Gazeable Objects

Mention: This is a research project to be carried out at the Perception Team, at Inria Grenoble.
Contact: radu.horaud@inria.fr, xavier.alameda-pineda@inria.fr.

Topic: Gaze is the direction towards which a person is looking. The automatic estimation of the gaze from a single image and from videos has been a hot research topic in previous years [1-4]. Often, researchers studied gaze from a human-centered perspective, trying to answer the question "where are people looking" or "what are people looking at." In this Masters thesis we propose to investigate an orthogonal direction: we would like to understand the automatic recognition of *gazeable* objects. In practice, this would mean to estimate the areas of the image from which a particular object can be gazed. The combination of this research with state-of-the-art methods on "standard" gaze estimation should boost the performance on many real-world applications, like human-robot interaction or end-to-end active recognition.

Environment: This project will be carried out in the Perception Team [5], at Inria Grenoble Rhône-Alpes, and in collaboration with University of Granada [6]. The research progress will be closely supervised by Dr. Xavier Alameda-Pineda [7], Dr. Pablo Mesejo-Santiago [8] and Dr. Radu Horaud [9], head of the Perception Team. At the perception team we have the necessary computational resources (GPU & CPU) to carry on the proposed research.

References:
[1] Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking?. In NIPS.
[2] Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. In CVPR.
[3] Recasens, A., Vondrick, C., Khosla, A., & Torralba, A. (2017). Following gaze in video. In ICCV.
[4] Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., & Rehg, J. (2018). Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. In arXiv.
[5] https://team.inria.fr/perception/
[6] https://sci2s.ugr.es/
[7] https://xavirema.eu/
[8] https://sites.google.com/site/pablomesejo/
[9] https://team.inria.fr/perception/team-members/radu-patrice-horaud/

Speech enhancement with deep neural networks

Mention: This is a research project to be carried out at the Perception Team, at Inria Grenoble.
Contact: radu.horaud@inria.fr, xavier.alameda-pineda@inria.fr, simon.leglaive@inria.fr.

Topic: Speech enhancement is an important preprocessing step to various speech information retrieval tasks such as automatic speech recognition. The goal of a speech enhancement method is to provide a clean speech signal from a noisy recording that contains interfering audio sources (other people talking, ambient noise, etc.). The goal of this project is to develop algorithms based on deep neural networks (DNNs) for speech enhancement. A specific focus will be made on using generative neural networks such as variational autoencoders [1, 2, 3]. Two kind of approaches may be considered and compared:
1.“Fully-supervised” methods which assume the knowledge of the possible noise types (traffic noise, nature noise, etc.).
2.“Semi-supervised” methods which do not rely on this knowledge.
After getting familiar with the literature, the intern will work on developing new methods for speech enhancement based on deep neural networks.

Information for applicants: Please send your complete CV and a motivation letter to Simon Leglaive (simon.leglaive [at] inria.fr). Feel free to contact Simon Leglaive for any further information about the internship.

References:
[1] Diederik P. Kingma and Max Welling, “Auto-encoding variational Bayes,” International Conference on Learning Representations (ICLR), 2014.
[2] Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara, “Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization”, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2018.
[2] Simon Leglaive, Laurent Girin, and Radu Horaud, “A variance modeling framework based on variational autoencoders for speech enhancement”, IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2018.

Robust voice activity detection with deep neural networks

Mention: This is a research project to be carried out at the Perception Team, at Inria Grenoble.
Contact: radu.horaud@inria.fr, xavier.alameda-pineda@inria.fr, simon.leglaive@inria.fr.

Short description:
Voice activity detection (VAD) is a segmentation problem of a given audio signal into speech and non-speech sections. It constitutes an essential part in many modern speech-based systems such as those for speech and speaker recognition, speech enhancement, emotion recognition and human-computer or human-robot interaction. In many realistic situations, the recorded speech signal is contaminated by interfering noise coming from other audio sources. This noise can strongly deteriorate the performance of the VAD system.
The goal of this project is to develop robust algorithms for VAD based on deep neural networks (DNNs). Due to the sequential nature of the data, a natural choice would be to work with recurrent neural networks (RNNs) such as long short-term memory (LSTM) networks [1, 2].
After getting familiar with the literature, the intern will work on developing new methods for robust VAD based on deep neural networks.

Information for applicants: Please send your complete CV and a motivation letter to Simon Leglaive (simon.leglaive [at] inria.fr). Feel free to contact Simon Leglaive for any further information about the internship.

References:
[1] Thad Hughes, and Mierle Keir, “Recurrent neural networks for voice activity detection”, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2013.
[2] Simon Leglaive, Romain Hennequin, and Roland Badeau, “Singing voice detection with deep recurrent neural networks”, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2015.

Coupled Audio-visual Multi-speaker Tracking

Mention: This is a research project to be carried out at the Perception Team, at Inria Grenoble.
Contact: radu.horaud@inria.fr, xavier.alameda-pineda@inria.fr.

Topic: Multi-speaker tracking has been widely investigated and the Perception team contributed with a consistent methodological framework based on variational Bayes techniques [1-4]. Often, audio-visual tracking methods first map all auditory and visual information in the same space, to later on run a tracking algorithm. However, in most of the cases the auditory and visual observations are of very different nature, and they would require different latent space models, that are obviously linked to each other. In this Master thesis we would like to investigate if it is possible to perform multi-speaker tracking with coupled natural latent spaces, rather than one single artificial latent space.

Environment: This project will be carried out in the Perception Team [5], at Inria Grenoble Rhône-Alpes. The research progress will be closely supervised by Dr. Xavier Alameda-Pineda [6] and Dr. Radu Horaud [7], head of the Perception Team. At the perception team we have the necessary computational resources (GPU & CPU) to carry on the proposed research.

References:

[1] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers, 2018.
[2] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environment, 2018.
[3] Y. Ban, X. Li, X. Alameda-Pineda, L. Girin, and R. Horaud, “Accounting for Room Acoustics in Audio-Visual Multi-Speaker Tracking,” in IEEE ICASSP, 2018.
[4] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, “Exploiting the Complementarity of Audio-Visual Data for Probabilistic Multi-Speaker Tracking,” in IEEE ICCVW, 2017.
[5] https://team.inria.fr/perception/
[6] https://xavirema.eu/
[7] https://team.inria.fr/perception/team-members/radu-patrice-horaud/

Speaker identity modeling with deep learning for re-identification

Mention: This is a research project to be carried out at the Perception Team, at Inria Grenoble.
Contact: radu.horaud@inria.fr, xavier.alameda-pineda@inria.fr, simon.leglaive@inria.fr.

Short description: Speaker identification is the task that aims at determining which speaker has produced a given utterance [1]. On the other hand, speaker verification or re-identification aims at determining whether there is a match between a given speech utterance and a target speaker identity model [2]. Re-identification becomes difficult in situations where multiple speakers interact with each other [3,4]. In this project, we propose to explore the use of Siamese networks for learning a speaker identity model, which can be then used for a re-identification task in a multi-speaker scenario. The goal is to develop a system that can handle previously unseen speakers entering an on-going recorded conversation. After getting familiar with the literature, the intern will work on developing new methods for modeling speaker identities in the context of this re-identification task.

Keywords: speaker re-identification, multi-speaker tracking, deep neural networks, signal processing.

Information for applicants: Please send your complete CV and a motivation letter to Simon Leglaive (simon.leglaive [at] inria.fr) and Xavier Alameda-Pineda (xavier.alameda-pineda [at] inria.fr). Feel free to ask questions for any further information.

References:

[1] Yanick Lukic et al. "Speaker identification and clustering using convolutional neural networks." IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2016.
[2] Arsha Nagrani et al. "Voxceleb: a large-scale speaker identification dataset." Interspeech, 2017.
[3] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers, 2018.
[4] Y. Ban, X. Li, X. Alameda-Pineda, L. Girin, and R. Horaud, “Accounting for Room Acoustics in Audio-Visual Multi-Speaker Tracking,” in IEEE ICASSP, 2018.