Introduction of Problem Context

According to **(Picard),** emotions are a mental state where one experiences pleasure or displeasure with high intensity. They are a very necessary semantic component in human interaction - often, without the context of the speakers emotion, the intentions behind utterances can be ambiguous **(Automatic Emotion Recognition)**. The field of Affective computing aims to understand this phenomenon so that interactions between humans and machines become more naturalistic. Automatic Emotion Recognition has made significant strides in the previous decade but there are still many areas of unexplored territory. Up until the Audio/Visual Emotion Challenge in 2019, there were few works in affective computing recognition literature that supported the common idea that emotions conveyed by facial expressions are mostly universal across cultures **(AVEC 2018, 2019).** There still exist, however, some barriers to universal emotion recognition. It has been found that training machines to recognise emotion from similar language families have shown more accurate results. **(AVEC 2018).**

The dataset we use in this project are audio-visual recordings that have been collected “in the wild”. This phrase simply refers to the idea that standard webcams have been used for recording in a natural setting (home/ work place). The data is not preselected and the behaviour exhibited by the subjects are wholly spontaneous and naturalistic. **(AVEC 2019).** Literature has demonstrated great success with “in the lab” data where the variables are controlled, but this does not account for the noise present in real life situations that “in the wild” data can mimic.

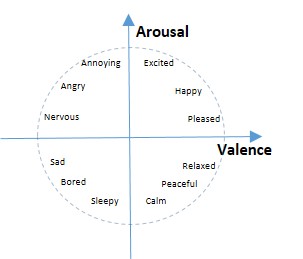
Multiple modalities are exploited in this project meaning that both audio and video data shall be used to extract information about the emotional cues demonstrated. Whilst vocal expressions of emotion have been showed to be less universal across cultures than facial expressions,multi modal systems have achieved relatively high cross cultural accuracies. **(AVEC 2019).**

**Emotional Representation Models**

There are two fundamental approaches when regarding emotion as a measurable concept:

1. Discrete Emotion theory
2. Dimensional models of emotion

Discrete emotion theory describes six basic emotions; anger, disgust, fear, happiness, sadness and surprise **(Ekman, Wikipedia emotion classification).**  Each emotion can be expressed with varying intensities within these discrete categories as opposed to an emotional state. **(Ekman**). There are many algorithms and techniques that realise discrete emotion theory as a classification problem e.g. Support Vector Machines, Naïve Bayes Classifiers and Neural Networks. The focus of this project is on the second approach – treating emotion as a point in a continuous space. This space is described by arousal, valence and dominance which are measures of affective activation, pleasure and control respectively **(Chen, multimodal multitask)**. This plane can be visualised by Thayer’s Arousal-Valence emotion plane **(R. E. Thayer, The Biopsychology of Mood and Arousal. New York: Oxford Univ. Press, 1989.)** Fig 1.

Emotion Recognition based on the emotion plane is a regression problem where observed variables and features are used to predict a real value describing the subject’s emotional state. This approach is free of the ambiguity that is present when treating emotion recognition as a classification problem **(Yang).** Regression is essentially an optimization problem where the value of parameters are set such that the “goodness” of a prediction is maximised in the resulting model. A “good” prediction could be a small prediction error or a large correlation coefficient.**(Continuous Emotion Recognition another look).**

**Machine Learning**

Automatic Emotion Recognition (AER) is typically tackled through machine learning models and algorithms. The process used by these techniques can be summarized as follows **(Acoustic feature selection)**:

* Feature Extraction Stage:

A feature is an input to a machine learning algorithm. Low Level Descriptor (LLD) features include spectral, prosodic and voice quality information for the audio channel and geometric details for the video channel e.g. face orientation and pixel coordinates for eye points and facial landmarks **(AVEC 2017)**. These features are compiled into feature vectors.

* Data Pre-processing Stage:

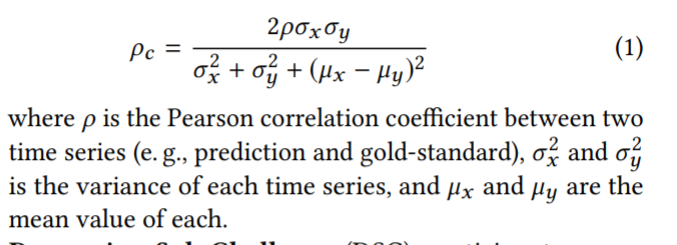
Either the most relevant subset of the entire feature set is selected or the number of dimensions in the dataset is reduced.

* Emotion Recognition Stage:

The emotional states of subjects in the recordings are determined by applying machine learning methods to the dataset. The output of these machine learning methods are compared to the labels of the dataset and the parameters of the model are tweaked if necessary. Usually the labels for each recording is merged into a single data series known as the gold standard, which can easily be handled by any machine learning algorithm. **(AVEC 2018)**

The output of the machine learning method is compared to the gold standard using a correlation coefficient. These are measures of how closely the gold standard and the predicted values are correlated. Values of -1.0, 0 and 1.0 indicate perfect negative correlation, no correlation and perfect positive correlation respectively.

Pearson’s correlation coefficient is defined as the covariance between two variables, x and y, divided by the product of the standard deviations of x and y. The concordance correlation coefficient is another measure that takes the Pearson coefficient of two time series and scaling it with their mean square distance. **(AVEC 2017)**

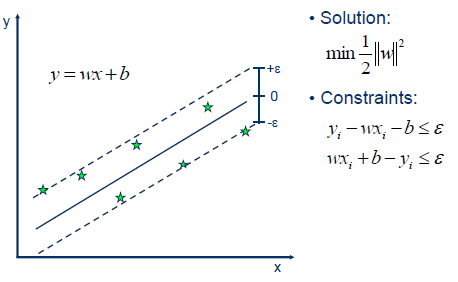


**Machine Learning Methods for Regression**

**Support Vector Regression Machines**

Support Vector Regression (SVR) is a supervised learning technique where the aim is to find a function that predicts real values that have a deviation from obtained values in the training data that is bounded by some \*epsilon\*, whilst the function itself is kept as flat as possible. **(Vapnik, 1995).** Since the output of a regression problem is a real number, it is very difficult to obtain a function that achieves perfect prediction, thus the need for a margin of tolerance. **(svrm)**

Own version of following chart:



**Artificial Neural Network**

Artificial Neural Networks (ANNs) are made up of some connected network of processing units and a learning algorithm. For every input connection of a unit, there is an associated weight value which determines the strength of the other connected nodes influence on itself. **(AI book, and 1991)**

Each node combines all the individual influences received on its input link with an activation function. These weights are adjusted to improve the output of the network. The most common method of adjusting these weights is back propagation where the values of the weights are moved in the direction of the greatest performance improvement by varying all the weights simultaneously in proportion to the observed improvement. (**AI book)**. There are two main connection topologies for an ANN - feedforward only architectures and recurrent architectures. (**1991)**

Feed forward networks only allow signals to travel in one direction so that the output of any one layer of units does not affect itself. Conversely, recurrent networks allow loops in the network, so that it simulates a form of “memory”. Feedforward networks assume that data in the training and testing set is independent. As a result, datasets where temporal information is important may not perform well on these networks – this is where recurrent networks are useful as they do not experience this problem. Since the dataset we use in this project contains video and audio data where the temporal information is important, we will be focusing on recurrent architectures.

**Recurrent Neural Networks**

Recurrent Neural Networks (RNNs) can selectively pass information across steps whilst sequential data is processed **(Lipton)**. There have been many successful machine learning models that do not explicitly model time. Instead they seek to implicitly capture time by using a sliding window to concatenate immediate predecessors and successors. This approach, however, does not provide a long term memory solution e.g. they cannot keep the complete context of a conversation for a long conversation in a call centre automation. (**Lipton)**

Training an RNN is a much more complex task than training a feedforward network and involves many more iterations. **(Picton)**. When attempting to train an RNN using methods that do not factor in the temporal aspect of the data, problems arise when the weight values are changed. This is due to the fact that the influence of a weight propagates throughout the time, rather than just in the current instant. Training an RNN by computing the weight correction for each iteration of a video dataset timeframe would very computationally expensive. The following are two examples of training methods that would work for an RNN. **(Dreyfus).**

**Back propagation through time (BPTT):**

The regular backpropagation algorithm is used with a cost function that shortens the computation time by using a sliding window.

**Real Time Recurrent Learning (RTRL):**

The gradient of previous states are approximated with respect to the current weights by the values of those gradients with respect to the previous weights.

**Long Short Term Memory**

When utilizing BPTT or RTRL algorithms on time-series with long term dependencies, errors that are propagated back in time through the network have the effect of either vanishing or exploding in size. **(Dieterrich)**. LSTM solves this by using specialized units called Constant Error Carousels (CECs) that enforce constant error flow through the network. CECs can store representations of recent input events as activations that do not decay over time, thus forming a form of long term memory. To prevent CECs from filling up with unnecessary information, access to it is “guarded” with input gates. These input gates learn to open and close at appropriate moments. Similarly, there also exist output gates which learn when to send out an output, and forget gates which resets the activation when the information stored is not useful. A CEC and its gates is collectively referred to as a memory cell. **(Hochreiter).**

**Attention based mechanisms**

When the human brain perceives a video, it does not give the same amount of attention to every pixel of the video, rather it selectively focuses on the important pixels. In the same way, ANNs can mimic this by adding an attention based mechanism that stochastically generates an output sequence to “focus” on given an input sequence. (**chorowski).** A temporal attention mechanism provides a larger weight to features from specific time frames. Attention mechanisms can also be adapted for multiple modalities. Modalities that are most helpful to determine the emotions at a particular instant can dynamically receive a stronger weight. Furthermore, the network can detect noise and other sources of uncertainty and down weight modalities affected by them to improve the certainty of decisions. (**Hori)**. This makes attention based mechanisms well suited for datasets that are captured “in the wild”.

**Conclusions**

SVRMs do not take into account the temporal relationships in the dataset and so is not as well suited for our project. They ignore crucial time-space relationships that ANNs can address with recurrent neural networks. **(Chen)** found that a temporal LSTM model significantly outperformed non temporal SVR models for predicting arousal and valence dimensions. The benefit of utilising attention based mechanisms is less clear; it has been shown that performance of such mechanisms degraded quickly when dealing with longer concatenated utterances for the audio modality **(Chorowski).** On the other hand, **(Hori)** found that combining multimodal attention models with standard temporal attention outperforms state-of-the-art methods, as the lower performance in the audio modality for longer utterances can be down weighted, as the attention is focused on the video modality instead.