PhD Proposal

INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex – France
http://www-sop.inria.fr/members/Francois.Bremond/

1. Title: Emotion Detection using Deep Learning

- Research axis of the 3IA: AI for Integrative Computational Medicine
- Supervisor: Francois Bremond - francois.bremond@inria.fr
- Potential co-supervisor: Antitza Dantcheva
- The laboratory and/or research group: STARS team at INRIA

Apply by sending an email directly to the supervisor.

The application should include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master's degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

Keywords : Emotion, Video Analysis, multimodal, fusion, bio-signal, Deep Learning

2. Scientific context

STARS group works on automatic sequence video interpretation. The "SUP" ("Scene Understanding Platform") Platform developed in STARS, detects mobile objects, tracks their trajectory and recognises related behaviours predefined by experts. This platform contains several techniques for the detection of people and for the recognition of human postures and gestures of one or more people using conventional cameras. However, there are scientific challenges in action detection when dealing with real word scenes populated with patients and doctors: cluttered scenes, handling wrong and incomplete person segmentation, handling static and dynamic occlusions, low contrasted objects, moving contextual objects (e.g. carts) ...

Existing work has either focused on simple activities in real-life scenarios, or the recognition of more complex (in terms of visual variabilities) activities in hand-clipped videos with well-defined temporal

boundaries. We still lack research on methods that can retrieve several instances of complex activity in a continuous video (untrimmed) flow of data. Existing methods that perform in online scenarios that can reason about the temporal and composite relations characterizing complex activities generally cannot handle uncertainty and tend to underperform in real life scenarios. Moreover, they have difficulties to distinguish similarly looking activities.
On the other hand, these methods are mostly dedicated to action detection and ignore the emotion component.

An emotion is a mental state that arises spontaneously and is often accompanied by cognitive, physical and physiological changes. Due to the complexity of human reactions, recognizing emotions is still limited to our knowledge and remains the target of many relevant scientific researches. In literature, the recognition of human behaviours, especially from facial expressions, often rely on the interpretation of dynamic scenes observed by video cameras. The accuracy of computer vision (CV) algorithms, as in the case of CNN, is typically limited by the identification of real emotion. A person may be happy even if she is not smiling and people differ widely in how expressive they are in showing their inner emotions. Recent multimodal sentiment analysis approaches focus on deep neural networks and propose multi-sensor data fusion methods. As emotions are complex set of reactions with multiple components [4], the idea is to compare/infuse/combine salient information from different modalities, coming from video cameras and biosensors. To lift the ambiguity, bio-signals (or Galvanic Skin Conductance (GSC) or electrodermal activity (EDA), ECG, EMG, Respiration Rate, etc.) with be used as ground truth (GT) for emotion.

3. General objectives of the PhD

This work consists in the improvement of Emotion Recognition/Detection algorithms using RGB video cameras at test time, but using multi-modalities at training time.
The objective is to develop and test a model on multiple datasets with various modalities to identify specific emotions, such as stress, anxiety, joy. The approach will consist of advanced Deep Learning methods for combining multimodal inputs, comparing various strategies such as multi-task learning, Knowledge Elicitation (infusion) using Student-Teacher paradigm, contrastive learning and co-training or Transformer. Several levels of ground truth (GT) supervision (e.g. weak-supervision) will be used to trained the model.

Typical pipeline can combine CNNs for 3D pose, eye-gaze and facial expression estimation, depending on the emotions to detect. Short temporal aspects of the actions can be handled through RNN or 3DCNN. The objective of this first step is to extract meaningful mid-level features that can be further processed thanks to more long-term reasoning based on TCN or Transformers or even ontology-based reasoning.
A challenge will be to propose an approach to leverage the knowledge acquisition process and the long-term reasoning with a weakly supervised setting.
This work aims at reducing the supervision in order to conceive a general and robust algorithm enabling the detection of the emotions of an individual (together with his/her facial expressions) living in an unconstrained environment and observed through a limited number of sensors (restricting to a single video camera).

To validate the work, we will assess the proposed approaches on videos from a set of applications in collaboration with Nice Hospital, such as the ones related to the monitoring of patients (e.g. autistic,

dementia, depressed) with behavioral disorders.

4. Pre-requisites:

Computer Vision, Strong background in C++/Python programming, Linux, Deep Neural Network frameworks (PyTorch, TensorFlow, Keras).

Knowledge on the following topics is a plus:
- Machine learning,
- Probabilistic Graphical Models and Optimization techniques,
- Mathematic (Geometry, Graph theory, Optimization),
- Artificial intelligence,
- Image processing and 3D Vision.

5. Schedule

1st year:
Study the limitations of existing emotion recognition/detection algorithms.
Depending on the targeted emotions, data collection might need to be carried out.
Propose an original algorithm that addresses current limitations on detection.
Evaluate the proposed algorithm on benchmarking datasets,
Write a paper

2nd year:
Investigation of feasibility/appropriateness of the framework in practical situations.
Propose an algorithm to address model learning task in weakly-supervised settings.
Write a paper

3rd year:
Optimize proposed algorithm for real-world scenarios.
Write a paper, and
PhD Manuscript

6. Bibliography:

- S. Das, S. Sharma, R. Dai, F. Bremond and M. Thonnat. VPN: Learning Video-Pose Embedding for Activities of Daily Living. In Proceedings of the 16th European Conference on Computer Vision, ECCV 2020, arXiv:2007.03056, online, UK, 23-28 August 2020.
- S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond and G. Francesca. Toyota Smarthome: Real-World Activities of Daily Living with supplementary. In Proceedings of the 17th International Conference on Computer Vision, ICCV 2019, in Seoul, Korea, October 27 to November 2, 2019.

- S. Das, F. Bremond and M. Thonnat. Looking deeper into Time for Activities of Daily Living Recognition. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass village, Colorado, March 2-5, 2020.
- Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV (2016).
- Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794−7803 CVPR (2018).
- Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018).
- Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In the IEEE International Conference on Computer Vision (ICCV) (Oct 2017).
- Zhang, S., Liu, X., Xiao, J.: On geometric features for skeleton-based action recognition using multilayer LSTM networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 148−157 (March 2017).
- Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: ECCV (2018) 59.
- Zhao, J., Snoek, C.G.M.: Dance with flow: Two-in-one stream action detection. In the IEEE Conference on Computer Vision and Pattern Recognition CVPR (June 2019)
- Shu, L. et al. A review of emotion recognition using physiological signals. Sensors 18, 2074 (2018).
- Chanthaphan, N., Uchimura, K., Satonaka, T. & Makioka, T. in 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). 117-124 (IEEE).
- Kahou, S. E. et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces 10, 99-111 (2016).
- Li, S. & Deng, W. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing (2020).
- Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019
- Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019
- Dai, Rui et al. "Self-Attention Temporal Convolutional Network for Long-Term Daily Living Activity Detection." 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (2019): 1-7.
- S.L. Happy, A. Dantcheva, A. Das, F. Bremond, R. Zeghari and P. Robert. Apathy Classification by Exploiting Task Relatedness. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Volume: 1, Pages: 733-738 DOI Bookmark:10.1109/FG47880.2020.00116, Buenos Aires, Argentina, 18-22 May, 2020.
- S.L. Happy, F. Bremond and A. Dantcheva. Semi-supervised Emotion Recognition Using Inconsistently Annotated Data. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Volume: 1, Pages: 477-484, DOI Bookmark:10.1109/FG47880.2020.00075, Buenos Aires, Argentina, 18-22 May, 2020.

7. Contact:

François Brémond <francois.bremond@inria.fr>
Sophia-Antipolis,
Research Director
Head of INRIA STARS Team
http://www-sop.inria.fr/members/Francois.Bremond/