.1.4 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

Contact: Xavier Alameda Pineda

7.1.5 exputils

Name: experiment utilities

Keywords: Python, Toolbox, Computer Science

Functional Description: Experiment Utilities (exputils) contains various tools for the management of scientific experiments and their experimental data. It is especially designed to handle experimental repetitions, including to run different repetitions, to effectively store and load data for them, and to visualize their results.

Main features: Easy definition of default configurations using nested python dictionaries. Setup of experimental configuration parameters using an ODF file. Running of experiments and their repetitions in parallel. Logging of experimental data (numpy, json). Loading and filtering of experimental data. Interactive Jupyter widgets to load, select and plot data as line, box and bar plots.

URL: https://gitlab.inria.fr/creinke/exputils

Contact: Chris Reinke

7.1.6 MixDVAE

Name: Source code for the article "Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation"

Keywords: Variational Autoencoder, Finite mixture

Functional Description: In this paper, we propose a latent-variable generative model called mixture of dynamical variational autoencoders (MixDVAE) to model the dynamics of a system composed of multiple moving sources. A DVAE model is pre-trained on a single-source dataset to capture the source dynamics. Then, multiple instances of the pre-trained DVAE model are integrated into a multi-source mixture model with a discrete observation-to-source assignment latent variable. The posterior distributions of both the discrete observation-to-source assignment variable and the continuous DVAE variables representing the sources content/position are estimated using a variational expectation-maximization algorithm, leading to multi-source trajectories estimation. We illustrate the versatility of the proposed MixDVAE model on two tasks: a computer vision task, namely multi-object tracking, and an audio processing task, namely single-channel audio source separation. Experimental results show that the proposed method works well on these two tasks, and outperforms several baseline methods.

URL: https://gitlab.inria.fr/robotlearn/dvae-umot-release

Contact: Xiaoyu Lin

7.1.7 Light-DVAE

Keyword: Variational Autoencoder

Functional Description: The dynamical variational autoencoders (DVAEs) are a family of latent-variable deep generative models that extends the VAE to model a sequence of observed data and a corresponding sequence of latent vectors. In almost all the DVAEs of the literature, the temporal dependencies within each sequence and across the two sequences are modeled with recurrent

neural networks. In this paper, we propose to model speech signals with the Hierarchical Transformer DVAE (HiT-DVAE), which is a DVAE with two levels of latent variable (sequence-wise and frame-wise) and in which the temporal dependencies are implemented with the Transformer architecture. We show that HiT-DVAE outperforms several other DVAEs for speech spectrogram modeling, while enabling a simpler training procedure, revealing its high potential for downstream low-level speech processing tasks such as speech enhancement.

URL: https://gitlab.inria.fr/robotlearn/light-dvae

Contact: Xiaoyu Lin

7.1.8 DDGM-SE

Keywords: Speech processing, Generative Models

Functional Description: This work builds on a previous work on unsupervised speech enhancement using a dynamical variational autoencoder (DVAE) as the clean speech model and non-negative matrix factorization (NMF) as the noise model. We propose to replace the NMF noise model with a deep dynamical generative model (DDGM) depending either on the DVAE latent variables, or on the noisy observations, or on both. This DDGM can be trained in three configurations: noise-agnostic, noise-dependent and noise adaptation after noise-dependent training. Experimental results show that the proposed method achieves competitive performance compared to state-of-the-art unsupervised speech enhancement methods, while the noise-dependent training configuration yields a much more time-efficient inference process.

URL: https://gitlab.inria.fr/robotlearn/ToBeProcessed/ddgm_se

Contact: Xiaoyu Lin

8 New results

The new results listed below are organised by research axis.

8.1 Deep Probabilistic Models

8.1.1 A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning

Participants: Samir Sadok, Laurent Girin, Xavier Alameda-Pineda.

High-dimensional data such as natural images or speech signals exhibit some form of regularity, preventing their dimensions from varying independently. This suggests that there exists a lower dimensional latent representation from which the high-dimensional observed data were generated. Uncovering the hidden explanatory features of complex data is the goal of representation learning, and deep latent variable generative models have emerged as promising unsupervised approaches. In particular, the variational autoencoder (VAE) which is equipped with both a generative and an inference model allows for the analysis, transformation, and generation of various types of data. Over the past few years, the VAE has been extended to deal with data that are either multimodal or dynamical (i.e., sequential). In this paper, we present a multimodal and dynamical VAE (MDVAE) applied to unsupervised audiovisual speech representation learning. The latent space is structured to dissociate the latent dynamical factors that are shared between the modalities from those that are specific to each modality. A static latent variable is also introduced to encode the information that is constant over time within an audiovisual speech sequence. The model is trained in an unsupervised manner on an audiovisual emotional speech dataset, in two stages. In the first stage, a vector quantized VAE (VQ-VAE) is learned independently for each modality, without temporal modeling. The second stage consists in learning the MDVAE model

on the intermediate representation of the VQ-VAEs before quantization. The disentanglement between static versus dynamical and modality-specific versus modality-common information occurs during this second training stage. Extensive experiments are conducted to investigate how audiovisual speech latent factors are encoded in the latent space of MDVAE. These experiments include manipulating audiovisual speech, audiovisual facial image denoising, and audiovisual speech emotion recognition. The results show that MDVAE effectively combines the audio and visual information in its latent space. They also show that the learned static representation of audiovisual speech can be used for emotion recognition with few labeled data, and with better accuracy compared with unimodal baselines and a state-of-the-art supervised model based on an audiovisual transformer architecture.

8.1.2 Unsupervised performance analysis of 3D face alignment with a statistically robust confidence test

Participants: Xavier Alameda-Pineda, Radu Horaud.

This paper addresses the problem of analyzing the performance of 3D face alignment (3DFA), or facial landmark localization. This task is usually supervised, based on annotated datasets. Nevertheless, in the particular case of 3DFA, the annotation process is rarely error-free, which strongly biases the results. Alternatively, unsupervised performance analysis (UPA) is investigated. The core ingredient of the proposed methodology is the robust estimation of the rigid transformation between predicted landmarks and model landmarks. It is shown that the rigid mapping thus computed is affected neither by non-rigid facial deformations, due to variabilities in expression and in identity, nor by landmark localization errors, due to various perturbations. The guiding idea is to apply the estimated rotation, translation and scale to a set of predicted landmarks in order to map them onto a mathematical home for the shape embedded in these landmarks (including possible errors). UPA proceeds as follows: (i) 3D landmarks are extracted from a 2D face using the 3DFA method under investigation; (ii) these landmarks are rigidly mapped onto a canonical (frontal) pose, and (iii) a statistically-robust confidence score is computed for each landmark. This allows to assess whether the mapped landmarks lie inside (inliers) or outside (outliers) a confidence volume. An experimental evaluation protocol, that uses publicly available datasets and several 3DFA software packages associated with published articles, is described in detail. The results show that the proposed analysis is consistent with supervised metrics and that it can be used to measure the accuracy of both predicted landmarks and of automatically annotated 3DFA datasets, to detect errors and to eliminate them. Source code and supplemental materials for this paper are publicly available at https://team.inria.fr/robotlearn/upa3dfa/.

8.1.3 MEGA: Masked Generative Autoencoder for Human Mesh Recovery

Participants: Xavier Alameda Pineda.

Human Mesh Recovery (HMR) from a single RGB image is a highly ambiguous problem, as similar 2D projections can correspond to multiple 3D interpretations. Nevertheless, most HMR methods overlook this ambiguity and make a single prediction without accounting for the associated uncertainty. A few approaches generate a distribution of human meshes, enabling the sampling of multiple predictions; however, none of them is competitive with the latest single-output model when making a single prediction. This work proposes a new approach based on masked generative modeling. By tokenizing the human pose and shape, we formulate the HMR task as generating a sequence of discrete tokens conditioned on an input image. We introduce MEGA, a MaskEd Generative Autoencoder trained to recover human meshes from images and partial human mesh token sequences. Given an image, our flexible generation scheme allows us to predict a single human mesh in deterministic mode or to generate multiple human meshes in stochastic mode. MEGA enables us to propose multiple outputs and to evaluate the uncertainty of the predictions. Experiments on in-the-wild benchmarks show that MEGA achieves state-of-the-art

performance in deterministic and stochastic modes, outperforming single-output and multi-output approaches.

8.2 Human Behavior Understanding

8.2.1 Autoregressive GAN for Semantic Unconditional Head Motion Generation

Participants: Louis Airale, Stéphane Lathuilière, Xavier Alameda Pineda.

Over the past years, semantic segmentation, as many other tasks in computer vision, benefited from the progress in deep neural networks, resulting in significantly improved performance. However, deep architectures trained with gradient-based techniques suffer from catastrophic forgetting, which is the tendency to forget previously learned knowledge while learning new tasks. Aiming at devising strategies to counteract this effect, incremental learning approaches have gained popularity over the past years. However, the first incremental learning methods for semantic segmentation appeared only recently. While effective, these approaches do not account for a crucial aspect in pixel-level dense prediction problems, i.e. the role of attention mechanisms. To fill this gap, in this paper we introduce a novel attentive feature distillation approach to mitigate catastrophic forgetting while accounting for semantic spatial- and channel-level dependencies. Furthermore, we propose a continual attentive fusion structure, which takes advantage of the attention learned from the new and the old tasks while learning features for the new task. Finally, we also introduce a novel strategy to account for the background class in the distillation loss, thus preventing biased predictions. We demonstrate the effectiveness of our approach with an extensive evaluation on Pascal-VOC 2012 and ADE20K, setting a new state of the art.

8.2.2 Semi-supervised learning made simple with self-supervised clustering

Participants: Pietro Astolfi, Elisa Ricci, Xavier Alameda-Pineda.

Self-supervised models have been shown to produce comparable or better visual representations than their supervised counterparts when trained offline on unlabeled data at scale. However, their efficacy is catastrophically reduced in a Continual Learning (CL) scenario where data is presented to the model sequentially. In this paper, we show that self-supervised loss functions can be seamlessly converted into distillation mechanisms for CL by adding a predictor network that maps the current state of the representations to their past state. This enables us to devise a framework for Continual self-supervised visual representation Learning that (i) significantly improves the quality of the learned representations, (ii) is compatible with several state-of-the-art self-supervised objectives, and (iii) needs little to no hyperparameter tuning. We demonstrate the effectiveness of our approach empirically by training six popular self-supervised models in various CL settings.

8.2.3 Motion-DVAE: Unsupervised learning for fast human motion denoising

Participants: Xavier Alameda-Pineda.

In this work, we address the task of unconditional head motion generation to animate still human faces in a low-dimensional semantic space from a single reference pose. Different from traditional audio-conditioned talking head generation that seldom puts emphasis on realistic head motions, we devise a GAN-based architecture that learns to synthesize rich head motion sequences over long duration while maintaining low error-accumulation levels. In particular, the autoregressive generation of incremental outputs ensures smooth trajectories, while a multi-scale discriminator on input pairs drives generation toward better handling of high- and low-frequency signals and less mode collapse. We experimentally

demonstrate the relevance of the proposed method and show its superiority compared to models that attained state-of-the-art performances on similar tasks.

8.2.4 Robust Audio-Visual Contrastive Learning for Proposal-based Self-supervised Sound Source Localization in Videos

Participants: Xavier Alameda-Pineda.

By observing a scene and listening to corresponding audio cues, humans can easily recognize where the sound is. To achieve such cross-modal perception on machines, existing methods take advantage of the maps obtained by interpolation operations to localize the sound source. As semantic object-level localization is more attractive for prospective practical applications, we argue that these map-based methods only offer a coarse-grained and indirect description of the sound source. Additionally, these methods utilize a single audio-visual tuple at a time during selfsupervised learning, causing the model to lose the crucial chance to reason about the data distribution of large-scale audio-visual samples. Although the introduction of Audio-Visual Contrastive Learning (AVCL) can effectively alleviate this issue, the contrastive set constructed by randomly sampling is based on the assumption that the audio and visual segments from all other videos are not semantically related. Since the resulting contrastive set contains a large number of faulty negatives, we believe that this assumption is rough. In this paper, we advocate a novel proposal-based solution that directly localizes the semantic object-level sound source, without any manual annotations. The Global Response Map (GRM) is incorporated as an unsupervised spatial constraint to filter those instances corresponding to a large number of sound-unrelated regions. As a result, our proposal-based Sound Source Localization (SSL) can be cast into a simpler Multiple Instance Learning (MIL) problem. To overcome the limitation of random sampling in AVCL, we propose a novel Active Contrastive Set Mining (ACSM) to mine the contrastive sets with informative and diverse negatives for robust AVCL. Our approaches achieve state-of-the-art (SOTA) performance when compared to several baselines on multiple SSL datasets with diverse scenarios.

8.2.5 A weighted-variance variational autoencoder model for speech enhancement

Participants: Xavier Alameda-Pineda.

We address speech enhancement based on variational autoencoders, which involves learning a speech prior distribution in the timefrequency (TF) domain. A zero-mean complex-valued Gaussian distribution is usually assumed for the generative model, where the speech information is encoded in the variance as a function of a latent variable. In contrast to this commonly used approach, we propose a weighted variance generative model, where the contribution of each spectrogram time-frame in parameter learning is weighted. We impose a Gamma prior distribution on the weights, which would effectively lead to a Student's t-distribution instead of Gaussian for speech generative modeling. We develop efficient training and speech enhancement algorithms based on the proposed generative model. Our experimental results on spectrogram auto-encoding and speech enhancement demonstrate the effectiveness and robustness of the proposed approach compared to the standard unweighted variance model.

8.2.6 Lost and found: Overcoming detector failures in online multi-object tracking

Participants: Xavier Alameda-Pineda.

Multi-object tracking (MOT) endeavors to precisely estimate the positions and identities of multiple objects over time. The prevailing approach, tracking-by-detection (TbD), first detects objects and then

links detections, resulting in a simple yet effective method. However, contemporary detectors may occasionally miss some objects in certain frames, causing trackers to cease tracking prematurely. To tackle this issue, we propose BUSCA, meaning 'to search', a versatile framework compatible with any online TbD system, enhancing its ability to persistently track those objects missed by the detector, primarily due to occlusions. Remarkably, this is accomplished without modifying past tracking results or accessing future frames, i.e., in a fully online manner. BUSCA generates proposals based on neighboring tracks, motion, and learned tokens. Utilizing a decision Transformer that integrates multimodal visual and spatiotemporal information, it addresses the object-proposal association as a multi-choice question-answering task. BUSCA is trained independently of the underlying tracker, solely on synthetic data, without requiring fine-tuning. Through BUSCA, we showcase consistent performance enhancements across five different trackers and establish a new state-of-the-art baseline across three different benchmarks. Code available at: https://github.com/lorenzovaquero/BUSCA.

8.3 Learning and Control for Social Robots

8.3.1 Navigating the Practical Pitfalls of Reinforcement Learning for Social Robot Navigation

Participants: Xavier Alameda-Pineda.

Navigation is one of the essential tasks in order forrobots to be deployed in environments shared with humans.The problem becomes increasingly complex when taking inconsideration that the robot's behaviour should be suitable tohumans. This is referred to as social navigation and it is acognitive task that us humans pay little attention to as it comesnaturally. Since crafting a model of the environment dynamicsthat faithfully characterises how humans navigate seems animpossible task, we look on the side of learning-based approachesand especially reinforcement learning. In this paper we areinterested in drawing conclusions on the vast number of designchoices when training a navigation agent using reinforcementlearning. To make this educated decisions, we offer a short surveyon recent papers addressing the social navigation problem usinglearning-based algorithms. Additionally, we take note of what worked best in our testing.

8.3.2 Socially Pertinent Robots in Gerontological Healthcare

Participants: Soraya Arias, Nicolas Turro, Alex Auternaud, Chris Reinke, Victor Sanchez, Xavier Alameda-Pineda.

Despite the many recent achievements in developing and deploying social robotics, there are still many underexplored environments and applications for which systematic evaluation of such systems by end-users is necessary. While several robotic platforms have been used in gerontological healthcare, the question of whether or not a social interactive robot with multi-modal conversational capabilities will be useful and accepted in real-life facilities is yet to be answered. This paper is an attempt to partially answer this question, via two waves of experiments with patients and companions in a day-care gerontological facility in Paris with a full-sized humanoid robot endowed with social and conversational interaction capabilities. The software architecture, developed during the H2020 SPRING project, together with the experimental protocol, allowed us to evaluate the acceptability (AES) and usability (SUS) with more than 60 end-users. Overall, the users are receptive to this technology, especially when the robot perception and action skills are robust to environmental clutter and flexible to handle a plethora of different interactions.

9 Partnerships and cooperations

9.1 International initiatives

Nothing to report.

9.2 European initiatives

9.2.1 H2020 projects

H2020 SPRING

Participants: Xavier Alameda-Pineda, Chris Reinke, Radu Horaud, Alex Auternaud, Victor Sanchez, Nicolas Turro, Gaetan Lepage, Soraya Arias.

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.

9.3 National initiatives

9.3.1 ANR JCJC Project ML3RI

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

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

9.3.2 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: https://project.inria.fr/avbot/

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Area Chair : Xavier Alameda-Pineda was Senior Area Chair at ACM Multimedia 2024, as well as Area Chair at ICASSP 2025.

General Chair Appointment : Xavier Alameda-Pineda was appointed General co-Chair of ACM Multimedia 2026.

Member of the organizing committees : Xavier Alameda-Pineda was co-Chair of the Award Committee at IEEE ICME 2024.

10.1.2 Scientific events: selection

Reviewer : The members of the team reviewer for top-tier international conferences such as IEEE CVPR 2024, ECCV 2024, IEEE ICASSP 2025, ACM MM 2024.

10.1.3 Journal

Member of the editorial boards : Xavier Alameda-Pineda was Associate Editor of IEEE TMM, ACM TOMM, ACM TIST, and CVIU.

Reviewer - reviewing activities : The members of the team reviewed for top-tier international journals such as IEEE TMM, IEEE TPAMI, IEEE TASLP, ACM TOMM, IJCV, CVIU, and TMLR.

Award Chair : Xavier Alameda-Pineda was co-Chair of the Award Committee for IEEE TMM 2023 Best Paper Award.

10.1.4 Invited talks

: Xavier Alameda-Pineda was invited to give two invited talks at CEA-LIST Days and at Valeo.ai, and was one of the Keynote Speakers at the French joint conference RFIAP/CAp 2024 in Lille.

10.1.5 Leadership within the scientific community

: Xavier Alameda-Pineda is the co-founder and Chair of the SIGMM European Chapter.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

In 2024, Xavier Alameda-Pineda was involved and responsible in teaching one course at Masters 2 level: Advanced Machine Learning, Applications to Vision, Audio and Text - at Master of Science in Informatics at Grenoble.

10.2.2 Supervision

Xavier Alameda-Pineda (co-)supervised Xiaoyu Lin (defended), Gaétan Lepage, Jordan Cosio, and Jean-Eudes Adylo.

10.2.3 Juries

Xavier Alameda-Pineda was a PhD reviewer for the Thesis of Hakim Benkirane, and was an examiner and appointed committee chair of the PhD of Ilyass Moummad.

10.3 Popularization

10.3.1 Productions (articles, videos, podcasts, serious games, ...)

: We have produced a video for the H2020 SPRING project: https://www.youtube.com/watch?v=JHDaQmK5HOg&t=10s&pp=ygUUc3ByaW5nIHByb2plY3QgaDIwMjA%3D.

11 Scientific production

11.1 Major publications

[1] L. Airale, D. Vaufreydaz and X. Alameda-Pineda. 'SocialInteractionGAN: Multi-person Interaction Sequence Generation'. In: *IEEE Transactions on Affective Computing* (11th May 2022). DOI: 10.1109/TAFFC.2022.3171719. URL: https://hal.inria.fr/hal-03163467.

[2] 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 (cit. on p. 4).

[3] X. Bie, S. Leglaive, X. Alameda-Pineda and L. Girin. 'Unsupervised Speech Enhancement using Dynamical Variational Autoencoders'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 30 (16th Sept. 2022), pp. 2993–3007. DOI: 10.1109/TASLP.2022.3207349. URL: https://hal.inria.fr/hal-03295630.

[4] 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.

[5] 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.

[6] Z. Kang, M. Sadeghi, R. Horaud and X. Alameda-Pineda. 'Expression-preserving face frontalization improves visually assisted speech processing'. In: *International Journal of Computer Vision* (12th Jan. 2023). DOI: 10.1007/s11263-022-01742-1. URL: https://hal.science/hal-03902610 (cit. on p. 3).

[7] 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.

[8] X. Lin, L. Girin and X. Alameda-Pineda. 'Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation'. In: *Transactions on Machine Learning Research Journal* (2024), pp. 1–19. URL: https://inria.hal.science/hal-03584014.

[9] 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 (cit. on p. 4).

[10] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda and R. Séguier. 'A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning'. In: *Neural Networks* 172 (Apr. 2024), p. 106120. DOI: 10.1016/j.neunet.2024.106120. URL: https://inria.hal.science/hal-04132316.

[11] L. Vaquero, Y. Xu, X. Alameda-Pineda, V. M. Brea and M. Mucientes. 'Lost and Found: Overcoming Detector Failures in Online Multi-Object Tracking'. In: ECCV 24 - 18th European Conference on Computer Vision. Milan (Italie), Italy, 14th July 2024, pp. 1–30. URL: https://inria.hal.science/hal-04650044.

[12] 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* 44.5 (1st May 2022), pp. 2673–2688. DOI: 10.1109/TPAMI.2020.3043781. URL: https://hal.inria.fr/hal-03328687.

[13] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus and X. Alameda-Pineda. 'TransCenter: Transformers With Dense Representations for Multiple-Object Tracking'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (28th Nov. 2022), pp. 1–16. DOI: 10.1109/TPAMI.2022.3225078. URL: https://hal.inria.fr/hal-03906940.

[14] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, M. Nabi, X. Alameda-Pineda and E. Ricci. 'Uncertainty-aware Contrastive Distillation for Incremental Semantic Segmentation'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (31st Mar. 2022), pp. 1–14. DOI: 10.1109/TPAMI.2022.3163806. URL: https://hal.inria.fr/hal-03908664.

[15] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, H. Tang, X. Alameda-Pineda and E. Ricci. 'Continual Attentive Fusion for Incremental Learning in Semantic Segmentation'. In: *IEEE Transactions on Multimedia* (14th Apr. 2022). DOI: 10.1109/TMM.2022.3167555. URL: https://hal.inria.fr/hal-03626393.

11.2 Publications of the year

International journals

[16] L. Airale, X. Alameda-Pineda, S. Lathuilière and D. Vaufreydaz. 'Autoregressive GAN for Semantic Unconditional Head Motion Generation'. In: *ACM Transactions on Multimedia Computing, Communications and Applications* (2024), pp. 1–11. DOI: 10.1145/3635154. URL: https://inria.hal.science/hal-03833759.

[17] X. Lin, L. Girin and X. Alameda-Pineda. 'Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation'. In: *Transactions on Machine Learning Research Journal* (2024), pp. 1–19. URL: https://inria.hal.science/hal-03584014 (cit. on p. 4).

[18] M. Sadeghi, X. Alameda-Pineda and R. Horaud. 'Unsupervised Performance Analysis of 3D Face Alignment with a Statistically Robust Confidence Test'. In: *Neurocomputing* 564 (Jan. 2024), pp. 1–16. DOI: 10.1016/j.neucom.2023.126941. URL: https://hal.science/hal-04265797.

[19] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda and R. Séguier. 'A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning'. In: *Neural Networks* 172 (Apr. 2024), p. 106120. DOI: 10.1016/j.neunet.2024.106120. URL: https://inria.hal.science/hal-04132316.

International peer-reviewed conferences

[20] G. Fiche, S. Leglaive, X. Alameda-Pineda, A. Agudo and F. Moreno-Noguer. 'VQ-HPS: Human Pose and Shape Estimation in a Vector-Quantized Latent Space'. In: *Proceedings of the 18th European Conference on Computer Vision ECCV 2024*. ECCV 2024 - 18th European Conference on Computer Vision. Milan, Italy, 31st May 2024, pp. 1–21. URL: https://hal.science/hal-04644818.

[21] A. Golmakani, M. Sadeghi, X. Alameda-Pineda and R. Serizel. 'A weighted-variance variational autoencoder model for speech enhancement'. In: ICASSP 2024 - International Conference on Acoustics Speech and Signal Processing. Seoul (Korea), South Korea, 2024, pp. 1–5. URL: https://inria.hal.science/hal-03833827.

[22] L. Vaquero, Y. Xu, X. Alameda-Pineda, V. M. Brea and M. Mucientes. 'Lost and Found: Overcoming Detector Failures in Online Multi-Object Tracking'. In: ECCV 24 - 18th European Conference on Computer Vision. Milan (Italie), Italy, 14th July 2024, pp. 1–30. URL: https://inria.hal.science/hal-04650044.

Conferences without proceedings

[23] X. Alameda-Pineda. 'Learning for Companion Robots: Preparation and Adaptation'. In: CAp - RFIAP 2024 - Joint Conférences sur l'Apprentissage automatique and Reconnaissance des Formes, Image, Apprentissage et Perception. Lille, France, 2024. URL: https://hal.science/hal-04634985.

[24] J.-E. Ayilo, M. Sadeghi, R. Serizel and X. Alameda-Pineda. 'Diffusion-based Unsupervised Audio-visual Speech Enhancement'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Hyderabad, India, 6th Apr. 2025. URL: https://hal.science/hal-04718254.

[25] D. Pikuli, J. Cosio, X. Alameda-Pineda, T. Fraichard and P.-B. Wieber. 'Navigating the Practical Pitfalls of Reinforcement Learning for Social Robot Navigation'. In: RSS 2024 - Workshop "Unsolved Problems in Social Robot Navigation. Delft / Netherlands, Netherlands, 19th July 2024, pp. 1–9. URL: https://inria.hal.science/hal-04639744.

Reports & preprints

[26] X. Alameda-Pineda, A. Addlesee, D. H. García, C. Reinke, S. Arias, F. Arrigoni, A. Auternaud, L. Blavette, C. Beyan, L. G. Camara, O. Cohen, A. Conti, S. Dacunha, C. Dondrup, Y. Ellinson, F. Ferro, S. Gannot, F. Gras, N. Gunson, R. Horaud, M. d'Incà, I. Kimouche, S. Lemaignan, O. Lemon, C. Liotard, L. Marchionni, M. Moradi, T. Pajdla, M. Pino, M. Polic, M. Py, A. Rado, B. Ren, E. Ricci, A.-S. Rigaud, P. Rota, M. Romeo, N. Sebe, W. Sieińska, P. Tandeitnik, F. Tonini, N. Turro, T. Wintz and Y. Yu. *Socially Pertinent Robots in Gerontological Healthcare*. Inria, 11th Apr. 2024, pp. 1–21. URL: https://inria.hal.science/hal-04737005.

11.3 Cited publications

[27] A. Ballou, X. Alameda-Pineda and C. Reinke. *Variational Meta Reinforcement Learning for Social Robotics*. 20th Dec. 2022. URL: https://hal.inria.fr/hal-03908505 (cit. on p. 4).

[28] C. Reinke and X. Alameda-Pineda. *Successor Feature Representations*. May 2022. URL: https://hal.inria.fr/hal-03426870 (cit. on p. 3).

[29] 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 (cit. on p. 6).

[30] 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 (cit. on p. 4).

[31] 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 (cit. on p. 4).

[32] 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 (cit. on p. 4).

[33] 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 (cit. on p. 6).

[34] 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 (cit. on p. 6).

[35] 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 (cit. on p. 6).

[36] 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 (cit. on p. 6).

[37] R. Gao and K. Grauman. 'Visualvoice: Audio-visual speech separation with cross-modal consistency'. In: *IEEE/CVF CVPR*. 2021 (cit. on p. 6).

[38] 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 (Dec. 2021), pp. 1–175. DOI: 10.1561/2200000089. URL: https://inria.hal.science/hal-02926215 (cit. on pp. 3, 4).

[39] 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 (cit. on p. 4).

[40] 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 (cit. on p. 6).

[41] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 7).

[42] 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 (cit. on p. 6).

[43] 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 (cit. on p. 4).