Emotion recognition through next-generation multimodal fusion

Encadrant : Francois Bremond
Fonction : Professeur, INRIA
Laboratoire/Entreprise : INRIA
Adresse : Sophia
Equipe : STARS

Keywords : Emotion, Video Analysis, multimodal, fusion, biosignal, Deep Learning

Context

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 littérature, the recognition of human behaviours [1], 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 [2, 3]. 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, Galvanic Skin Conductance (GSC) or electrodermal activity (EDA) with be used as ground truth (GT).

Internship Objective

The objective of the internship is to develop and test a model on multiple datasets with various modalities to identify specific emotions, as stress, anxiety, joy. The student will be guided through the implementation 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. Several levels of ground truth (GT) supervision will be used to trained the model.

Typical pipeline can combine CNNs for 3D pose, eye-gaze and facial expression - estimation [5 – 10] 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 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, such as related to patients (e.g. autistic, dementia, depressed) and customers monitoring.

Prerequisites

Computer Vision, Strong background in C++/Python programming, Linux.

Knowledge on the following topics is a plus:

- Machine learning,
- Deep Neural Networks frameworks (PyTorch, TensorFlow, Keras),
- Probabilistic Graphical Models and Optimization techniques,
- Mathematic (Geometry, Graph theory, Optimization),
- Artificial intelligence,
- Image processing and 3D Vision.

References

1 Shu, L. *et al.* A review of emotion recognition using physiological signals. *Sensors* **18**, 2074 (2018).

2 Chanthaphan, N., Uchimura, K., Satonaka, T. & Makioka, T. in 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). 117-124 (IEEE).

3 Kahou, S. E. *et al.* Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* **10**, 99-111 (2016).

4 Li, S. & Deng, W. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020).

5 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

6 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

7 Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In The IEEE International Conference on Computer Vision (ICCV), October 2019

8 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.

9 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.

10 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.