AVEC 2019: Cross-Cultural Affect Recognition

Dataset: large volume of un-segmented, non-prototypical and non-preselected data of wholly naturalistic behaviour. This is the kind of data that new generation of human-robot communication interfaces have to face in the real world.

Audio-visual recordings collected *“in the wild”*. Standard webcams, at home/work place. Knowledge of German/Hungarian culture leveraged to infer knowledge on Chinese culture.

Ranking on labels relies on CCC – most suitable as it is not biased by changes in scale and location and elegantly includes information on both precision and accuracy in a single statistical measure.

CCC can be easily exploited as a loss function for training neural networks.

# Current State of the Art in cross cultural emotion recognition

A common idiom in facial expression recognition is that emotional expressions have a large degree of universality across cultures.

Vision only or multi modal systems achieved higher cross culture accuracies than speech only approaches.

Techniques used in AVEC 2018:

* Model based on emotional salient detection to identify emotion markers invariant to socio cultural context.
* Data driven approach based on long short term memory recurrent neural networks LSTM-RNN

All entrants in AVEC 2018 observed a drop in system performance when testing on Hungarian data.

# Cross-cultural Emotion Database (SEWA)

Spontaneous behavior

Video chats annotated w.r.t emotional dimensions: *arousal, valence* and *liking* (sentiment). Annotation contours (traces) are combined into a single gold-standard using *evaluator weighted estimator*.

# Baseline Features

Audiovisual representations can be learnt from expert driven information extracted from raw signals or directly from the raw signals.

Also can be generated using adversarial networks or CNNs trained on out-of-domain data and for a different task e.g. audio representations extracted by a model trained for objects classification in images.

Traditional approach in affect sensing consists in summarising low level descriptors of audiovisual signals over time with a set of measures computed over a sliding analysis window. For audio channel descriptors include spectral, cepstral, prosodic and voice quality info. For video channel they are appearance geometric and FAUs.

## Bag-of-Words

Represents distribution of LLDs according to a dictionary learned from them.

<https://machinelearningmastery.com/gentle-introduction-bag-words-model/>

## Deep Representations

Deep spectrum features – audio baseline feature representation.

Inspired by deep representation learning paradigms common in image processing. Spectral images of speech instances are fed into pre-trained image recognition CNNs, resulting activations are extracted as feature vectors

## Cross Cultural Sub challenge

Results of AVEC 2019 baseline features confirmed the idiom that facial expressions of emotion have a large degree of universality across cultures compared to vocal expressions where the acoustic and prosodic dimensions already play a key role in communication e.g. tonal languages like Mandarin.