Multi-modal multi-view emotion detection using non-negative matrix factorisation

Nistor Grozavu[1], **Yasser Khalafaoui** [1], **Nicoleta Rogovschi**[1]
[1] CY Cergy Paris University, France,
nistor.grozavu@cyu.fr, mykhalafaoui@alteca.fr, nicoleta.rogovschi@parisdescarrtes.fr

Through this work we explore the unsupervised topological learning of multimodal data presenting a complex structure allowing to learn their representations. We are particularly interested in heterogeneous data whose representation may have been informed in different ways: expert representation which may be complex. Most classical machine learning and statistical inference systems dedicated to multimodal and/or complex data, whether they are based on random models, empirical measures or prototype-based models, rely on a strong hypothesis, consisting in supposing at least that the structure of the data generating process for the observed scene is fixed, though it can be supposed unknown. In an unsupervised context, some existed works on Ensemble and Collaborative machine learning approaches were proposed but are limited to the same data distribution, i.e. in a multi-view context.

Non-negative Matrix Factorization (NMF) is a data mining technique that splits data matrices by imposing restrictions on the elements' non-negativity into two matrices: one representing the data partitions and the other to represent the cluster prototypes of the data set. This method has attracted a lot of attention and is used in a wide range of applications, including text mining, clustering, language modeling, music transcription, and neuroscience (gene separation). The interpretation of the generated matrices is made simpler by the absence of negative values. In this work, we propose a study on multi-modal clustering algorithms and present a multi-modal multi-view non-negative matrix factorization, in which we analyze the collaboration of several local NMF models.

The validation of the proposed approach is done on fusion of several emotion detection models covering multiple modalities: visual, acoustic and textual based on a dual-layered attention architecture.

The obtained results will be also presented on a demonstration starting by downloading video from YouTube. From the video we extract the audio track and the transcription for processing using and finally, as far as the visual aspect is concerned, we use two approaches, the first being based on "MediaPipe" if the trained model requires input markers, otherwise we extract directly from the video the images containing a well-framed face and clear. After-that the proposed multi-modal Non-negative matrix factorisation method is used for emotional detection.