AVEC 2018: Bipolar Disorder and Cross Cultural Affect Recognition

# Ranking metrics for regression:

CCC is preferred over Pearson’s Correlation Coefficient as it is sensitive to bias and scaling and it permits discriminative training when used as a cost function.

# Facial Recognition idioms and previous results

A common idiom in facial expression recognition is that due in part to consistencies across all humans in terms of make-up of facial muscles [12], emotional expressions have a large degree of universality across cultures. It was hard to find works in affective computing recognition literature which supported this claim [17], until the AVEC 2019 paper lol.

It has been argued that results of facial expression perception studies can be easily biased by the manner in which the answers are elicited [65].

Training with cultures from similar language families has been shown to have better accuracy than dissimilar [28, 67].

# Generation of Emotion Labels

Emotion Recognition requires a large amount of labels of sufficient quality to train systems to learn mapping from input data and labels describing the emotion. Labels need careful attention in definition as they are subjective and highly variable as it depends on human judgement.

Humans seem to be more efficient at discriminating among options that assigning absolute values to subjective variables [50, 84]. The dominant approach in affect modelling relies on absolute values of dimensional attributes such as arousal or valence that are annotated time continuously over the recordings.

Dominant practice – summarise the annotations for each recording into single time series known as the gold-standard. This can easily be processed by any ML algorithm.

Issues during fusion of individual annotations: inconsistencies between values reported by annotators, delay is present between emotional event expressed in the data and the corresponding annotation value.

Method to process noisy time continuous labels reported by humans on dimensional attributes of humans – winning contribution of AVEC 2019 [71], maximise PCC between audio-visual features and the gold standard to estimate the delay used to compensate the **reaction time of annotators.** In AVEC 2015 [61] participants proposed to estimate an overall reaction time for each emotional dimension by maximising recognition performance while varying the delay in a grid search [35, 36].

# Expert knowledge

The traditional approach consists in summarising LLDs of speech and video over time with a set of statistical measures compute over a fixed duration sliding window. Features can be brute forced with large ensemble of LLDs that are all combined with a large set of statistical measures – ComParE acoustic feature set [70], or they can be reduced to smaller expert knowledge based info.