The Effect of Temporal Supervision on the Prediction of Self-reported Emotion from Behavioural Features

T.M. Rietveld

The Effect of Temporal Supervision on the Prediction of Self-reported Emotion from Behavioural Features

by

T.M. Rietveld

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday August 24, 2020 at 15:30 PM.

Student number:	4472926	
Project duration:	December 1, 2019 – August 24, 2020	
Thesis committee:	Dr. H. Hung,	TU Delft, supervisor
	Dr. C. Oertel,	TU Delft
	Dr. K. Hildebrandt,	TU Delft
	Mr. A. Gudi, M.Sc.,	VicarVision
	Mr. B. Dudzik, M.Sc.,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

Continuous affective self-reports are intrusive and expensive to acquire, forcing researchers to use alternative labels for the construction of their predictive models. The most predominantly used labels in literature are continuous perceived affective labels obtained using external annotators. However an increasing body of research indicates that the relation between expressed emotion and experienced emotion might not be as apparent as previously assumed. Retrospective self-reports provided by participants do capture experienced emotion, but models applied on these labels suffer from the lack of continuous annotations during training. In this work, we aim to answer whether this lack of temporal information can be remedied by using continuous external annotations as proxies for experienced emotion over time. Furthermore, we investigate whether weakly-supervised models can generate accurate continuous annotations to reduce the annotation burden for large datasets. Our results indicate that external annotation sequences bear little significant information for the prediction of self-reports. However, forcing models to reflect changes in external annotations by training models in a multitask fashion improves model performance, suggesting that such temporal supervision helps models to distinguish relevant segments in input data. Besides this, we find that weakly-supervised models can to a certain extent capture changes over time, but in general yield poor results compared to fully-supervised models.

Preface

This document is the conclusion of my Master Thesis conducted at the Delft University of Technology in collaboration with VicarVision. During this 9 month process I've had the opportunity to work with very knowledgeable people without whom this thesis would not have been the same.
I would first and foremost like to thank my supervisors at the TU Delft, Dr. Hayley Hung and Bernd Dudzik. Your feedback, useful insights and expertise have helped me during this journey to shape my research into its final form. I would also like to express my gratitude towards Amogh Gudi, my supervisor at VicarVision. I appreciate how you were always available for brainstorming sessions and feedback whenever i needed it and have enjoyed both our on- and off-topic discussions. Another word of thanks goes out to all of the other colleagues at the SMR group; I've truly enjoyed my time with you all and you have made my internship a great experience.
A last and special word of thanks goes out to all my friends and family who were there for me during this process, both when I needed some motivation as well as when I needed to disconnect.

T.M. Rietveld
Delft, August 2020

The Effect of Temporal Supervision on the Prediction of Self-reported Emotion from Behavioural Features

T.M. Rietveld
tim.rietveld@box.nl
Delft University of Technology
Delft

ABSTRACT

Continuous affective self-reports are intrusive and expensive to acquire, forcing researchers to use alternative labels for the construction of their predictive models. The most predominantly used labels in literature are continuous perceived affective labels obtained using external annotators. However an increasing body of research indicates that the relation between expressed emotion and experienced emotion might not be as apparent as previously assumed. Retrospective self-reports provided by participants do capture experienced emotion, but models applied on these labels suffer from the lack of continuous annotations during training. In this work, we aim to answer whether this lack of temporal information can be remedied by using continuous external annotations as proxies for experienced emotion over time. Furthermore, we investigate whether weakly-supervised models can generate accurate continuous annotations to reduce the annotation burden for large datasets. Our results indicate that external annotation sequences bear little significant information for the prediction of self-reports. However, forcing models to reflect changes in external annotations by training models in a multitask fashion improves model performance, suggesting that such temporal supervision helps models to distinguish relevant segments in input data. Besides this, we find that weakly-supervised models can to a certain extent capture changes over time, but in general yield poor results compared to fully-supervised models.

1 INTRODUCTION

Affective state estimation has been a common goal for computer scientists and psychologists alike. Systems capable of accurately predicting which affective state its users are in are valuable in various domains, such as interactive multi-media applications, education and healthcare, but also in research and marketing. Obtaining the continuous ground truth for experienced emotion is however intrusive and disruptive, as it can only be acquired by continuously asking participants to report their experience through self-reports.

Having access to temporal labels is valuable to account for intra-video variations of emotion, as multiple emotions can be experienced during the same stimulus. Due to the high cost and intrusiveness of obtaining continuous self-reported annotations, current affect prediction models are often trained using external annotations. While still costly, these labels are obtainable for moderately sized datasets, as one annotators can label multiple response videos. However, by doing so models implicitly assume that displayed expressions equate to the experienced affective state, for example interpreting smiling as happiness. Although various works argue that coherence between bodily responses and experienced emotion exists[17, 22, 54, 65], displayed expressions are not equal to experienced emotion as emotion is not always expressed[15]. Furthermore, there might be no prototypical expressions for experienced emotion[26] (see Barrett et al. [3] for an extensive review).

If one wants to be sure that experienced emotion is captured, another alternative is to sacrifice the continuous annotations and use self-reports collected upon task completion. Doing so requires participants to recall their experience over the whole task and report a 'summarised' experience label. Although less informative than continuous experience labels, these retrospective self-reports are also of large interest, as they capture the overall emotion a certain stimulus induced. Models predicting retrospective self-reports are valuable in domains such as personalized multimedia recommendation, marketing and empathic agent design.
A major issue that these models face is the lack of temporal information available during training, as only one self-report per video is present. On top of this, these summarised self-reports are subject to a steep loss of episodic information[85]; Due to biases in episodic and semantic memory, the peak and end effects of an experience will exert a disproportionate influence on retrospective self-reports respectively[46, 85].

Incorporating knowledge about external annotations during the prediction of retrospective self-reports would remedy the lack of temporal knowledge due to the unavailability of continuous labels. Because external annotations are assumed to be proxies for experienced emotion, using these as continuous labels could be beneficial during prediction. However, because of the complex relationship between expressions and experienced emotion it is unclear how emotion is exactly

expressed through bodily behaviour. Adding to this, as external annotations are subject to annotators' interpretation of bodily behaviour, it is unclear whether these annotations are good proxies for retrospective self-reports at all. This leads to one of the research questions we aim to answer in this work:

Research Question 1: How do visual behaviours relate to retrospective self-reports of experienced emotion? Does the utilisation of external annotations improve the predictive power?

Psychological works finding coherence between bodily response systems is a good indicator that expressions might carry significant information for the prediction of experienced emotion. However, the exact relationship between them is unclear. Various studies indicate that emotions might have no prototypical expression, and that various expressions are shared between emotions[3, 26]. External annotations, which are the annotators' interpretations of participants' displayed behaviour, might therefore not necessarily find similar coherence between the interpretations of expressions and experienced emotion. With the following sub-question we aim to gain a better understanding of how the relationship between external annotations and retrospective self-reports compares to the relationship between visual behaviours and these self-reports.

Research Question 1.1: Are external annotations significant predictors for retrospective self-reported emotional experiences? How do they compare to raw visual behaviours?

Because external annotations are interpretations of expressions and are thus perceiver-dependant, the attribution of emotion to these expressions might be off. However, changes in external annotations are likely to be caused by changes in the participants behaviour. This notion of change might be valuable for the prediction of retrospective self-reports as it could indicate variations of emotion within the stimulus. As models trained for the prediction of self-reports only have access to a single video-level label, adding these external annotations as temporal supervision provides the model with a sense of intra-video emotion variations that would otherwise be absent. This could help models distinguish relevant segments from background data, which could boost model performance. We pose the following subquestion:

Research Question 1.2: Do external annotations help to seperate emotional episodes from neutral segments?

In all previous questions we have assumed external annotations are available. However obtaining these temporal annotations is time-consuming and costly, hindering the creation of large in-the-wild datasets with continuous annotations. However, most affective state prediction models are trained on external annotations and do require continuous labels during training. Various other domains cope with similar annotation costs, such as action or object localisation, protein function prediction and text categorisation. Approaches in these fields have tried to overcome this by using weakly-supervised models[72, 106, 109, 113], achieving promising results. A major benefit of these models is that they are specifically designed to cope with label sparsity, noisy data or annotations. Utilising weakly-supervised models in the domain of affect estimation allows these models to be trained on a single video-level label, while being able to predict on frame or segment resolution. However, up to now few works have applied weakly supervised models for affective state estimation. We therefore aim to investigate the viability of these models for the prediction of retrospective self-reports, as well as the prediction of continuously perceived emotion.

Research Question 2 How well can weakly supervised models be deployed to predict continuous annotations from video-level annotations?

As mentioned above, weakly-supervised models are capable of predicting at a higher resolution than the label they were trained on. This property has the benefit that localisation can be done simultaneously with classification and that insights into relevant segments for classification can be generated. A prominent example of this is weakly supervised action localisation, where a model is trained on a video-level label, but is capable of indicating in which segments of the video this label is present. Having accurate models capable of performing localisation allows annotators to label significantly less, allowing for the creation of larger datasets. To evaluate the viability of transferring such weakly supervised models to the domain of affect estimation, we pose the following subquestions:

Research Question 2.1 How does retrospective self-report prediction performance compare to fully-supervised models?
Research Question 2.2 How well can weakly supervised models predict continuous external annotation labels from video-level labels?

Answering these questions helps us to gain insight in the value of obtaining external annotations for the prediction of retrospective self-reports, and whether they can be obtained in a less costly manner. The remainder of this work is structured as follows. Section 2 describes related literature and gives a background to emotion theory. Section 3

describes the different datasets that are utilized in this work. In Section 4, the proposed classification pipeline and selected models are introduced. After this, Section 5 will describe a set of experiments and their results to investigate the predictive value of external annotations for the prediction of retrospective self-reports. This is followed by an analysis of the predictiveness of behavioural patterns in Section 6. A modality ablation study is presented in Section 7. Section 8 contains the results of experiments aiming to address the viability of utilising weakly-supervised models in affective computing. Lastly, Section 9 discusses the joint outcomes of these experiments and their implications. Furthermore, possible directions for future work are proposed.

2 RELATED WORK

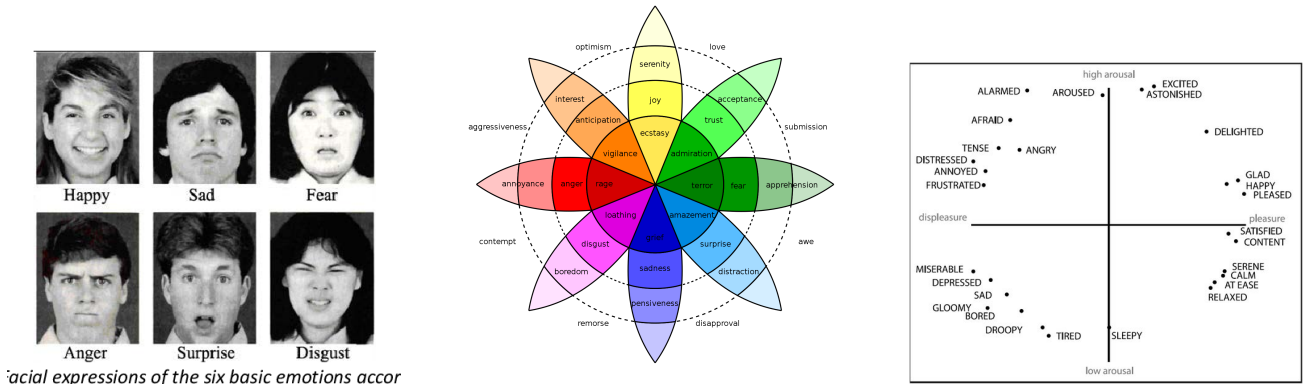
Emotion Representations

Before one can attempt to build emotion estimation models, a quantifying definition for emotion needs to be selected. Various representations for emotion exist in psychological literature, ranging from categorical to fully continuous. In the work of Darwin and Prodger [17] and Ekman et al. [24], emotions are seen as distinguishable evolutionarily evolved traits leading to a categorical representation for emotions. Ekman distinguishes six 'basic' emotions; emotions that can be distinguished from each other across cultures based on their corresponding facial expressions. Plutchik argues there are eight primary emotions, and introduces families of related emotions that evolve into each other based on the intensity of the emotion. This leads to a two-dimensional model, with one discrete dimension and one continuous dimension, called the Plutchik Wheel of Emotion[78]. A fully continuous representation is proposed by Russell and Mehrabian [88]. They propose a 3D model for representing emotions with Valence, Arousal and Dominance as its dimensions. In this model, Valence describes the pleasantness of the emotion, Arousal the energy of the emotion and Dominance describes how controlling and dominant an emotion is. Closely related is the 2D "core affect" model proposed by Russell [87]. Here the Dominance axis is omitted from the 3D model, hence emotions are modelled using two continuous dimensions: Arousal and Valence. The author suggests that different emotions can be placed in a circular spatial field in which emotions that are closely related lie close to each other. This continuous representation enables affective constructs to be defined as a combination of others, allowing for fuzzier definitions and to define more subtle variations in emotions. Ekman's basic emotions, Plutchik's wheel of emotion and the core affect model can be seen in Figure 1.

Communication and Attribution of Emotion

Psychologists have been trying to gain insight into how people experience and communicate emotion for years. One of the most prominent theories is the one proposed by Ekman et al. [24], arguing for the existence of six basic emotions that can uniquely be identified from the face. Based on his findings, Ekman and Friesen [23] developed the Facial Action Coding System (FACS), in which combinations of Action Units (AUs), activations of individual facial muscles, are mapped to emotion. Various other works argue for the importance of the face during the communication of emotion. In the work of Mehrabian [67], the author used verbal, vocal and facial cues to convey a message and investigated how they influenced the effect of the message. He found that when verbal and nonverbal cues communicated different affective states, facial cues contributed 55 percent to the interpreted effect of the message[67]. This dependency on the face during the conveying of emotion is also found in the works of Barrett et al. [3], Fölster et al. [28], Zhang et al. [112].

Besides the face, scientists have focused on the contribution of body posture and movement for the communication of emotion. Various works have in found evidence for the expressiveness of different body parts in static poses[16, 43]. Wallbott [104] shows that significant differences in body posture exists between 14 investigated emotion categories. In [18] the author researches the expressiveness of body movements for affective states. He found that particular body movements are expressive for specific emotional states, and that combinations of movements can predict emotion attribution. Shafir et al. [91] find significant effects between movement sequences and the emotion that participants reported after performing these movements. Furthermore, multiple works argue that face and body modalities might be connected during emotion attribution. Ambady and Rosenthal [2] found that human judgements of behaviours that were based on both face and body cues were 35% more accurate than those based solely on the face. Furthermore, works report that incongruency between modalities hinders correct attribution of acted out emotions[32, 60, 100]. As both the face and body appear to play important roles during the communication of emotions, we apply a combination of both modalities in our predictive models.



(a) categorical representation[23] **(b) Plutchik wheel of emotion[78]** **(c) Core Affect model[87]**

Figure 1: Different emotion representations

Given the vast amount of works linking emotion and facial and bodily responses, external annotations have been a popular choice for obtaining emotional labels to train predictive models. External annotators are tasked with attributing emotion by interpreting behavioural cues in response videos. However, over the years concerns have been voiced regarding the direct connection between experienced emotion and facial expressions. For example, in the work of Fernandez-Dols et al. [26], the authors find no coherence between participants' self-reports and the prototypical facial expressions corresponding to the reported emotion. Similarly, Ortony and Turner [74] argue that emotional experience is linked to individual muscular movement, rather than to full facial configurations. In the work of Hirt et al. [40], the authors directly compare the output of an emotion classification model and their relation to self-reports and find low coherence between the two.
Besides this, multiple works have shown that emotion attribution is a difficult task for humans. In the work of Elfenbein [25], the authors perform an extended meta-analysis over multiple studies investigating human emotion recognition capabilities, finding an overall attribution accuracy of 58% across cultures. Alongside this, works have found that annotators will overestimate the duration of emotional events compared to neutral ones[19, 21]. While training annotators is possible and results in fewer missatributions[64], doing so significantly increases annotations costs and hinders the annotation of large datasets.

Self-reports on the other hand are not subject to interpretation by others, as they are directly reported by participants themselves. However, these self-reports are difficult to obtain in a continuous fashion, as doing so would require participants to label their emotion during a stimulus which distracts them from this stimulus. For this reason, self-reports are often gathered before and upon task completion, asking participants to recall and summarise their experience over a stimulus. Multiple works have shown that emotional experiences are subject to a quick loss of information and that when multiple emotions are experienced within the stimulus, these mixed emotions become increasingly difficult to recall over time[1, 85]. Biases in episodic and semantic memory cause participants to base their self-reports on emotional episodes of high intensity and moments closer towards the end of the stimuli[46, 82, 90]. Besides the influence of memory biases, self-reported emotion might be susceptible to intentional or unintentional distortions by the participant. For example, participants might not report their true emotions due to social desirability biases[38], altering their reports to hide negative emotions such as shame and embarrassment[53, 89].

Summarising, ample research has shown that facial and bodily behaviour play an important role in the interpretation of emotions. However a growing body of research argues that the relation between expressed behaviours and experienced emotions appears to be more involved. For this reason we conduct a set of experiments to investigate how visual behaviours relate to retrospective self-reports. Furthermore, we study the relationship between interpretations of these visual behaviours in the form of external annotations and self-reported emotion to investigate whether these interpretations serve as good proxies for experienced emotion.

Predictive models

Over the years ample approaches have been proposed for the classification of emotion from video. Most of these models are trained on externally annotated label sequences, and hence attempt to predict emotion labels that match with the annotator's attributed emotion to participant behaviours. Early versions of these models used the FACS from Ekman and Friesen [23] and based their predictions solely on detected facial landmarks[56, 101]. More recent approaches often apply deep learning techniques to deduce emotions directly from raw images or video of faces. Convolutional Neural Networks (CNNs) are a popular component of these models, as they can exploit local structures and are translation equivariant. For example, Burkert et al. [8] use a standard CNN approach to learn visual features for emotion recognition from videos, which can then be passed to a traditional classifier. Gudi et al. [34] apply a similar approach to predict AU occurrence and intensity. This move towards deep models can also be observed for models that utilise the body modality. Initial models often used skin-color tracking algorithms to extract the locations of body parts of interest, and used these coordinates for their predictive models[11, 30]. Newer models apply deep models to extract keypoint locations instead[27, 79, 80], or directly attempt to learn relevant body parts or movements from data in an end-to-end fashion, often applying similar CNN-based models[4, 105, 108]. Learning from raw images or videos does however require a large amount of data during training, which might not be available. For this reason various approaches have attempted to use the output of pretrained models and build a classification network on top of these outputs showing promising results[27, 77, 79]. Doing so allows for leveraging the generalising capabilities from such pretrained models, significantly reducing the required amount of training data. Taking inspiration from these approaches, we apply two CNN-based pretrained networks to represent facial and bodily information.

With the shift towards end-to-end models, fusing modalities has become more common as training on modalities can be done jointly, incorporating information between modalities during optimisation. Various approaches have combined modalities for affect estimation and report improved model performance of these models[4, 27, 36, 98]. Besides exploring the effect of fusing modalities, affective computing researchers have also attempted to incorporate temporal dimensions into their models, as the temporal dynamics of expressions can provide valuable information for emotion classification. Gunes et al. [35] apply a Bidirectional-Long short-term memory (LSTM) and compare model performance to a Support Vector Machine (SVM). They report superior model performance for the temporal model, and conclude that temporal dynamics are crucial for affect prediction. This is supported by various other works[4, 98], showing a large temporal dependency on onset, apex and offset frames of expressions. Due to the shown importance of temporal information, we include time-based models in our experiments.

Comparatively little work attempts to predict retrospective self-reported emotion instead of using externally annotated data. In the work of Liu et al. [59], the authors attempt to predict retrospective self-reported emotion from recognised facial expressions, they report significantly above random chance model performance for predicting self-reports from facial expressions. In a follow-up study, the authors also model for the relation between retrospective self-reports and the expected induced emotion of a stimulus video. They report that including the expected induced emotion only provides a minor improvement, which they attribute to these relationships varying per subject and emotion[108]. These works do however not train on complete response videos, but are trained on the onset, apex and offset frames of expressions. Other studies focus on the usage of physiological signals for the prediction of retrospective self-reports, instead of using facial or bodily behaviour, such as [50], [95] and [47]. To the best of our knowledge, the work by Li et al. [55] is the only approach attempting to predict continuous self-reports. They evaluate their model on an internally collected dataset, in which participants were asked to continuously annotate their Arousal and Valence during the exposure to movie scenes. Correlations between predictions based on Galvanic Skin Response (GSR) signals and these continuous self-reports was then investigated, revealing a low correlation (Pearson correlation = 0.26).
Above works either require temporal annotations in the form of onset, apex and offset annotations or require sensors often unavailable outside labaratory settings to estimate experienced emotion. As these problems hinder the creation of large in-the-wild datasets, in this work we test whether models can predict self-reported emotion from video without the need such temporal annotations.

Because of the cost of collecting continuous self-reports and continuous external annotations, models that do not require these labels during training but are capable of predicting them are an interesting option. However little research has been conducted that investigates the viability of applying such models in the domain of affective computing. Some approaches in the domain of implicit tagging exist, applying various unsupervised techniques on collected physiological signals to detect emotional highlights (i.e. [10, 69, 93]). However, due to their unsupervised nature these models are only capable of detecting possible emotional episodes, they do

not predict a corresponding emotional label.
Weakly-supervised models on the contrary are capable of performing detection and prediction simultaneously. This allows for the localisation of temporal or spatial segments as well as the classification of global labels using the same model. In Sikka et al. [92] the authors apply such a weakly-supervised Multiple Instance Learning model to classify and localise expressions of pain in videos. Similar models have been applied outside the domain of affective computing. Wu et al. [109] apply an Multiple Instance Learning (MIL) model to perform protein function classification. In Zhou et al. [113], the authors apply MIL models to various domains, such as image and text categorisation. More recently, weakly supervised deep models have been proposed and have primarily been applied to the domain of object and action localisation in videos. These models, just like previous weakly-supervised models, are capable of predicting both temporal and global labels simultaneously, but eliminate the need for feature extraction, as they can be trained in an end-to-end fashion[72, 94, 106]. As these weakly-supervised models have shown promising results in different domains, we propose to apply these models in the domain of emotion prediction. Inspired by the approach presented in [92], we test a time-invariant Multiple Instance Learning model alongside a recurrent weakly supervised deep model as applied by Wang et al. [106].

Concluding, many works have been proposed for the prediction of perceived emotion annotation sequences and show good results. This is in line with psychological literature indicating that visual behaviours play an important role during emotion attribution. However, little affective computing works have focused on predicting self-reported experience from bodily behaviours. To address this gap in literature we test various models to investigate whether models can capture relevant visual behaviours for the prediction of retrospective self-reported emotion. As these models suffer from the lack of temporal information during training, we test if external annotation sequences can serve as proxies for relevant emotional segments. Obtaining these annotation sequences through external annotators is however expensive and hinders the creation of large in-the-wild datasets. To help alleviate this problem, we conduct a set of experiments to test whether weakly-supervised models can accurately predict external annotations from a single global label.

3 DATASETS

In this work experiments are mainly conducted on the RECOLA dataset[84]. This dataset was selected as it contains response videos of spontaneous interactions from a collaborative task captured in a controlled environment, in combination with continuous external annotations and self-reports. However as this dataset is relatively small, experiments that do not require external annotations will be verified on the larger Mementos dataset[20]. Characteristics and differences of these datasets will be explained in the following subsections.

RECOLA

The RECOLA dataset contains spontaneous collaborative and affective interactions in French[84]. Participants were grouped in dyads, but placed in seperate rooms. Both were seated behind a desk and had electrodes placed to record physiological signals. Participants were first individually tasked with solving a survival task: rank a number of items according to their importance for survival. After participants created their individual ranking, they were asked to reach a consensus through a group discussion. This discussion is what is captured in the dataset.
The dataset contains audio, video, Electro-Cardiogram (ECG) and Electro-Dermal Activity (EDA) signals of a duration of 5 minutes. Participants were asked to report their mood before and upon task completion, as well as to report the mood of their teammate through Self-Assessment Manikins[6] with a 9 point scale. Furthermore, the corpus contains continuous Valence-Arousal labels, obtained through external annotators. Six Fresh speaking annotators separately annotated Arousal and Valence in a continuous fashion using an annotation slider with values ranging from -1 to +1. These annotations where normalised and synchronised between annotators and averaged to form the gold-standard external annotation. The resulting corpus contains the data of 27 participants, their self-reports and the continuous annotation sequences. In this work we only focus on the captured video data and corresponding labels, although we believe that incorporation of the other available modalities could improve model performance.

Mementos

The Mementos dataset contains videos of spontaneous responses to one minute clips of music videos[20]. Three hundred participants were shown a random 7 video sample from a 42 video subset of the DEAP corpus[49]. After each music video participants were asked to report their affective state through the AffectButton by Broekens and Brinkman [7], resulting in a Pleasure, Arousal and Dominance score all ranging from -1 to +1. Participants were also asked whether this stimuli elicited particular memories, and to report the affect associated with these memories. Data were collected using the Mechanical Turk platform, hence response videos are captured in an uncontrolled environment, leading to varying lighting, pose and video quality conditions. In total, 2098 videos and their corresponding self-reports were successfully collected. No external annotations are available for this corpus. From these 2098 videos, we filter out all videos with a length of 65 seconds or more. The motivation for this

is that the video recordings should only capture responses to the stimuli, which have a duration of a minute. Therefore, recordings which are significantly longer than a minute are the result of problems with the recording setup, which the authors attribute to participants' varying internet connections or used web browser. Applying this filter, we obtain 1988 unique videos which are used to train our models.

4 METHOD

This section will provide an overview of the flow from raw input videos to the final affective label that is predicted by the models. We will first provide a description about how response videos are represented as features for our predictive models. After this, the label preparation procedure is discussed. Lastly, we introduce the predictive models that are selected for our experiments. A visual representation of the classification pipeline can be seen in Figure 2.

Feature extraction

The first step in any classification approach is choosing a representation for the original data in terms of feature vectors that the model can learn from. As learning from raw pixel values requires large amounts of data, we use the outputs of pretrained models to represent our videos. This allows us to obtain meaningful descriptors from these models without the need for a large-scale dataset. Similar to [27, 77, 79], we build our classification network on top of these outputs. We apply two different pretrained models to represent our input videos. The first model is a domain specific facial expression classification model, which returns emotion categories as well as Action Unit activation. The second is a body pose model capable of returning coordinates for specific body parts. By fusing the outputs of these two models, we can capture both fine-grained facial movement as well as body orientation and movements. These separate models will be discussed in more detail in the following paragraphs.

Facial feature extraction. Given an input video, we apply the pretrained DeepFace model from VicarVision's FaceReader software package[33, 34] to represent facial behaviour. For each frame in the video, a cropped grayscale image of the face is extracted. An alignment transformation is applied to this image using the locations of the eyes to correct for rotations of the face. From this aligned image the model predicts probabilities for Ekman's six basic emotions and a neutral class. Furthermore, it predicts demographics such as gender and age, alongside Action Unit activations for 20 AUs. For frames where the eyes are occluded or out-of-frame, we simply return a array containing only masked values. We end up with an $N \times 32$ representation of facial behaviour, where N denotes the number of frames. An example of an output on a single frame can be seen in Figure 3a. This figure shows detected activated Action Units overlapped on the face, as well as a categorical emotion estimation.

Body feature extraction. In parallel to the extraction of body features, we apply a pretrained model to represent bodily behaviour. For this we use the pretrained OpenPose model[9] to obtain human pose from an input video. Each frame in the video is passed through the model, which returns the 2D location of 25 body parts. Since participants in our datasets are likely to be seated behind a desk, we decided to omit all body parts below the elbow, as these are occluded or out-of-frame most of the time. We are left with the location of 10 body parts (Ears, Eyes, Nose, Neck, Shoulders and Elbows) for each frame in the video. Whenever one of these selected body parts is occluded or out-of-frame, it is substituted with a masking value in a post-processing step. After passing a complete video through the OpenPose model, the representation of this video is an $N \times 20$ feature vector, where N represents the total number of frames per video. An example of the OpenPose model's output can be found in Figure 3b. In this figure, numbered points denote detected body keypoints, which are connected to form the upperbody skeleton.

Joining Features. From these two sets of extracted features, we form our final input vector for each frame by joining these two sets of features based on their respective frame number. Whenever either of the models report missing values, so occluded body parts in case of the OpenPose model, or at least one occluded eye in the FaceReader model, we replace these values with a masking value of -2. As all outputs of either models are strictly positive, this masking value will not interfere with any legitimate data points, and can thus safely be used for masking. The final result is an $N \times 52$ feature vector, representing both facial and bodily features.

Creating uniform sequence lengths. As videos in both datasets vary in their amount of frames, the feature embedding for each video will be of different sizes. However, most classification models require uniform input sizes between inputs. To achieve this, we apply two different approaches. The first approach is to compute statistical features over the time dimension of the feature embedding. As features are aggregated over the temporal dimension, the output dimension will be irrespective of the original video length. The resulting feature vector includes time-independent statistics, such as mean, max and standard deviation, but also contains features that do take the sequential order into account such as autocorrelation and partial autocorrelation for different

[1]Image retrieved and modified from VicarVision's FaceReader Demo video: https://www.youtube.com/watch?v=emqhpMNcoRk

[2]Image modified from [51].

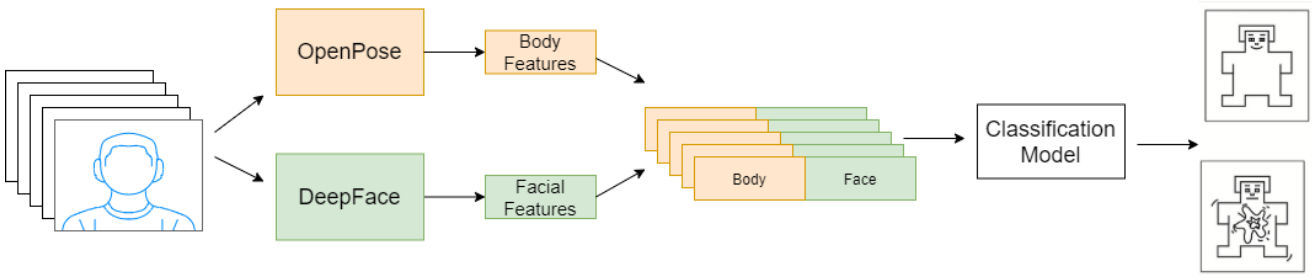
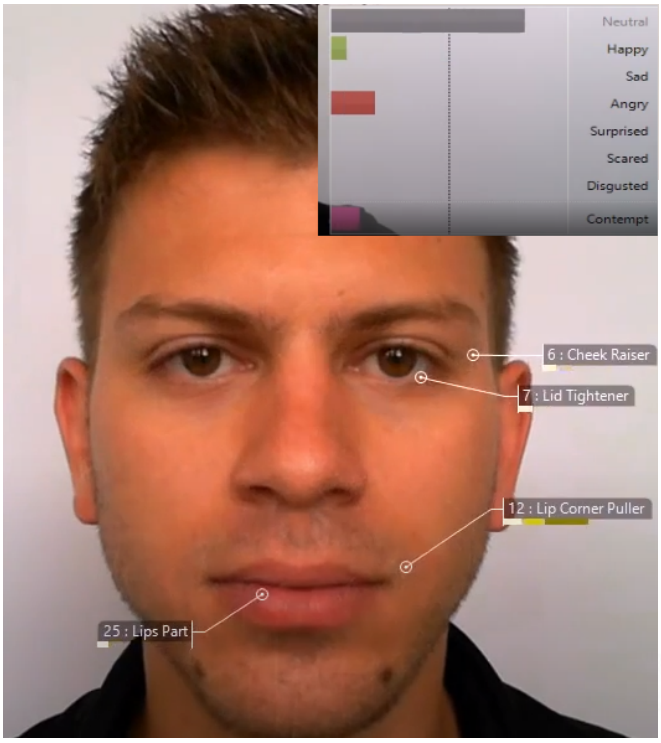
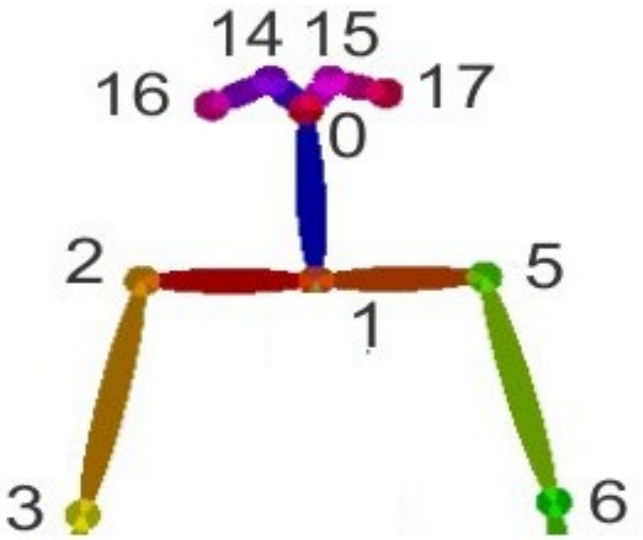


Figure 2: Proposed self-report classification pipeline utilising pretrained models to represent input videos.



(a) DeepFace[1]



(b) OpenPose[2]

Figure 3: Examples of the outputs of the applied pretrained DeepFace and OpenPose models

lags, Fourier analysis and seasonality[3]. This approach will be referred to as *Statistics* aggregation in the remainder of this work.

Our second approach uses a combination of resampling and padding to achieve similar sequence lengths. We first resample each video to a desired frame rate to account for varying frame rates between videos. After this we pad each video to the maximum number of frames occurring in the resampled dataset. This padding value is then masked within the models to account varying sequence sizes. This approach will be denoted by *Resample* aggregation in the following sections.

[3]For a complete feature overview, see https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html

Label preparation

Before classification can be performed using the extracted features described in the previous subsection, we first need to define what labels will be used during training. We discuss how we handle both the retrospective self-reports as well as the external annotation labels.

Binning of retrospective self-report values. Due to the low amount of datapoints available in the RECOLA dataset (n=18), attempting to classify the self-reports into the original 9 levels is infeasible due to the low amount of examples per class. Therefore, we bin the original labels for Arousal into 3 distinct bins, representing low arousal, medium arousal and high arousal respectively. The bin boundaries for these distinct bounds set at [1-3], (3,6] and (6,9] respectively. Following from this binning approach, resulting bins contain 7,7,4 samples respectively. We apply a similar binning strategy for the Mementos dataset. However, since values for self-reports in this dataset range from -1 to +1, we first linearly project these values to a 1 to 9 scale before bins are determined using the same bin boundaries as defined for RECOLA. Arousal and Valence distributions of the datasets before and after binning can be found in Figures 18 to 21 in Appendix C.

As becomes apparent in Figure 19, the Valence dimension in the RECOLA dataset is severely skewed, leading to little label variance. Applying our binning approach to the Valence distribution results in 0, 2 and 16 samples in the three respective bins. Choosing different bin boundaries would lead to difficulties during result interpretation, and could cause misinterpretations regarding the predictability of each of the Arousal-Valence dimensions. For these reasons, the prediction of valence has been omitted from further experiments.

Synchronising External annotations and behavioural features. Besides retrospective self-reports, external annotation sequences are available in the RECOLA dataset. These external annotations are collected at a 40ms rate, and authors report that this corresponds to the collected video frame rate of 25FPS. This frame rate is however not exact, leaving a small asynchrony between annotations and total amount of frames. To illustrate this, each annotation sequence has a length of 7501 (5 minute video reported at 40ms), whereas the length of the videos range from 7331 to 7501 frames. To account for this, we compute a timestamp per frame by assuming a stable frame rate throughout the video and label each frame with the annotation that is closest to its time stamp. A visual description of the synchronisation of video frames and annotations can be seen in Figure 4. As external annotations for the Mementos dataset are unavailable, this step is omitted.

We are left with a three category label per video, which represents self-reported Arousal, and in the case of RECOLA a $N \times 2$ label sequence representing the continuous external Arousal-Valence annotations. These labels will be used throughout the various classification models that are discussed in the following paragraphs.

Affect classification models

Now that both input and outputs have been obtained, we describe how we attempt to model the relationship between input and output. To do so, we apply several machine learning approaches, which will seperately be introduced in the following subsections. For a more thorough description and background on these models, we refer the reader to Appendix A.

Support Vector Classifier. SVMs have been a popular model choice in various domains, due to its robustness and ability to perform non-linear classification. A SVM model is a binary classification model that applies a predefined kernel to the input data to project this data into a hyper-dimensional space[5]. It then attempts to find a decision boundary that maximises the distance (or so-called margin) between the class instances and the decision boundary.
In our work we apply the SVM model to predict self-reported emotion. As the SVM is a binary classification model, the model we apply to our 3-class classification problem is essentially an ensemble of SVMs. Each of the models in the ensemble is trained in a one-versus-all fashion, training three different models for the three categories respectively. For each video, the final label is obtained by assigning the label of the model with the highest positive class probability. Models were trained using the scikit-learn python package[76]. Each of our models utilise the following parameters during training, which have been obtained by applying a combination of cross-validation and gridsearch over the RECOLA dataset:

- RBF kernel
- RBF kernel coefficient $\gamma = 0.005$
- Regularisation parameter $C = 1e3$

Multi-layer Perceptron. Whereas SVMs uses predefined kernels to transform the model's input into high-dimensional feature space, MLPs are purely data-driven. First introduced by McCulloch and Pitts [66], these models work by multiplying the input with one or more weight matrices. Non-linear relations can be captured by applying so-called activation functions to the result of each consecutive multiplication. During training the model's output is compared with the ground truth. This difference is then used to compute a loss-value, from which weight gradients can be computed using backpropegation.
We apply two different MLP models in our experiments.

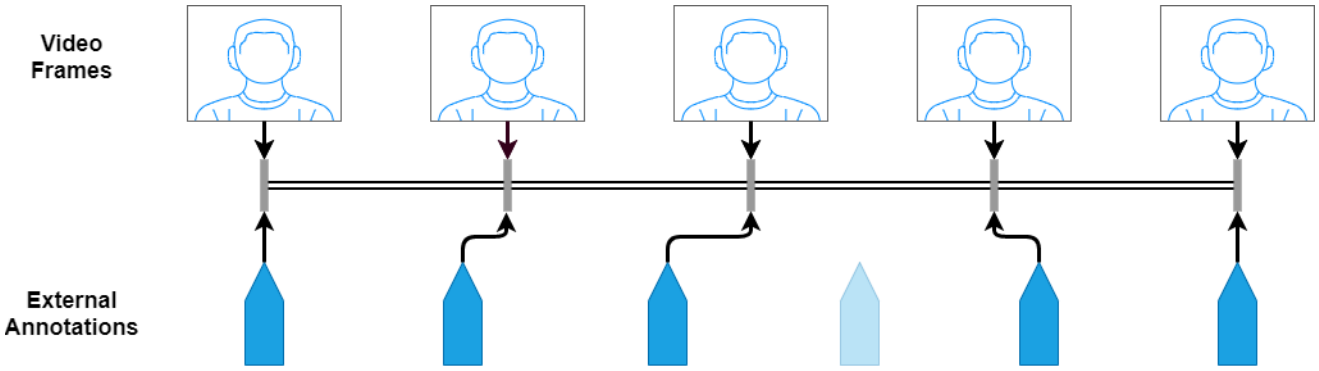
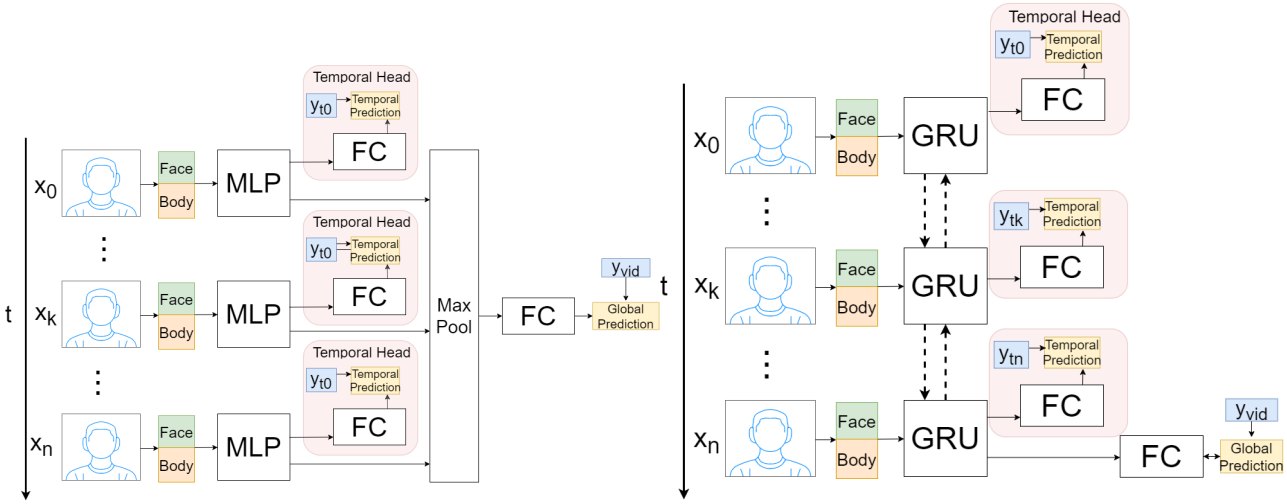


Figure 4: Synchronisation of video frames and external annotation sequences.



(a) Multilayer Perceptron (MLP) **(b) Gated Recurrent Unit (GRU)**

Figure 5: Model Architectures of the MLP and GRU models. The red boxes contain the temporal prediction heads and are omitted for models solely predicting self-reports. FC denotes a Fully Connected output layer

The first model is solely tasked with the prediction of self-reported arousal. The second model has an additional output head, allowing us to predict continuous external annotations simultaneously with global self-reported arousal. Differences between the two model architectures can be found in Figure 5a. The red boxes denote the temporal prediction head, which are omitted for models solely predicting self-reports.

Both models apply 2 hidden layers with relu activations[29] to each timestep of the input. The outputs of each timestep are combined using a maximum pool operation over the temporal dimension. This is fed into an output layer using a softmax activation to obtain the video-level label. To predict continuous annotations, we apply a separate output layer with a hyperbolic tangent activation function to the outputs of each timestep. We apply the hyperbolic tangent as output activation as our continuous target labels vary within the range of -1 and +1, which matches with the output domain of the hyperbolic tangent function. We apply Gaussian

noise and Dropout[97] as regularisation techniques to prevent models from overfitting.
Models were trained using the following training hyperparameters:

- Adam optimiser[48] with learning rate $\alpha = 0.01$
- A learning rate decay of 0.1 based on plateaus in the validation loss
- Input augmentation using Gaussian noise ($\sigma = 0.1$)
- Dropout between hidden layers (rate = 0.2)
- Video-level loss: Categorical cross-entropy
- Temporal loss: Mean Squared Error (Model 2 only)

Gated Recurrent Unit Network. Previously discussed models were not capable of exploiting the temporal dynamics that can be present in sequential data. As emotional expressions can usually be divided into a neutral, onset, apex and offset phase, the dynamics between these phases might be crucial for accurate predictions. Recurrent Neural Network models are capable of using the temporal dimension and explicitly model for temporal relationships over time. One of such Recurrent Neural Network (RNN) models is the Gated Recurrent Unit model introduced by Cho et al. [12]. In this model, sequential information is kept between timesteps in a hidden state that propagates information over time. For each timestep the model applies an update and forget gate to the combination of this hidden state and the input to compute the hidden state for the following timestep. Similar to the MLP model, the model is trained using backpropegation.
In our experiments we utilise two different versions of the GRU model, one only predicting video-level labels, the other predicting both video-level labels and continuous temporal labels. Figure 5b shows the difference between the model architectures, where the temporal prediction heads denoted by the red boxes are only included for the models predicting temporal annotations.

Both models utilise a Bidirectional GRU layer with a latent dimension of size 100. In the model predicting only video-level labels, we pass the hidden state of the GRU layer for the last timestep as input to an output layer with softmax activation. For the model generating temporal predicting, we pass the hidden state of the gru layer of each timestep to another output layer using a hyperbolic tangent activation function. This approach closely resembles the architecture of the MLP model with the key distinction being that temporal patterns between timesteps can be considered for learning a sequence representation. This is illustrated by the dotted lines in Figure 5b and the absense of such lines for the MLP model shown in Figure 5a. Similar to the MLP model described above, we apply Gaussian noise and Dropout to prevent overfitting on training data. Additional training parameters are reported below.

- Adam optimiser with learning rate $\alpha = 0.001$
- A learning rate decay of 0.1 based on plateaus in the validation loss
- Input augmentation using Gaussian noise ($\sigma = 0.1$)
- Recurrent dropout between timesteps (rate = 0.05)
- Video-level loss: Categorical cross-entropy
- Temporal loss: Mean Squared Error (Model 2 only)

MILBoost model. During the prediction of retrospective self-reports, all models are tasked with learning a relationship between the representation of a whole video and the corresponding affective label. However, since affective episodes in response videos are likely to be sparse, the label for the complete video might only depend on several small segments in the input. MIL models inherently embed this notion of redundancyinto their designs. They operate on so called bags, where each bag contains multiple instances. For our task, we define each video as a separate bag, and each frame of that video as a separate instance in the bag. The MILBoost model as introduced by Viola et al. [103] classifies a bag as positive if at least one of its instances is classified as positive. It predicts every instance using a boosted ensemble of Decision Trees. As this ensemble is only capable of binary classification, we train 3 different models for our 3-class classification problem in similar fashion to the SVM model. Again, labels are assigned based on the model with the highest label probability. However in contrast to previously described models, this model by design predicts a label for each instance, so tasks such as localisation can be performed in parallel to predicting global bag-level labels. In our experiments we use the same model for the prediction of self-reports as well as the prediction of continuous annotations. Each of the models in our ensemble apply are trained using the following parameters:

- Boosting ensemble of 50 iterations
- Decision Tree Boosting classifier (max depth =5)
- Generalized Mean softmax as in [92]

Weakly supervised Action Localisation model. A drawback of the MILBoost model is that it cannot handle sequential data, forcing its users to create sequential embeddings if they aim to utilise temporal information. Over the last years, weakly supervised deep models have gained popularity in the domain of action recognition in videos and object localisation in images. Similar to regular MIL models, these models also predict an output for each instance. However, temporal dynamics can be leveraged through the usage of recurrent models. The model we apply to our input data is an adaptation of the work of Wang et al. [106]. During training, this model randomly samples sequences of frames from the input video. These so called shots are fed to a bidirectional LSTM layer of size 256 to obtain individual feature embeddings

for each shot. This representation is passed to two separate output heads, one determining the temporal output value for each sequence, the other determining a relevance score of this sequence for the global label. Outputs of both heads are multiplied for each shot to form temporal predictions. Video-level predictions are obtained by summing these temporal predictions. The model's training parameters can be found below:

- Adam optimiser[48] with learning rate $\alpha = 0.01$
- Early stopping based on increases of the validation loss
- Recurrent dropout between timesteps (rate = 0.1)
- Dropout in output heads (rate=0.3)
- Video-level loss: Mean Squared Error

5 PREDICTING SELF-REPORT FROM EXTERNAL ANNOTATIONS

Traditionally, affect classification tasks attempt to predict a label for each time-step, where the labels of these time-steps are often collected from external annotators. These external annotations are formed by interpreting behavioural cues and expressions that participants display. However, if one is more interested in predicting experienced emotion, self-reports can be used as alternative form of annotation. However collecting continuous self-reports is intrusive and at best costly. Therefore, experiment designers often opt for asking for a single global self-report before and after task completion. This does however require participants to recall and summarise their experience during the task. As only a single self-reported label is available, models have no knowledge about variations of emotion within the stimuli that participants might experience, which is assumed to be captured in the external annotations. Therefore, we investigate to what extent external annotations can be applied for the prediction of self-report. If external annotations are good proxies for experienced emotion, then correlations between these external annotations and retrospective self-reports are likely to exist.

Experiment 1: Significance of external annotations on self-reports

To investigate whether perceived emotion sequences can directly be mapped to self-reported emotion, we perform a statistical analysis using self-reported emotion as the dependant variable. We describe our external annotations as statistical features similar to the ones described in Section 4. By doing so we avoid testing individual time-steps for significance, but rather test for significant properties of the annotation sequences. We obtain 763 interdependent features, for which we individually compute P values using Kendall's tau and Mann-Whitney U tests. To account for Type 1 errors (false significant features) introduced by performing a large number of independent tests, a Benjamini-Hochberg post-hoc procedure is applied to test each feature for statistical significance. This procedure assumes that falsely significant features occur with a predefined chance Q (set to 0.05 in our work) and tries to accounts for this. P values for each feature are ordered, determining a significance rank i per feature. This rank is multiplied by the expected false discovery rate Q and divided by the total number of tested features m to form the Benjamini-Hochberg critical value as shown in Equation 1.

$$BH_{crit}(i) = \frac{i * Q}{m} \tag{1}$$

Features are assumed to be significant if P values are smaller than the critical value, or if features with a lower rank comply with this criteria. Table 1a shows the most significant features for the prediction of Arousal, while Table 1b reports the most significant features for Valance.

Although we observe multiple statistical significant P values for both arousal and valence, we reject all of these features as being significant by means of the Benjamini-Hochberg procedure. This indicates that external annotation sequences bear little predictive value for retrospective self-reports. Although this result could be caused due to a too aggressive correction on significance values, this seems unlikely as decreasing the probability of false discovery Q by a factor 10 would still lead to the rejection of all features for valence, and all but one feature for arousal.

Experiment 2: Predicting self-reports from external annotation sequences

To further investigate the relation between self-reports and external annotation sequences, we perform a set of experiments to discover how well classification models can predict self-reports from external annotations. Applying non-linear models to the statistical representation as well as the raw annotation sequences allows us to test whether non-linear relationships might be missed during statistical analysis. Furthermore, comparing performances between models trained on the raw annotations sequences and model trained on the computed statistical feature representation enables us to evaluate whether relevant information was excluded from the statistical representation.

In this experiment we apply the MLP to both the statistical features and the raw annotation sequence. The GRU model is solely applied to the raw annotation sequence due to its time-based nature. Lastly, we apply the SVM only to the statistical features, to avoid rigidly fixing each timestep to a different dimension. Model performance was computed by averaging accuracy scores over 10 sessions of 3-fold cross-validation

Table 1: Top 5 significant external annotation features for self-reported Arousal and Valence. BH_{crit} denotes the Benjamini-Hochberg critical value to test P value against for significance. No significant features were found for both Arousal and Valence.

(a) Arousal

Feature	Specifics	P value	BH_{crit}
Fast Fourier Transform	coefficient 67	3.3e-4	6.6e-5
Fast Fourier Transform	coefficient 78	1.8e-3	1.3e-4
Fast Fourier Transform	coefficient 15	5.2e-3	2.0e-4
Fast Fourier Transform	coefficient 81	8.4e-3	2.6e-4
Fast Fourier Transform	coefficient 68	1.1e-2	3.3e-4

(b) Valence

Feature	Specifics	P value	BH_{crit}
$\sigma_{\|change\|}$	Quantiles 0.2-0.6	2.8e-3	6.6e-5
$\mu_{\|change\|}$	Quantiles 0.4-0.6	2.8e-3	1.3e-4
σ_{change}	Quantiles 0.2-0.6	2.8e-3	2.0e-4
$\mu_{\|change\|}$	Quantiles 0.2-0.6	3.6e-3	2.6e-4
Fast Fourier Transform	coefficient 65	3.6e-3	3.3e-4

Table 2: Model performance for the classification of retrospective self-reported Arousal from external annotations. Models fail to predict above random chance performance, indicating that external annotations contain little predictive value.

Model	Modality	Aggregation	ACC	STD
Majority	None	None	0.38	0.18
SVC	Ext. Annotation	Statistics	0.2	0.12
MLP	Ext. Annotation	Statistics	0.27	0.18
MLP	Ext. Annotation	Resample	0.3	0.17
GRU	Ext. Annotation	Resample	**0.33**	0.21

on the RECOLA dataset. The results of this experiment can be found in Table 2. In this table, the Aggregation column defines whether models were trained on statistical representations of the external annotation sequences as in the statistical test (denoted by Statistics), or were applied on raw unprocessed annotation sequences instead (denoted by Resample).

Similar to our first experiment, this experiment reveals that external annotations contain no significant predictive power for retrospective self-reports. Applying increasingly complex models did not boost performance. This result indicates that predictive power is not bound by model complexity or the model's ability to model for non-linear relations. Furthermore, a comparison between the MLP model trained on statistical features and a similar model trained on raw data shows that no substantial performance increase was found by using raw annotation sequences, suggesting that the statistical representation contains similar levels of information as the raw annotation sequences.

Both sets of experiments therefore find that external annotations are of low predictive value for the prediction of retrospective self-reports. Various explanations can be thought of for these findings. Firstly it could be that there exists low coherence between bodily behaviours and retrospective self-reports. This would cause external annotations, which are purely based on these visual behaviours, to have little to no predictive value. This explanation is however in contrast to various other works, where authors are able to capture summarised self-reported emotion labels from bodily responses[59, 96, 108].

Another possible explanation could be that annotators are insufficiently able to attribute the experienced emotion from expressions alone. Various works have researched how well humans are able to attribute emotion from visual cues, showing that humans across cultures are capable of recognising basic emotion at better than chance levels. Although above chance, recognition levels for basic emotions were estimated to be 58% percent (see [25] for an extensive meta-analysis). Similar results were found by Matsumoto et al. [64], who found a recognition rate of 47% for untrained annotators, which jumped to 65% after training. Although these results were obtained for discrete emotion categories, similar misattributions are at least plausible for continuous emotion representations. Annotators' misattribution of arousal and valence could therefore be a possible cause for the lack of predictive power of external annotations on self-reports.

A last explanation can be found in the size and characteristics of the used dataset, as only 18 external annotation sequences with their respective self-reports are available. This amount may be insufficient for a statistical analysis to reveal significant effects, and for models to have too little examples to learn meaningful decision boundaries. Besides this, a single video and retrospective self-report is available per participant. Because of this, it is not possible to account for personal biases that might exist in these self-reports. Due to these biases, participants might rate similar emotions with different levels of arousal and valence. As continuous annotations are obtained through external annotators who rated

the complete dataset, this participant-specific bias is not reflected in the continuous annotations, adding to a reduced coherence between self-reports and said external annotations.

6 PREDICTING SELF-REPORT FROM VISUAL BEHAVIOUR

Our previous experiments revealed that external annotation sequences by themselves contain little predictive value for retrospective self-reports. This indicates that dedicated models are required to be able to estimate self-reports. In this section we investigate whether models can predict retrospective arousal from visual behaviours. Furthermore, we investigate whether external annotations can serve as proxies for relevant affective behaviour to provide the model with additional temporal information during training.

Experiment 3: Predicting self-reports from behaviour cues

To further investigate whether the low predictive power of self-reports is due to incorrect assumptions made on coherence of visual behaviour and retrospective self-reports we apply models directly to uninterpreted behavioural data. We use the output of the FaceReader model to capture facial behaviour and apply the OpenPose model to obtain bodily features, using the combined set of outputs per frame as input to our models to predict Arousal self-reports. We apply similar models as in the previous experiment, only adjusting the models' sizes to account for the larger input space. External annotations are therefore not considered in this experiment and are not part of the input to the models. Results of this experiment can be seen in Table 3. Reported values on the both dataset denote the average accuracy and standard deviation after 10 sessions of 3-fold cross-validation.

Table 3: Model performance for the classification of retrospective self-reported Arousal from behavioural cues on the RECOLA and Mementos dataset. Models are able to predict self-reported Arousal with above chance performance on both datasets. The time-based GRU model outperforms other time-invariant models.

Model	Modality	Aggregation	RECOLA	Mementos
Majority	Face+Body	None	0.39±0.18	0.45±0.02
SVC	Face+Body	Statistics	0.27±0.12	0.51±0.02
MLP	Face+Body	Statistics	0.48±0.24	0.46±0.02
MLP	Face+Body	Resample	0.48±0.11	0.50±0.03
GRU	Face+Body	Resample	**0.57**±0.14	**0.53**±0.03

Results indicate that behavioural features are capable of capturing self-reported emotion with better than chance performance. This does provide an additional indication that the poor predictive power of external annotations found in previous experiments is not due to a low coherence of visual behaviour and retrospective self-reports.
Similar to other affective computing works[4, 35, 98] we find a positive effect of utilising temporal dynamics in our models, as our time-based GRU model outperforms the time-invariant MLP model by a fair margin on the RECOLA dataset. A similar but smaller performance increase is observed for the Mementos dataset indicating that this effect transfers between datasets and persists under the more challenging lighting and pose conditions present in this dataset.

Experiment 4: External annotations as temporal supervision

In the previous experiment we investigated the predictive power of behavioural patterns for self-reports. In these model architectures, models had to learn the relation of a complete input sequence to a single self-report label. As affective episodes are likely to occur sparsely throughout the video, finding these episodes without temporal supervision might be troubling for the model. To remedy this, we add external annotations as temporal supervision to the models. Although we found that external annotations bear a low predictive power for self-reports, they do provide a notion of affective variation over time. Even when the attributed emotion value of these external annotations is completely off, relative changes in these annotations are likely to be caused by changes in participants' behaviour. These changes might help the model with distinguishing relevant segments from background expressions. Therefore, adding the external annotations as an additional supervision, forcing the model to reflect temporal changes in its output could help to place more focus on segments where behavioural changes occur, leading to increased performance. As we aim to predict a temporal affective labels alongside the video-level self-reports we only apply models to the raw annotation sequences, and omit models trained on statistical features from this experiment. As the Mementos dataset does not contain external annotation sequences, results will only be reported on the RECOLA dataset.

Statistical tests to verify whether the addition of temporal supervision causes a significant effect were considered, however doing so would require a larger amount of training data. Standard statistical tests such as the Student t-test assume independence between samples, which is violated by performing cross-validation due to shared training instances between folds. Other tests, such as the McNemar symmetry chi-square test require a too large amount of false negatives or positives to accurately estimate the underlying chi-square distribution[44, 81]. A 5x2 cross-validation scheme would

result in too few train examples to accurately learn the relation between inputs and self-reports. For this reason we have decided to omit such tests in this analysis and draw our conclusions based on the accuracy distributions of the model alone.

Table 4: Model performance for the classification of retrospective Arousal self-reports with different levels of supervision. MLPs models benefit from additional temporal supervision, whereas GRU models are less affected by the level of supervision.

Model	Modality	Temporal supervision	ACC	STD
Majority	None	No	0.39	0.18
MLP	Face+Body	No	0.48	0.11
GRU	Face+Body	No	0.57	0.14
MLP	Face+Body	External Annotation	0.61	0.13
GRU	Face+Body	External Annotation	0.61	0.17

Table 4 shows the performance of the models for different levels of supervision. Figure 6 visualises the model's accuracy distributions. We observe that adding temporal supervision causes a substantial increase in model performance for the MLP model. This indicates that temporal supervision helps the model distinguish relevant cues from background data. As our previous experiments revealed that external annotations bear little predictive value for the prediction of self-reports, this improvement is not likely to be caused by the raw values of the temporal annotation sequences. Instead, it seems more plausible that interactions between the location of changes in the annotation sequences and the model's input are responsible for this increase. As changes in external annotations are likely caused by notable affective changes in a participant's behaviour, training the model to reflect such changes forces it to find patterns in the input sequence explaining this difference. Therefore, doing so provides the model with a sense of what annotators attributed as relevant behaviours. The increase in model performance when adding temporal supervision indicates that differences between segments as indicated by annotators are indeed relevant for the prediction of self-reported emotion.

Interestingly, a smaller performance gain is observed for the GRU model. This result could be explained by the fact that the GRU model is a temporal model which explicitly models for changes over time. It could therefore be that the model can already distinguish similar behavioural patterns over time in the input sequence without specific temporal supervision. This would cause external annotations to convey less relevant information which leads to a smaller performance gain.

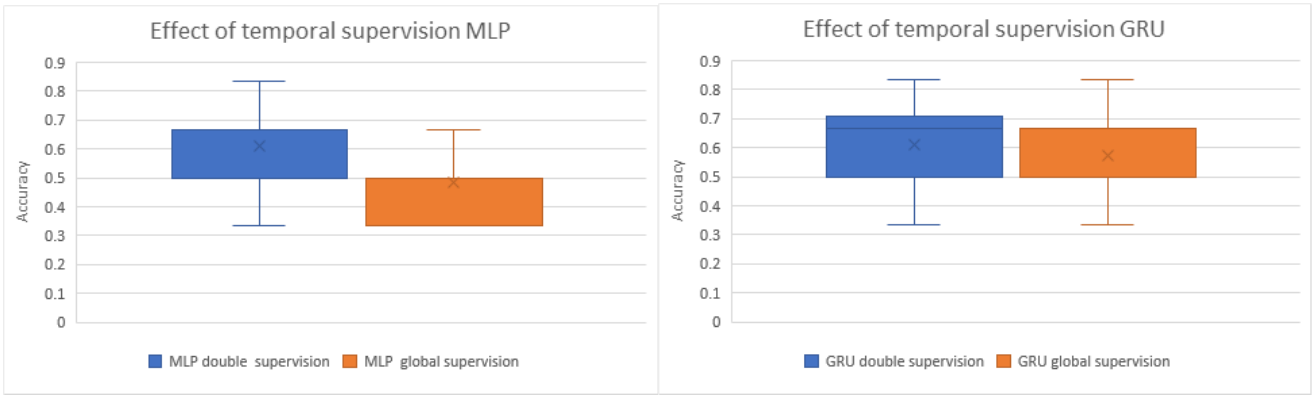
Although our analysis shows a beneficial effect of temporal supervision for both models, care has to be taken during interpretation of our results, as we can not show the significance of this effect due to the size of the dataset. However, inspection of the boxplots in Figure 6a showing the accuracy distributions for both versions of the MLP model gives us a good indication that it is unlikely that these results stem from the same underlying distribution. We therefore conclude that the addition of temporal supervision to the time-invariant MLP model leads to a substantial performance gain, while the time-dependant GRU model benefits less or not at all, potentially due to its underlying capability to capture variations over time without the need for temporal supervision.

7 MODALITY ABLATION STUDY

To investigate how the different modalities contribute to model performance we perform an ablation study on the previously described models. Various psychological works have focused on revealing how emotions manifest themselves through facial and bodily behaviour, showing that both play important roles during both the conveying and interpretation of emotion[2, 16, 52, 100]. For this reason we expect that utilising both modalities within the same model leads to improved model performance. This hypothesis has also been adopted in various other works[4, 27, 36, 98], which show improved model performance, indicating that both modalities contain unique relevant information. These works do however show the contribution of different modalities for the predicting of perceived emotion in the form of external annotations. In this work we research whether similar relationships can be observed for the prediction of retrospective self-reports.

Experiment 3.2: Effect of face and body modalities on prediction of self-reported arousal

To better understand how the face and body modalities separately contribute to the prediction of self-reported arousal, we apply our classification models to both modalities separately. As models in this case only have information regarding one modality, we can compare the performances to the models where both modalities were available to understand the contribution of the removed modality. We report the results for the prediction of self-reported arousal on both the RECOLA and Mementos datasets in Tables 5 and 6 respectively. Similar to previous experiments, values represent average accuracy scores and standard deviations obtained by performing 10×3 cross-validation.



(a) Multilayer Perceptron **(b) Gated Recurrent Unit**

Figure 6: Model Accuracy distributions for different levels of supervision. A substantial increase in performance for the MLP model is observed, whereas the increase is absent for the GRU model.

Table 5: Model performance for the classification of retrospective self-reports on the RECOLA dataset using different input modalities. Values represent accuracy and standard deviation over repeated k-fold validation

Model	Face	Body	Full
Majority	0.39±0.18	0.39±0.18	0.39±0.18
SVC	0.29±0.12	0.24±0.16	0.27±0.12
MLP_{stat}	0.37±0.15	0.30±0.12	0.48±0.24
MLP_{seq}	0.48±0.22	0.46±0.14	0.48±0.12
GRU	**0.52**±0.21	**0.48**±0.14	**0.57**±0.18

Table 6: Model performance for the classification of retrospective self-reports on the Mementos dataset using different input modalities. Values represent accuracy and standard deviation over repeated k-fold validation

Model	Face	Body	Full
Majority	0.45±0.02	0.45±0.02	0.45±0.02
SVC	0.51±0.01-	0.45±0.01	0.51±0.02
MLP_{stat}	0.47±0.02	0.43±0.02	0.46±0.02
MLP_{seq}	0.50±0.03	0.46±0.04	0.50±0.03
GRU	**0.51**±0.03	**0.46**±0.02	**0.53**±0.03

We observe that in the RECOLA dataset, the body modality does convey some significant information as both the MLP and GRU models applied to the sequential body data achieve above majority class accuracy scores. This effect is however not observed for the Mementos dataset, as model performance lies very close to majority class prediction scores. This indicates that the body modality does convey little predictive value on its own in Mementos, possibly due to the in-the-wild nature of the dataset. As illumination conditions strongly vary in this dataset, body poses as predicted by the OpenPose model might contain a significant amount of noise. This is in contrast to the RECOLA dataset, where data was collected in a controlled environment.

The face appears to be a better predictor for self-reported arousal, as we achieve above random chance performance for multiple models in the Mementos dataset. We also observe that models using only the facial modality consistently outperform models trained solely on the body modility. It does therefore seem that our models are better able to capture relationships between facial expressions and self-reported arousal than capturing such relationships between body pose and self-reports. Although the body modality conveyed relevant information about self-reported arousal for the RECOLA dataset, training models on the combined modalities in general does not seem to improve model performance. Our results reveal that only the GRU model consistently benefits from modality fusion. We attribute this to the models ability to model for temporal dependencies between the face and body modalities. All other models either assume time points are independent (the MLP model trained on sequence data), or can only model temporal variations within the same modality (the models trained on the statistical representations). As various works found that there might exist an asynchrony between bodily movements and corresponding facial behaviour [37, 112], they might therefore not be able to benefit from the addition the body modality.

Experiment 4.2: Effect of modalities on prediction of self-reported arousal with temporal supervision

In our earlier experiments we found that the addition of temporal supervision led to a substantial increase model performance for the MLP model, and to a lesser extend to the GRU model. To investigate whether particular modalities benefit more from this supervision, we perform a similar ablation study as in our previous experiment. That is, we train the models on the individual modalities and compare their performance to models trained with both modalities available. Model performance for the different modalities and levels of supervision on the RECOLA dataset is reported in Table 7.

In general we observe a similar relative performance pattern between the face and body modalities as in our previous experiment; the face modalities outperforms the body modality for all tested models. Interestingly, we find an increase in performance for all modalities when temporal supervision is added. This provides additional empirical evidence that applying external annotation sequences as temporal supervision increases model performance for self-reported arousal. Models trained on the face modality do however show a more substantial increase in performance from the addition of temporal supervision than models trained on the body modality. This indicates that temporal supervision helps the models to better distinguish facial expressions that are relevant for self-reported arousal. The lower performance increase in the body modality could be caused by multiple factors. A first explanation could be that external annotation sequences are more influenced by facial behaviour than by body movements. As participants in the RECOLA dataset were seated and had electrodes attached to their hands, body movements could be restricted and therefore be used more sparsely to convey emotion. A likely result of this would be that changes in external annotation sequences are primarily caused by changes in facial behaviour, and are therefore very relevant for the models using the face modality. However, models trained on the body modality would benefit substantially less from these annotations sequences as changes in the temporal annotations might have no matching changes in body movement. A second but related explanation could be that due to the restricted body movements, the body modality might simply contain less predictive information, causing models to be unable to better discriminate between arousal levels from body movement patterns alone.

In contrast to our previous experiment were models performed equally good or better when modalities were fused, fusing modalities for models with access to temporal supervision appears to slightly degrade performance for both models when compared to the face modality. This could be caused by various factors. First it could be that information of the body modality is largely redundant when adding temporal supervision, as the temporal annotation sequences used for this supervision stem from interpretations of both the face and body modality. When the body modality is primarily used by models to regulate the temporal relevance of facial behaviours, models trained with external annotations can use these annotations for this purpose, rendering the body modality less useful. As a result of this, model performance is likely to be equal or slightly less for models trained on both modalities due to the possibility of incongruency between modalities or noise introduced by the body modality. This hypothesis could however not be verified due to the black-box nature of our models.
Another possibility is that the performance difference is an artefact introduced by the low number of training samples. As the difference in model performances is small and the standard deviations on the accuracy scores relatively high, it could be that both the accuracy distributions for both models are samples stemming from the same underlying distribution. As we cannot perform statistical tests due to the low dataset size and cannot verify whether similar behaviour occurs on the Mementos dataset due to its lack of temporal annotations, no definitive conclusions about the nature of this difference can be drawn.

8 WEAKLY SUPERVISED MODELS FOR AFFECT RECOGNITION

As temporal annotation cost is high for both external annotations and self-reports, weakly-supervised models that do not require these labels during training but are capable of predicting them could help to alleviate annotation costs on large spontaneous in-the-wild datasets. Similar practice is applied in the domain of object or action localisation from videos, where videos or images are temporally annotated with the help of weakly-supervised models[72, 105, 110]. Transferring these types of models to the domain of affective state estimation not only enables training on large sparsely labelled datasets, but can also be used to simultaneously temporally annotate said data. To get an insight in the viability of transferring these models to the domain of affect estimation, we perform a twofold of experiments. The first experiment aims to evaluate the global video-level prediction, while the second experiment focusses on the model's performance during temporal label generation. These experiments will be described in more detail in the following subsections.

Table 7: Model performance for the classification of retrospective self-reports using different input modalities. Values represent accuracy and standard deviation over repeated k-fold validation

Model	Temporal Label	Face	Body	Face+Body
Majority	-	0.39±0.18	0.39±0.18	0.39±0.18
MLP	No	0.48±0.22	0.46±0.14	0.48±0.12
	External Annotation	**0.64**±0.23	0.50±0.16	**0.61**±0.13
GRU	No	0.52±0.21	0.48±0.14	0.57±0.18
	External Annotation	0.63±0.18	**0.53**±0.15	0.61±0.17

Experiment 5: Predicting self-reports using weakly supervised models

Traditional supervised models such as the ones used in Section 5 often start with the assumption that every feature or time-step is equally important and attempt to distinguish relevant features during training. On the contrary, weakly-supervised models such as MIL models inherently assume redundancy is present in data. To test whether this helps models during prediction, we deploy two types of weakly supervised models; the MILBoost model as used by Sikka et al. [92] and an Action Localisation model, which is an adapted version of the UntrimmedNet[106]. We compare these models against the models predicting self-reported Arousal as used in Section 6. Similar to previous experiments, models are trained on the RECOLA dataset and attempt to predict a binned Arousal label from behavioural cues. Weakly-supervised models were not applied to the Mementos dataset due to time constraints and is considered as future work. Table 8 shows a comparison between the weakly-supervised models and fully-supervised models in terms of accuracy and standard deviation on retrospective self-report predictions.

Table 8: Comparison of weakly supervised model performance on the prediction of retrospective self-reported Arousal for the RECOLA dataset. Weakly-supervised models do not outperform traditional models for video-level predictions.

Model	Modality	ACC	STD
Majority	None	0.39	0.18
MLP	Face+Body	0.48	0.11
GRU	Face+Body	**0.57**	0.14
MILBoost	Face+Body	0.45	0.24
Action Localisation Net	Face+Body	0.50	0.19

The results of this experiment show that applying weakly supervised models for the prediction of self-report results in equal or worse performance than our tested fully supervised models. This indicates that explicitly modeling for sparsity or irrelevance of time-steps in the model's architecture does not lead to improved classification results. Various explanations can be thought of to explain these results. Firstly, it could be that the fully supervised models are sufficiently capable of dealing with sparsity in the input and explicitly modelling for irrelevant time-steps is not necessary. Secondly, it could be that too little data is available during training to learn meaningful distinctions between relevant and irrelevant time-steps, resulting in comparable performances. However, since models are capable of predicting self-reported Arousal with better than random chance performance, this explanation seems less likely as the model is successful in learning meaningful predictive patterns from the input sequence. However, it would be interesting to research whether this behaviour is observed on other larger datasets, such as Mementos.

Comparing both weakly-supervised models, we observe higher model performance for the time-dependant Action Localisation Net. This increase in performance could be caused by the ability of the Action Localisation Net to model for temporal dynamics in the input sequence by means of its recurrent layer. Alternatively, differences in the underlying assumptions on sparsity these models make could have caused this difference in performance. In the MILBoost model a single positive instance is enough for a video to be classified as that positive label. As such, videos might be classified as high Arousal based on a single segment containing high Arousal. Participants might however take more than a single episode of high Arousal into account in their self-reported score. The Action Localisation Net on the other hand obtains video-level classifications by a weighted average over the segments. Therefore, video-level predictions are based on the context of the complete video. From our results it seems that the Action Localisation model's capabilities to model for temporal dependencies in both its input and output leads to increased performance. This is in accordance with our results from our previous experiments, where we find that the recurrent GRU model consistently outperforms the time-invariant MLP model when no temporal supervision is available.

Experiment 6: Predicting temporal annotations using weakly supervised models

In our previous experiment we tested the performance of weakly-supervised models on video-level predictions. However, these models are inherently capable of predicting on a temporal level while being trained solely on video-level labels. To predict informative temporal labels, models have to be capable of learning the complex relation between a video-level label and the temporal manifestation of these labels in the video itself. Whereas labels are objective in domains such as action localisation (a video either does or does not contain an action), global labels for affective computing are harder to interpret, as they are subjective measures and might have no clear manifestation in a video. Besides this, episodic information is quickly forgotten in retrospective reports, leading to biases towards peak and end effects[46, 85]. Lastly, we found no significant relation between the continuous annotations and retrospective self-reports in our previous experiments.

For these reasons, localisation of perceived emotional episodes from a retrospective self-report label might be intractable at this time. Therefore, we construct a simplified experiment where we disregard the complex relationship introduced by biases in memory and differences in annotation sources. Instead of using retrospective self-reports, we construct a new video-level label which is the result of a simple max aggregation over the external annotation sequences. This newly constructed label is what will be used as supervision to the weakly supervised models. We compare these weakly supervised models with the MLP and GRU models with temporal heads that have been solely trained on the external annotation sequence. Furthermore, we include variations of these MLP and GRU models in our experiment who are solely trained on the newly constructed video-level label. Doing so allows us to compare dedicated weakly-supervised architectures against more naive models trained on similar labels.

The metric used in this experiment is Lin's Concordance Correlation Coefficient (CCC)[57], which is the evaluation metric used in the the Audio/Visual Emotion Challenges (AVEC) measuring agreement between temporal model predictions and a target sequence. This metric is unbiased to changes in scale and location and includes both information on precision and accuracy[57]. Lin's Concordance Correlation Coefficient is defined as in Equation 2:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{2}$$

In this equation ρ_c defines the CCC, ρ denotes the Pearson correlation between the two time series, μ_z and σ_z are the mean and standard deviation of time series z respectively. This metric can interpreted as a penalised Pearson correlation, where the penalty is determined by the squared perpendicular deviation from the 45° line drawn between pairs of the two time series to penalise for differences in scales. The CCC is equal to +1 if sequences are in perfect agreement, and -1 if sequences are in perfect disagreement.

We report our results on the test partition of the RECOLA dataset, where models are trained using both the train and test partitions. Because continuous annotation sequences can contain a delay as annotators are unable to respond to displayed changes instantly[62, 73], for each model we compute the optimal forward shift on the validation set. To do so, we shift our predictions forward by 0 to 8 seconds in steps of 400 milliseconds. Missing values introduced by this shift are filled with the prediction of the first timestep. A similar approach is applied to the baseline model of the AVEC 2017 challenge [83]. Figure 7 provides a visual explanation of this shift. The shift value obtained on the validation set is then applied to the predicted test sequences after which the CCC value is computed. Besides the value of the CCC metric, we report the applied shift value in Table 9.

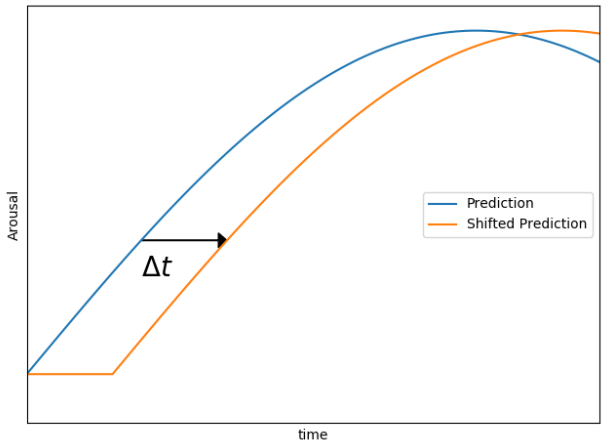


Figure 7: Visual description of shifting temporal model predictions forward in time.

From these results we observe that the MILBoost model seems to be unable to capture meaningful information, as its Concordance Correlation Coefficient value lies very close to zero indicating no agreement between the model predictions and the true annotation labels. Interestingly, model performance of the MILBoost model is similar to those of naive temporal prediction models that were solely trained

Table 9: Comparison of model performance on the prediction of temporal Arousal annotations. Weakly supervised models are able learn some variations in the external annotation sequences, but are substantially outperformed by fully-supervised variants.

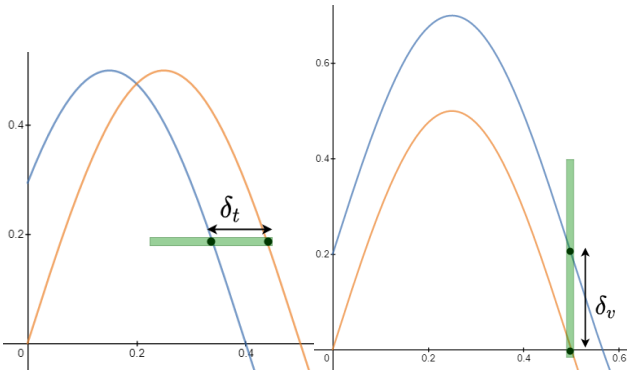
Model	Modality	Temporal Label	Global Label	CCC	delay
SVR[83]	BoVW	External Annotation	No	**0.308**	600ms
MLP	Face+Body	External Annotation	No	0.211	800ms
GRU	Face+Body	External Annotation	No	0.274	0ms
MLP	Face+Body	No	$\max(y_{ext})$	0.019	2800ms
GRU	Face+Body	No	$\max(y_{ext})$	0.038	1600ms
MILBoost	Face+Body	No	$\max(y_{ext})$	0.020	3600ms
Action Localisation	Face+Body	No	$\max(y_{ext})$	**0.088**	800ms

on video-level label. This indicates that the MILBoost's specialised architecture does not yield a performance benefit for this task. This could be caused by the way we constructed our temporal outputs for the MILBoost models. As the MILBoost model is a classification model, we need to convert its temporal outputs to continuous values in order to compute the CCC value. This conversion from categorical predictions to regression values might introduce a significant amount of noise to the temporal prediction, causing a low agreement score.

Contrasting this, the Action Localisation Net model does seem to capture some correlations with the true external annotation sequences and outperforms the naive models, indicating that its architecture allows for building a more accurate representation of relevant segments over time. However, when compared to models that had access to temporal supervision during training, we observe that it is significantly outperformed by such models. Therefore it seems currently unfeasible to apply weakly-supervised models to generate accurate temporal labels as an alternative for external annotators. However, as the dataset size is relatively small, the model's ability to achieve positive agreement with the external annotation sequences without explicit temporal supervision indicates that these models could show promising results when trained on larger datasets.

Experiment 7: Effects of time and value relaxation on model performance

To gain a deeper understanding where weakly-supervised models lack performance compared to their fully-supervised counterparts, we perform an additional set of experiments. In these experiments we research the relation between the model's predicted signal and the true annotation sequence in terms of differences in time and value. As the CCC metric applies a penalty for differences in both time and value between signals, the decrease in model performance could be caused by either factor. Therefore, in this experiment we try to make the relationship between these differences and model performance more explicit. To do so, we apply a tolerance based accuracy metric. Given a list of predictions, target labels and tolerance values, this metric is computed as in Pseudocode 1. The metric can be interpreted by overlapping each point on the predicted sequence with a box whose width and height are equal to an allowed time and value offset respectively. A point is considered as an accurate prediction if any point in the true annotation sequence falls within this box. Figure 8 provides a visual interpretation of this metric for a single point in the predicted sequence. Results of this experiment can be found in Tables 9a to 9c. Visual representations of these tables can be found in Figure 10.



(a) Time tolerance **(b) Value tolerance**

Figure 8: Examples for time and value tolerances for a single prediction point. Blue lines indicate model predictions, orange lines denote target labels. The green box denotes the range in which a target label point needs to lie for the prediction to be considered correct.

	0s	2s	4s	6s	8s	10s	20s
0.01	0.03	0.26	0.4	0.51	0.58	0.64	0.78
0.04	0.13	0.37	0.51	0.6	0.67	0.71	0.84
0.08	0.24	0.47	0.59	0.67	0.73	0.77	0.88
0.11	0.34	0.55	0.66	0.74	0.79	0.82	0.91
0.15	0.44	0.63	0.73	0.79	0.83	0.86	0.94
0.18	0.52	0.69	0.78	0.83	0.87	0.89	0.95
0.35	0.83	0.91	0.95	0.96	0.97	0.98	1.0

(a) MILBoost

	0s	2s	4s	6s	8s	10s	20s
0.01	0.04	0.15	0.22	0.28	0.33	0.37	0.5
0.04	0.22	0.31	0.41	0.44	0.49	0.54	0.63
0.08	0.37	0.48	0.54	0.59	0.62	0.67	0.78
0.11	0.53	0.61	0.67	0.71	0.75	0.77	0.86
0.15	0.66	0.73	0.76	0.81	0.85	0.87	0.94
0.18	0.76	0.8	0.84	0.88	0.9	0.91	0.97
0.35	0.99	1.0	1.0	1.0	1.0	1.0	1.0

(b) Action Localisation Net

	0s	2s	4s	6s	8s	10s	20s
0.01	0.05	0.28	0.4	0.48	0.53	0.58	0.71
0.04	0.23	0.44	0.55	0.61	0.66	0.7	0.79
0.08	0.4	0.58	0.66	0.72	0.76	0.78	0.84
0.11	0.54	0.69	0.76	0.8	0.82	0.84	0.88
0.15	0.66	0.78	0.82	0.85	0.87	0.88	0.9
0.18	0.75	0.83	0.86	0.88	0.89	0.9	0.91
0.35	0.93	0.94	0.95	0.95	0.96	0.96	0.97

(c) GRU with temporal supervision

Figure 9: Average accuracy values for different levels of time and value tolerance for different models.

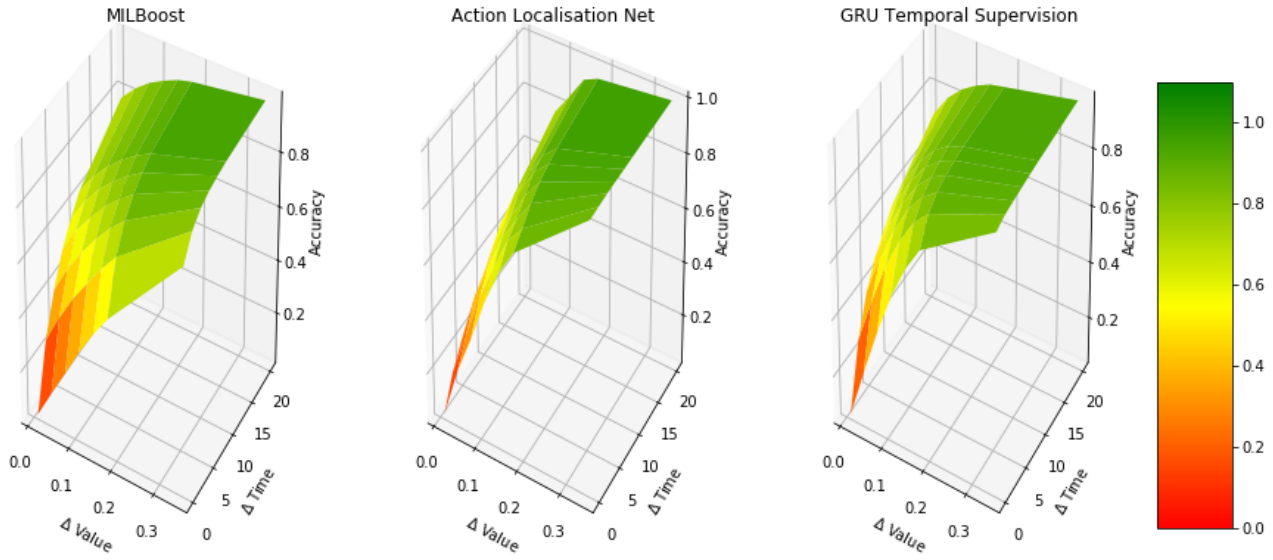


Figure 10: Surface plot of prediction accuracy for different levels of value and time tolerances for different models.

We observe that the MILBoost model preformance appears to increase linearly as both the time and value thresholds get relaxed. This indicates that the models original model predictions are off in both the time and value domains, leading to low quality predictions. The Action Localisation Net on the other hand seems to be less influenced by increases of the time tolerance. We find that increasing the time tolerance when no value tolerance is allowed has very little effect on accuracy scores. This indicates that the model systematically predicts a value that is either too high or low compared to the true labels. This is verified by the sharper increase of accuracy values along the value tolerance axis for low time offsets. This indicates that the model is already relatively capable of capturing some temporal variations in a relatively close time proximity, but attributes incorrect values to these patterns. This ability to better capture temporal variations could be explained by the models ability to model for temporal dependencies in the input, which are not modelled for in the MILBoost model.

When comparing the fully supervised GRU model to the Action Localisation network, we observe that the fully supervised model achieves roughly similar accuracy performance when no time tolerance is allowed. However, compared to the Action Localisation Net, the fully supervised model's accuracy values at low time and value tolerances increase faster in both directions as more tolerance is allowed. This indicate that predictions of the fully supervised model on average lie closer to the target label, both in time proximity as in value proximity, leading to superior performances.

Algorithm 1: Computation of tolerance based Accuracy matrix

input : Prediction X; Target Y, Time-tolerance list TT, Value-Tolerance list VT
Result: Tolerance based accuracy matrix
$P \leftarrow$ length X;
$N \leftarrow$ length TT;
$M \leftarrow$ length VT;
acc_tol_sum $\leftarrow N \times M$ Matrix;
for $t \leftarrow 0$ **to** P **do**
  for $i \leftarrow 0$ **to** N **do**
    $\delta_t \leftarrow TT_i$;
    time_range $\leftarrow Y_{t-\delta_t}$ to $Y_{t+\delta_t}$;
    for $j \leftarrow 0$ **to** M **do**
      $\delta_v \leftarrow VT_j$;
      prediction_diff $\leftarrow |time\_range - X_t|$;
      if $\exists$ *point* $\in$ *prediction_diff: point* $\leq \delta_v$ **then**
        acc_tol_sum[i,j] + = 1;
      end
    end
  end
end
return $\frac{acc\_tol\_sum}{P}$

Summarising, we find that the performance differences between the MILBoost model and the Action Localisation Net can be attributed to the Action Localisation model's ability to predict more relevant values at low time tolerances. Furthermore, training models in a fully-supervised manner allows the model to form predictions that are closer to the target sequence in terms of both value and time.

Overall, we find that applying weakly supervised models for the prediction of self-reported emotion yields similar performance to unspecialised model architectures, indicating that explicitly modelling for sparsity does not lead to improved classification results. Besides this, we find that tasking the weakly supervised models with the prediction of external annotation sequences seems unfeasible at this time. Although the Action Localisation Net is able to achieve a positive agreement with the external annotation sequences, performance is low compared to fully supervised models. Analysis revealed that training with temporal supervision allows fully supervised models to make estimations that are both closer in value and time compared to weakly supervised models. However, as the Action Localisation model was able to capture positive agreement with external annotation sequences from little training data, this model could be promising when applied to larger datasets.

9 DISCUSSION

A major in challenge in the domain of affect classification is the inability to obtain continuous ground truth labels for experienced emotion. In this work we have focused on the relation between the two main alternative label types researchers adopt in their models: continuous external annotation sequences and retrospective self-reports. Through various conducted experiments we are able to answer the research questions that were posed in the introduction of this work. In this section, we will address each question separately and discuss the findings of the related experiments.

Research Question 1.1: Are external annotations significant predictors for retrospective self-reported emotional experiences? How do they compare to raw visual behaviours?

In our first set of experiments, we evaluate the predictiveness of external annotations using a statistical analysis on aggregated features, as well as applying increasingly powerful predictive models. We find that neither the statistical analysis nor the predictive models reveal a significant predictive effect. This implies that external annotations alone do not carry sufficient significant information to accurately capture self-reports. This finding has several implications for researchers attempting to predict retrospective self-reports. First of all, external annotations seem to capture a different psychological construct than retrospective self-reports, as is indicated by the lack of feature significance. This could have multiple explanations, among which are annotator misattribution of experienced emotion from expressions, dataset size, and the fact that we cannot account for personal biases in self-reports. Therefore, these results should be treated with care and further experiments on larger datasets should be conducted to verify these results.
Second, results from applying the external annotations as direct input to predictive models indicate that transfer learning using existing models trained to predict these external annotations is not a viable option on its own. Therefore, dedicated models have to be constructed to capture self-report, leading to the necessity of larger datasets as relations have to be learned from scratch.

Although transfer learning from models predicting external annotations seems infeasible, using pretrained models that predict facial expressions and bodily keypoint locations show promising results. Our results indicate that training on the outputs of such pretrained models allows for better than chance performance for the predicting of self-reported

arousal. This indicates that relevant visual behaviour can indeed be captured by such models and that visual behaviour can successfully be used for the prediction of self-reports. Results of our ablation study revealed that the facial modality conveys the most relevant information for the prediction of self-reports and that bodily behaviour alone seems to contain insufficient information for accurate self-report predictions. Furthermore, we found that only the GRU model consistently benefited from the fusion of modalities, which we attribute to its ability to model for dynamics between modalities.

Research Question 1.2: Do external annotations help to seperate emotional episodes from neutral segments?
The prediction of self-reported emotion is in itself a challenging field, as only a single video-level label is available for models to learn from. This forces the models to learn the relationship between temporal behavioural cues and retrospective self-reports without temporal supervision. We researched whether applying external annotations sequences as an additional level of supervision helps to overcome this issue, by forcing the model to reflect changes in these annotation sequences over time. Doing so resulted in improved model performance, which indicates that although external annotations in itself might not be viable predictors for self-report, forcing models to reflect changes in the annotation sequences helps to distinguish relevant segments in the input video.

With these joined results of our sets of experiments, we are able to formulate an answer to our first research question:

Research Question 1: How do visual behaviours relate to retrospective self-reports of experienced emotion? Does the utilisation of external annotations improve the predictive power?

Results on two datasets showed that the facial modality is most predictive of self-reported arousal and that visual behaviours can predict self-reported arousal with better than chance performance. External annotations by themselves seem to contain insufficient information to capture self-reported arousal, but applying them as temporal supervision consistently leads to improved model performance. This indicates that interactions between external annotation sequences and the visual behaviours help the model to distinguish relevant sections for the prediction of self-reports.

Obtaining continuous external annotations is however very costly which hinders the creation of large-scale datasets. In this work we proposed two weakly-supervised models to predict such temporal annotations automatically without the need for continuous labels during training. We conducted experiments to investigate the performance of weakly supervised models for both the prediction of self-reported emotion, as well as their performance on the prediction of continuous annotations. This allows us to answer our second set of research questions:

Research Question 2.1 How does retrospective self-report prediction performance compare to fully-supervised models?
Our results on the RECOLA dataset indicate that weakly-supervised models can be used for the prediction of self-reported emotion with above chance performance. However, they are outperformed by our tested fully-supervised GRU model. This indicates that explicitly modelling for sparsity in the input does not help these models for the prediction of self-reported arousal. This could indicate that fully-supervised models are sufficiently capable of dealing with the sparsity of emotional displays in the input. However, this behaviour could also be caused by the low amount of training samples and results should therefore be verified on larger datasets such as Mementos.

Research Question 2.2 How well can weakly supervised models predict continuous external annotation labels from video-level labels?
As obtaining continuous temporal affective labels using external annotators is time-consuming and costly, the automatic generation of such annotations using predictive models has gained significant attention. However up to now, all of these models require the availability of temporal labels during training. We tested whether our weakly-supervised models are able to predict accurate temporal annotations trained solely on video-level labels. Our results show that the proposed Action Localisation model is able to achieve some agreement with the true external annotation sequences (Lin's $CCC = 0.09$), but in general is substantially outperformed by fully-supervised models (Lin's $CCC = 0.27$). An analysis of the effects of time and value tolerances on the performance of the Action Localisation model revealed that model performance was more dependant on value tolerances than on time tolerance. This indicates that the model is relatively capable of capturing temporal variations with a close time proximity, but attributes incorrect values to these patterns. When compared to fully-supervised models, we find that training on continuous labels results in predictions that are closer to the target sequence in both value and time proximity.

With these results on the performance of weakly supervised models, we can address our final research question:

Research Question 2 How well can weakly supervised models be deployed to predict continuous annotations from video-level annotations?

Our results indicate that applying weakly supervised models to the prediction of self-reported arousal does not lead to improved accuracy. Furthermore, when applied to the domain of predicting perceived emotion sequences in the form of external annotations we observe that these models are substantially outperformed by fully-supervised variants. For these reasons it seems that the application of weakly-supervised models in this domain is currently unfeasible. However, since the weakly-supervised Action Localisation model was able to capture some agreement with external annotation sequences from a small amount of input data, we feel that such models could be promising when applied to larger amounts of training data.

limitations & future work

Several limitations might hinder the generalisability of our results to different datasets. The first and most apparent limitation is the size of the used RECOLA dataset. Due to the low amount of available data in this dataset, obtained results indicating that external annotation sequences are not predictive for self-reported arousal should be treated with care. Although these results are in line with previous research of Hirt et al. [40], who found that automated systems trained to predict external annotations are not predictive of retrospective self-reports of interest, boredom and valence, additional experiments on larger corpora should be conducted to gain a better understanding of the complex relationship between the two annotation types.

Another limitation of our work is that we did not account for personal biases in self-reports. As works in literature have found that the experience of emotion is highly subjective[70], accounting for these personal biases could help models during training. Furthermore, experiments conducted on the relation between perceived emotion and bias-corrected self-reports could provide additional insights into the relationship between experienced and perceived emotion. Future works could attempt to account for personal biases through the usage of the relative difference between reports collected at the start and the end of stimuli or through the usage of personalized models trained per individual.

Lastly, in this work we applied weakly-supervised models for the prediction of perceived annotation sequences that were trained on an aggregated video-level label. The choice of this aggregation function might have significantly influenced the model's ability to predict temporal annotation sequences. Further research should be conducted to investigate whether the usage of different aggregation functions improves the model's performance. Ideally however, datasets are constructed containing both continuous and global level labels stemming from the same annotator source, removing the need for a manually defined aggregation function. The construction of datasets containing continuous self-reported emotion is however costly and involved as methods need to be found to obtain such ground truths without significantly distracting participants from the stimuli they are exposed to. An alternative and perhaps more feasible option would be to construct datasets containing continuous perceived annotation labels obtained from external annotations in addition to a global retrospective label provided by said annotators. Applying weakly supervised models to such a dataset would indicate the viability of generating continuous annotations from a single video-level perceived emotion label. Positive results would significantly reduce the annotation burden on external annotators, as they would only need to report a single label per video. Although our results suggest that it might be too early to successfully apply weakly-supervised models to the domain of affect classification, rapid technological advances in these models are made in domains such as Action Localisation in videos. As such, a successful transfer of these models to the domain of affect estimation might simply be a matter of time.

REFERENCES

[1] Jennifer Aaker, Aimee Drolet, and Dale Griffin. 2008. Recalling Mixed Emotions. *Journal of Consumer Research* 35, 2 (2008), 268–278. https://doi.org/10.1086/588570

[2] Nalini Ambady and Robert Rosenthal. 1992. *Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis.* Technical Report 2. 256–274 pages.

[3] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (jul 2019), 1–68. https://doi.org/10.1177/1529100619832930

[4] Pablo Barros, Doreen Jirak, Cornelius Weber, and Stefan Wermter. 2015. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks* 72 (2015), 140–151. https://doi.org/10.1016/j.neunet.2015.09.009

[5] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. [n.d.]. *A Training Algorithm for Optimal Margin Classifiers.* Technical Report.

[6] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (mar 1994), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

[7] Joost Broekens and Willem Paul Brinkman. 2013. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human Computer Studies* 71, 6 (jun 2013), 641–667. https://doi.org/10.1016/j.ijhcs.2013.02.003

[8] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. 2015. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371* (2015).

[9] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January (dec 2018), 1302–1310. arXiv:1812.08008 http://arxiv.org/abs/1812.08008

[10] Prithwi Raj Chakraborty, Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. 2015. Using viewer's facial expression and heart rate for sports video highlights detection. In *ICMR 2015 - Proceedings of the 2015 ACM International Conference on Multimedia Retrieval.* Association for Computing Machinery, Inc, 371–378. https://doi.org/10.1145/2671188.2749361

[11] Shizhi Chen, Yingli Tian, Qingshan Liu, and Dimitris N. Metaxas. 2011. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* https://doi.org/10.1109/CVPRW.2011.5981880

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference.* Association for Computational Linguistics (ACL), 1724–1734. https://doi.org/10.3115/v1/d14-1179 arXiv:1406.1078

[13] Junyoung Chung, Caglar Gulcehre, and Kyunghyun Cho. [n.d.]. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* Technical Report. arXiv:1412.3555v1

[14] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid Inria. [n.d.]. *Multi-fold MIL Training for Weakly Supervised Object Localization.* Technical Report.

[15] James A. Coan and John J.B. Allen. 2008. The Handbook of emotion elicitation and assessment. *Choice Reviews Online* 46, 03 (2008), 46–1769–46–1769. https://doi.org/10.5860/choice.46-1769

[16] Mark Coulson. 2004. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior* 28, 2 (2004), 117–139.

[17] Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals.* Oxford University Press, USA.

[18] Marco de Meijer. 1989. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior* 13, 4 (dec 1989), 247–268. https://doi.org/10.1007/BF00990296

[19] Sylvie Droit-Volet, Sophie Brunot, and Paula M Niedenthal. 2004. Perception of the duration of emotional events. *Cognition and Emotion* 18, 6 (2004), 849–858. https://doi.org/10.1080/02699930341000194

[20] Bernd Dudzik, Hayley Hung, Mark Neerincx, and Joost Broekens. [n.d.]. Investigating the Influence of Personal Memories on Video-Induced Emotions. ([n. d.]).

[21] Daniel A Effron, Paula M Niedenthal, Sandrine Gil, and Sylvie Droit-Volet. 2006. Embodied Temporal Perception of Emotion. (2006). https://doi.org/10.1037/1528-3542.6.1.1

[22] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[23] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.

[24] Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science* 164, 3875 (1969), 86–88. https://doi.org/10.1126/science.164.3875.86

[25] Hillary Anger Elfenbein. 2016. On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis. (2016). https://doi.org/10.1037/0033-2909.128.2.203

[26] Jose-Miguel Fernandez-Dols, Flor Sanchez, Pilar Carrera, and Maria-Angeles Ruiz-Belda. 1997. *ARE SPONTANEOUS EXPRESSIONS AND EMOTIONS LINKED? AN EXPERIMENTAL TEST OF COHERENCE.* Technical Report.

[27] Panagiotis P Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. 2019. Fusing Body Posture with Facial Expressions for Joint Recognition of Affect in Child-Robot Interaction. 1 (2019). https://doi.org/10.1109/LRA.2019.2930434 arXiv:1901.01805v3

[28] Mara Fölster, Ursula Hess, and Katja Werheid. 2014. Facial age affects emotional expression decoding. , 30 pages. https://doi.org/10.3389/fpsyg.2014.00030

[29] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. *Deep Sparse Rectifier Neural Networks.* Technical Report.

[30] Donald Glowinski, Antonio Camurri, Gualtiero Volpe, Nele Dael, and Klaus Scherer. 2008. Technique for automatic emotion recognition by body gesture analysis. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* IEEE, 1–6.

[31] Arepalli Peda Gopi, • R Naga, Sravana Jyothi, • V Lakshman Narayana, and • K Satya Sandeep. [n.d.]. Classification of tweets data based on polarity using improved RBF kernel of SVM. ([n. d.]). https://doi.org/10.1007/s41870-019-00409-4

[32] Yuanyuan Gu, Xiaoqin Mai, and Yue-jia Luo. 2013. Do Bodily Expressions Compete with Facial Expressions? Time Course of Integration of Emotional Signals from the Face and the Body. *PLoS ONE* 8, 7 (jul 2013), e66762. https://doi.org/10.1371/journal.pone.0066762

[33] Amogh Gudi and Vicarvision Amsterdam. [n.d.]. *Recognizing Semantic Features in Faces using Deep Learning.* Technical Report. arXiv:1512.00743v2 http://www.vicarvision.nl/

[34] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. [n.d.]. *Deep Learning based FACS Action Unit Occurrence and Intensity Estimation.* Technical Report.

[35] Hatice Gunes, Mihalis A Nicolaou, and Maja Pantic. 2011. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing* (2011), 1. https://doi.org/10.1109/T-AFFC.2011.9

[36] Hatice Gunes and Massimo Piccardi. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications* 30, 4 (nov 2007), 1334–1345. https://doi.org/10.1016/j.jnca.2006.09.007

[37] Hatice Gunes, Massimo Piccardi, and Senior Member. 2008. Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display Affective gaming in VR for cognitive training View project Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display. (2008). https://doi.org/10.1109/TSMCB.2008.927269

[38] P Paul Heppner, Bruce Wampold, Jesse Owen, Mindi Thompson, and Kenneth Wang. 2016. *Research Design in Counseling*. 334 pages.

[39] Geoffrey Hinton, Ni@sh Srivastava, and Kevin Swersky. [n.d.]. *Neural Networks for Machine Learning Lecture 6a Overview of mini--batch gradient descent*. Technical Report.

[40] Franziska Hirt, Egon Werlen, Ivan Moser, and Per Bergamin. 2019. Measuring emotions during learning: Lack of coherence between automated facial emotion recognition and emotional experience. *Open Computer Science* (2019). https://doi.org/10.1515/comp-2019-0020

[41] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[42] Shujun Huang, C. A.I. Nianguang, Pedro Penzuti Pacheco, Shavira Narandes, Yang Wang, and X. U. Wayne. 2018. Applications of support vector machine (SVM) learning in cancer genomics. , 41–51 pages. https://doi.org/10.21873/cgp.20063

[43] William T James. 1932. A Study of the Expression of Bodily Posture. *The Journal of General Psychology* 7, 2 (1932), 405–437. https://doi.org/10.1080/00221309.1932.9918475

[44] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: A classification perspective*. Vol. 9780521196000. Cambridge University Press. 1–406 pages. https://doi.org/10.1017/CBO9780511921803

[45] Rafal Jozefowicz and Wojciech Zaremba. [n.d.]. *An Empirical Exploration of Recurrent Network Architectures*. Technical Report.

[46] Daniel Kahneman, Ed Diener, and Norbert Schwarz. 1999. Well-being: The foundations of hedonic psychology. *Health San Francisco* (1999), xii, 593. https://doi.org/10.7758/9781610443258

[47] Joep J.M. Kierkels, Mohammad Soleymani, and Thierry Pun. 2009. Queries and tags in affect-based multimedia retrieval. In *Proceedings - 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*. 1436–1439. https://doi.org/10.1109/ICME.2009.5202772

[48] Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv:1412.6980 https://arxiv.org/abs/1412.6980v9

[49] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (jan 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[50] Sander Koelstra, Ashkan Yazdani, Mohammad Soleymani, Christian Mühl, Jong Seok Lee, Anton Nijholt, Thierry Pun, Touradj Ebrahimi, and Ioannis Patras. 2010. Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6334 LNAI. 89–100. https://doi.org/10.1007/978-3-642-15314-3_9

[51] Santiago Gerling Konrad, Mao Shan, Favio R. Masson, Stewart Worrall, and Eduardo Nebot. 2018. Pedestrian Dynamic and Kinematic Information Obtained from Vision Sensors. In *IEEE Intelligent Vehicles Symposium, Proceedings*, Vol. 2018-June. Institute of Electrical and Electronics Engineers Inc., 1299–1305. https://doi.org/10.1109/IVS.2018.8500527

[52] Mariska E. Kret, Karin Roelofs, Jeroen J. Stekelenburg, and Beatrice de Gelder. 2013. Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. *Frontiers in Human Neuroscience* 7, DEC (2013). https://doi.org/10.3389/fnhum.2013.00810

[53] Ivar Krumpal. 2011. Determinants of social desirability bias in sensitive surveys: a literature review. (2011). https://doi.org/10.1007/s11135-011-9640-9

[54] Robert W Levenson. 1994. Human emotion: A functional view. *The nature of emotion: Fundamental questions* 1 (1994), 123–126.

[55] Ting Li, Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, and Liming Chen. 2015. Continuous arousal self-assessments validation using real-time physiological responses. In *ASM 2015 - Proceedings of the 1st International Workshop on Affect and Sentiment in Multimedia, co-located with ACM MM 2015*. Association for Computing Machinery, Inc, 39–44. https://doi.org/10.1145/2813524.2813527

[56] James J. Lien, Jeffrey F. Cohn, Takeo Kanade, and Ching Chung Li. 1998. Automated facial expression recognition based on FACS action units. In *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*. IEEE Computer Society, 390–395. https://doi.org/10.1109/AFGR.1998.670980

[57] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (mar 1989), 255. https://doi.org/10.2307/2532051

[58] Yuan Pin Lin, Chi Hong Wang, Tien Lin Wu, Shyh Kang Jeng, and Jyh Horng Chen. 2007. Multilayer perceptron for EEG signal classification during listening to emotional music. In *IEEE Region 10 Annual International Conference, Proceedings/TENCON*. https://doi.org/10.1109/TENCON.2007.4428831

[59] Zhilei Liu, Shangfei Wang, Zhaoyu Wang, and Qiang Ji. 2013. Implicit video multi-emotion tagging by exploiting multi-expression relations. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. https://doi.org/10.1109/FG.2013.6553767

[60] Hanneke K M Meeren, Corné C R J van Heijnsbergen, and Beatrice de Gelder. 2005. *Rapid perceptual integration of facial expression and emotional body language*. Technical Report 45. www.pnas.orgcgidoi10.1073pnas.0507650102

[61] Larry M Manevitz, Malik Yousef, Nello Cristianini, John Shawe-Taylor, and Bob Williamson. 2001. *One-Class SVMs for Document Classification*. Technical Report Dec. 139–154 pages.

[62] Soroosh Mariooryad and Carlos Busso. 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing* 6, 2 (apr 2015), 97–108. https://doi.org/10.1109/TAFFC.2014.2334294

[63] Llew Mason, Peter Bartlett, Jonathan Baxter, and Marcus Frean. [n.d.]. *Boosting Algorithms as Gradient Descent*. Technical Report.

[64] David Matsumoto, • Hyi, and Sung Hwang. [n.d.]. Evidence for training the ability to read microexpressions of emotion. ([n. d.]). https://doi.org/10.1007/s11031-011-9212-2

[65] Iris B. Mauss, Loren McCarter, Robert W. Levenson, Frank H. Wilhelm, and James J. Gross. 2005. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion* 5, 2 (jun

2005), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

[66] Warren S McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5, 4 (1943), 115–133.

[67] Albert Mehrabian. 1968. Communication without words. *Psychology today* 2, 4 (1968).

[68] Mihir Narayana Mohanty, Mahesh Chandra, H K Palo, and Narayana Mohanty. 2015. Use of Different Features for Emotion Recognition Using MLP Network. 332 (2015). https://doi.org/10.1007/978-81-322-2196-8_2

[69] Arthur G. Money and Harry Agius. 2009. Analysing user physiological responses for affective video summarisation. *Displays* 30, 2 (apr 2009), 59–70. https://doi.org/10.1016/j.displa.2008.12.003

[70] Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review* 5, 2 (apr 2013), 119–124. https://doi.org/10.1177/1754073912468165

[71] Rodrigo Moraes, João Francisco Valiati, and Wilson P. Gavião Neto. 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 40, 2 (feb 2013), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059

[72] Phuc Nguyen, Bohyung Han, Ting Liu, and Gautam Prasad. 2018. Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 6752–6761. https://doi.org/10.1109/CVPR.2018.00706 arXiv:1712.05080

[73] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. [n.d.]. *Robust Continuous Prediction of Human Emotions using Multiscale Dynamic Cues*.

[74] Andrew Ortony and Terence J. Turner. 1990. What's basic about basic emotions? *Psychological Review* 97, 3 (1990), 315–331. https://doi.org/10.1037/0033-295X.97.3.315

[75] Michalis Papakostas, Konstantinos Tsiakas, Theodoros Giannakopoulos, and Fillia Makedon. 2017. Towards predicting task performance from EEG signals. In *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, Vol. 2018-January. Institute of Electrical and Electronics Engineers Inc., 4423–4425. https://doi.org/10.1109/BigData.2017.8258478

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[77] Guoqin Peng and Dan Xu. 2019. Weakly Supervised Learning of Image Emotion Analysis Based on Cross-spatial Pooling. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11818 LNCS. Springer, 116–125. https://doi.org/10.1007/978-3-030-31456-9_13

[78] Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion* 1984 (1984), 197–219.

[79] Akilesh Rajavenkatanarayanan, Ashwin Ramesh Babu, Konstantinos Tsiakas, and Fillia Makedon. 2018. Monitoring Task Engagement using Facial Expressions and Body Postures. (2018). https://doi.org/10.1145/3191801.3191816

[80] Akilesh Rajavenkatanarayanan, Fillia Makedon, Ashwin Ramesh Babu, and James Robert Brady. 2018. Multimodal Approach for Cognitive Task Performance Prediction from Body Postures, Facial Expressions and EEG Signal. (2018). https://doi.org/10.1145/3279810.3279849

[81] Sebastian Raschka. 2018. *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. Technical Report. arXiv:1811.12808v2

[82] Donald A. Redelmeier and Daniel Kahneman. 1996. Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66, 1 (jul 1996), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

[83] Fabien Ringeval, Jonathan Gratch, Sharon Mozgai, Bjorn Schuller, Roddy Cowie, Nicholas Cummins, Maja Pantic, Michel Valstar, Stefan Scherer, and Maximilian Schmitt. 2017. AVEC 2017 - Real-life depression, and affect recognition workshop and challenge. In *AVEC 2017 - Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, co-located with MM 2017*. Association for Computing Machinery, Inc, 3–9. https://doi.org/10.1145/3133944.3133953

[84] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. 1–8. https://doi.org/10.1109/FG.2013.6553805

[85] Michael D. Robinson and Gerald L. Clore. 2002. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin* 128, 6 (2002), 934–960. https://doi.org/10.1037/0033-2909.128.6.934

[86] David E Ruineihart, Geoffrey E Hint, and Ronald J Williams. 1985. *LEARNING INTERNAL REPRESENTATIONS BERROR PROPAGATION two*. Technical Report.

[87] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[88] James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality* 11, 3 (1977), 273–294.

[89] Nora Cate Schaeffer. 2000. Asking questions about threatening topics: A selective overview. In *The science of self-report: Implications for research and practice*. Lawrence Erlbaum Associates Publishers, 105–121. https://psycnet.apa.org/record/1999-04118-007

[90] Charles A. Schreiber and Daniel Kahneman. 2000. Determinants of the remembered utility of aversive sounds. *Journal of Experimental Psychology: General* 129, 1 (2000), 27–42. https://doi.org/10.1037/0096-3445.129.1.27

[91] Tal Shafir, Rachelle P. Tsachor, and Kathleen B. Welch. 2016. Emotion Regulation through Movement: Unique Sets of Movement Characteristics are Associated with and Enhance Basic Emotions. *Frontiers in Psychology* 6, JAN (jan 2016), 2030. https://doi.org/10.3389/fpsyg.2015.02030

[92] Karan Sikka, Abhinav Dhall, and Marian Bartlett. 2013. Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. IEEE, 1–8. https://doi.org/10.1109/FG.2013.6553762

[93] Fernando Silveira, Brian Eriksson, Anmol Sheth, and Adam Sheppard. 2013. Predicting audience responses to movie content from electrodermal activity signals. In *UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 707–716. https://doi.org/10.1145/2493432.2493508

[94] Krishna Kumar Singh and Yong Jae Lee. 2017. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2017-Octob. Institute of Electrical and Electronics Engineers Inc., 3544–3553. https://doi.org/10.1109/ICCV.2017.381 arXiv:1704.04232

[95] Mohammad Soleymani, Sander Koelstra, Ioannis Patras, and Thierry Pun. 2011. Continuous emotion detection in response to music videos. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*. 803–808. https://doi.org/10.1109/FG.2011.5771352

[96] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing* 3, 2 (2012), 211–223. https://doi.org/10.1109/T-AFFC.2011.37

[97] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. 2014. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Technical Report 56. 1929–1958 pages. http://jmlr.org/papers/v15/srivastava14a.html

[98] Bo Sun, Siming Cao, Jun He, and Lejun Yu. 2018. Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. *Neural Networks* 105 (sep 2018), 36–51. https://doi.org/10.1016/j.neunet.2017.11.021

[99] Guoyun Tu, Yanwei Fu, Boyang Li, Jiarui Gao, Yu-Gang Jiang, and Xiangyang Xue. 2019. A Multi-task Neural Approach for Emotion Attribution, Classification and Summarization. *IEEE Transactions on Multimedia* (dec 2019), 1–1. https://doi.org/10.1109/tmm.2019.2922129 arXiv:1812.09041

[100] Jan Van der Stock, Ruthger Righart, and Beatrice de Gelder. 2007. Body Expressions Influence Recognition of Emotions in the Face and Voice. (2007). https://doi.org/10.1037/1528-3542.7.3.487

[101] Hans Van Kuilenburg, Marco Wiering, and Marten Den Uyl. [n.d.]. *A Model Based Method for Automatic Facial Expression Recognition*. Technical Report. http://home.wanadoo.nl/van.kuilenburg/

[102] Vladimir N. Vapnik. 2000. *The Nature of Statistical Learning Theory*. Springer New York. https://doi.org/10.1007/978-1-4757-3264-1

[103] Paul Viola, John C Platt, and Cha Zhang. [n.d.]. *Multiple Instance Boosting for Object Detection*. Technical Report.

[104] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology* 28, 6 (1998), 879–896.

[105] Guolong Wang, Zheng Qin, and Kaiping Xu. 2017. Recognizing emotions based on human actions in videos. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10133 LNCS. Springer Verlag, 306–317. https://doi.org/10.1007/978-3-319-51814-5_26

[106] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. UntrimmedNets for weakly supervised action recognition and detection. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Vol. 2017-Janua. 6402–6411. https://doi.org/10.1109/CVPR.2017.678 arXiv:1703.03329

[107] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. [n.d.]. *Temporal Segment Networks for Action Recognition in Videos*. Technical Report. arXiv:1705.02953v1 https://github.com/yjxiong/temporal-segment-networks/

[108] Shangfei Wang, Zhilei Liu, Yachen Zhu, Menghua He, Xiaoping Chen, and Qiang Ji. 2015. Implicit video emotion tagging from audiences' facial expression. *Multimedia Tools and Applications* 74, 13 (jun 2015), 4679–4706. https://doi.org/10.1007/s11042-013-1830-0

[109] Jian Sheng Wu, Sheng Jun Huang, and Zhi Hua Zhou. 2014. Genome-wide protein function prediction through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11, 5 (2014), 891–902. https://doi.org/10.1109/TCBB.2014.2323058

[110] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, and Eric I.Chao Chang. 2014. Deep learning of feature representation with multiple instance learning for medical image analysis. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 1626–1630. https://doi.org/10.1109/ICASSP.2014.6853873

[111] Hamwira Yaacob, Wahab Abdul, and Norhaslinda Kamaruddin. 2013. Classification of EEG signals using MLP based on categorical and dimensional perceptions of emotions. In *2013 5th International Conference on Information and Communication Technology for the Muslim World, ICT4M 2013*. https://doi.org/10.1109/ICT4M.2013.6518914

[112] Mingming Zhang, Tiantian Liu, Yule Jin, Weiqi He, Yuxia Huang, and Wenbo Luo. 2019. The asynchronous influence of facial expressions on bodily expressions. *Acta Psychologica* 200 (sep 2019), 102941. https://doi.org/10.1016/j.actpsy.2019.102941

[113] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. [n.d.]. *Multi-Instance Learning by Treating Instances As Non-I.I.D. Samples*. Technical Report.

A MODEL PRELIMINARIES

In this section each model used in our experiments will be discussed in more detail. Model background and descriptions of their inner workings will be described in the sections below.

Support Vector Classifier

Model background and applications. Support Vector Machines are binary classification models that are trained to perform maximum margin classification. First introduced by Vapnik [102], these models attempt to find a decision boundary that maximises the distance, or so called margin, of this boundary to the nearest point from each class. Points are classified by the side of the decision boundary these points are located in hyperspace. However, a problem for these models is that in their most naive form, they cannot capture non-linear relationships. To overcome this issue [5] proposed to use non-linear projections of the input before fitting a decision boundary. This projection is widely known as the kernel-trick and allows SVM models to model for non-linear relations. Various different kernel functions have been proposed, among which the Polynomial and Radial Basis Function kernel are most notable.

Another issue that needs to be accounted for in SVMs is the fact that these models in their original definition do not allow for misclassifications and are therefore very sensitive to outliers which significantly shift the location of the decision boundary. To overcome this, the concept of soft-margin SVMs were introduced. By a reformulation of the optimization objective function, the model is allowed to make misclassifications and is therefore less suspetible to outliers in the data. SVMs have been extensively applied in various domains, showing their robustness and ability to work with a high feature dimension or sparsity. This has made them a popular choice in domains such as document classification and gene classification[31, 42, 61, 71]. Affective computing researchers have also deployed them in tasks such as multi-modal affect recognition[35, 96], the prediction of task performance[75] and emotion attribution to videos[99].

Mathematical Definition. Soft-margin SVMs are tasked with minimising the average hinge loss function, which is introduced in Equation 3.

$$\mathcal{L}_{hinge}(\overrightarrow{x_i}) = \max(0, 1 - y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} - b)) \tag{3}$$

In this equation, y_i is the target label for data point $\overrightarrow{x_i}$, $\overrightarrow{w}$ is the models learned normal vector to the decision hyperplane and b determines the offset of the hyperplane to the origin along the direction of $\overrightarrow{w}$. The model estimates these parameters $\overrightarrow{w}$ and b by minimising the average hinge loss plus a term to allow for misclassifications, as can be seen in Equation 4:

$$\mathcal{L}_{hinge} = \left[\frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} - b))\right] + \lambda\|\overrightarrow{w}\|^2 \tag{4}$$

In this equation the λ parameter controls the how far misclassifications are allowed to be from the classification boundary. For values of λ approaching zero, we obtain the hard-margin variant of the SVM. This equation can be solved using quadratic programming by maximising the coefficients c_i as defined in Equation 5.

$$\overrightarrow{w} = \sum_{i=1}^{n} c_i y_i \varphi(\overrightarrow{x_i}) \tag{5}$$

where $\varphi(\overrightarrow{x_i})$ denotes data point $\overrightarrow{x_i}$ transformed into hyperspace by a predefined kernel function.

Applied model. The support vector we applied in our tests uses the Radial Basis Function kernel function to transform data points into a higher dimensional space. This kernel is applied to statistical features computed on the joined output sequences of the OpenPose and DeepFace models. To compute the statistical features from our input sequences, we use the tsfresh python package[4]. From a $Videos \times 7501 \times 52$ input sequence representing containing the output of the DeepFace and OpenPose model for each frame, we compute a $Videos \times 64092$ feature vector. These feature vectors are then standardised by removing the mean and scaling to unit variance before being fed to the support vector classifier for classification. Optimal model parameters were estimated using a gridsearch over the regularization parameter C ([1,1000] using exponential increases) and kernel coefficient γ ([1e-4, 0.1] using 5e-4 increases).

Multi-layer Perceptron

MLPs models are heavily inspired by the biological structure of the brain. In the 1940s McCulloch and Pitts [66] suggested that simple calculation units in the brain called neurons work together and exchange information to make decisions. In MLP models, neurons are organised in layers, where neurons between layers are densely connected with a certain corresponding weight. The model obtains outputs by multiplications of the input with the weight matrixes that connect neurons between layers. However, since the model's output is purely derived from linear combinations of inputs and weights, MLP models often use activation functions on the outputs of each layer to introduce non-linearity. Common choices for activation functions are the Sigmoid, Hyperbolic tangent and Rectified Linear Unit[29] functions.

[4]https://tsfresh.readthedocs.io/en/latest/text/introduction.html

During training the model attempts to find suitable weights for the connections between neurons. This is done using a process called back-propagation in which the model's predictions are iteratively compared to target labels to compute a predefined loss function from which weight gradients can be estimated. Various different methods to obtain this weight gradient exist, among which are Stochastic Gradient Descent, Adam[48] and RMSprop[39]. Due to their ability to learn relations from input to output without the need for manual feature engineering, various works have applied the MLP for affective state estimation from various inputs, such as Electroencephalography (EEG) signals and speech[58, 68, 111]

Mathematical Definitions. We demonstrate the mathematical formulation and calculations for update gradients for the most simple variant of the MLP model, the single-layer perceptron using Gradient Descent. For variants of the MLP model with more layers between input and output, model definition and weight updates resolve around the same mathematical concepts, but require more intermediate computations. For the single-layer perceptron, the relation between input x and output o is denoted by the following formula:

$$o = g(\overrightarrow{w} \cdot \overrightarrow{x} + b) \tag{6}$$

Where g denotes the activation function of the model, $\overrightarrow{w}$ denotes the weight matrix of size $(output\_dim \times input\_dim)$ between the input and output neurons and b denotes learnable bias of size $(output\_dim \times 1)$. During training weight gradients $\Delta\overrightarrow{w}$ and Δb can be computed using the formula defined in Equation 7.

$$\begin{aligned} \Delta\overrightarrow{w} &= \alpha \frac{\partial L(X)}{\partial \overrightarrow{w}} \\ \Delta b &= \alpha \frac{\partial L(X)}{\partial b} \end{aligned} \tag{7}$$

In this formulation α denotes the learning rate, L is the defined loss function and $\frac{\partial L(X)}{\partial z}$ denotes the partial derivative of L with respect to z. Model weights are updated by subtracting the gradients, after which this update procedure repeats for a predefined number of epochs.

$$\begin{aligned} \overrightarrow{w}_{t+1} &= \overrightarrow{w}_t - \Delta\overrightarrow{w}_t \\ b_{t+1} &= b_t - \Delta b_t \end{aligned} \tag{8}$$

Applied model. The model applied in this work is a four layered Multilayer Perceptron for the prediction of self-reports with hidden layer shapes of sizes 100 and 50 respectively. Due to the continuous nature of video data, we tested two different model architectures. In the first model, each time-step of the input is concatenated to form the input to the model, resulting in a $(timesteps * features) \times 1$ input vector. This input is then passed through the hidden layers and output layer to form a global prediction. This model architecture can be seen on the left in Figure 11. Our other approach applies a four-layered MLP to each time-step of the input. In this model, the input is therefore not concatenated, and has a dimension of $timesteps \times features$. This MLP essentially outputs a latent representation per time-step, which are then passed through a max-pooling layer to combine them into a single representation for the complete video. A final single fully connected layer with a softmax activation is used to transform this sequence embedding to a video-level label. This approach is similar to the one applied in [27], where the feature extraction phase is replaced to work the OpenPose and DeepFace model outputs instead of learning from raw frames. The architecture of the second approach can be seen in on the right hand side of Figure 11

Comparisons between these models revealed no significant difference in performance. However, the model using a time-distributed MLP has several benefits over the concatenation model. First of all, the total number of parameters present in the time-distributed model is far less compared to the concatenation model. This is due to the fact that the same MLP model is applied to each time-step, where each time-step has an input dimension of $features \times 1$. This results in a weight matrix of dimension $latent\_dim \times features$ between the input and first hidden layer. In comparison the first weight matrix of the concatenation model requires a weight matrix of size $(latent\_dim \times (timesteps * features))$. Besides this, the second model allows for easier incorporation of temporal supervision, as a temporal prediction head can easily be built on top of the latent temporal representation before this is aggregated with the max-pooling operation into a video-level representation. For these reasons, all experiments are conducted using the time-distributed architecture. The model architecture for including temporal supervision is depicted in Figure 12. As described, it adds a temporal output head to each latent time-step representation before they are aggregated for final prediction. The temporal output branch is denoted by the red "Temporal head" box. During training, both the global label and temporal labels can be fed to the model, as is depicted by the blue boxes.

Gated Recurrent Unit

GRUs models stem from the same core concepts as the MLP model. However, in contrast to the MLP model, these models are allowed to contain cycles between neurons. This allows for the modelling of temporal patterns in the input. Initial simple designs of Recurrent Neural Networks were introduced in the paper of Ruineihart et al. [86]. These models do however contain two key issues regarding the

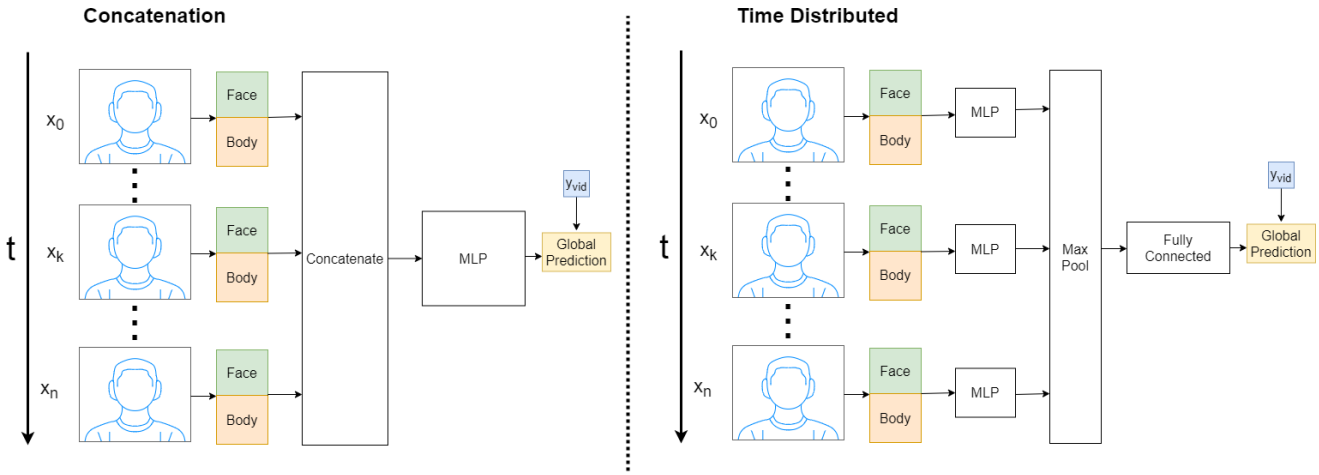


Figure 11: The two different MLP model architectures. The left model utilizes feature concatenation, the right model a temporally distributed MLP model.

weight gradients: the vanishing and exploding gradient problems. For this reason Hochreiter and Schmidhuber [41] introduced the LSTM network, which attempts to solve such issues by the utilisation of gates, which are responsible for controlling what information will be passed on to further timesteps. In 2014, Cho et al. [12] introduced the GRU model, which is closely related to the LSTM model but is computationally more efficient. Empirical results have shown that model performance between the GRU and LSTM models is similar[13, 45].

Mathematical Definitions. GRU models resolve around the concept of a hidden state that propagates over time. In general, this be hidden state can be defined by the following formula:

$$h_t = f(h_{t-1}, x_t) \tag{9}$$

where h_t defines the hidden state at time t, f is a non-linear function that defines what information should be kept over time and x_t is the models input for time t. In a GRU model this function f is defined by means of an update and reset gate applied to the hidden representation h_{t-1} and input x_t. These gates are defined as follows:

$$\begin{aligned} z_t &= \sigma(W^z x_t + U^z h_{t-1}) \\ r_t &= \sigma(W^r x_t + U^r h_{t-1}) \end{aligned} \tag{10}$$

Where z_t and r_t are the update and reset gate values for timestep t, *sigma* denotes the sigmoid function, W^k defines the input weight matrix for gate k and U^k defines the hidden state weight matrix for gate k. With these gate values, we can obtain a memory matrix that captures relevant information from the past timesteps. Using this memory we can define the function f from Equation 9.

$$\begin{aligned} h_t^{'} &= tanh(Wx_t + r_t \circ Uh_{t-1}) \\ f(h_{t-1}, x_t) &= z_t \circ h_{t-1} + (1 - z_t) \circ h_t^{'} \end{aligned} \tag{11}$$

In these equation $h_t^{'}$ denotes the memory content as time t, $tanh$ denotes the Hyperbolic Tangent and $\circ$ denotes the Hadamard product (element-wise product). Similar to the MLP model, weight updates can be obtained by taking the partial derivatives of the loss function with respect to the layer. However as these computations are more involved and quite extensive for the GRU model, they have been omitted from this report.

Applied model. In our models, we stack two recurrent GRU layers on top of each other; one applied on the sequence from past to future, the other applied from future to past. Doing so allows the model to utilise past and future information to create a latent sequence representation by combining the outputs of the two unidirectional models. This latent representation is passed through a single fully-connected layer with softmax activation to obtain the final video-level classification, in similar fashion as the MLP model. Temporal supervision can be added by building a regression head on top of the temporal outputs of the bidirectional GRU model, as denoted by the red "temporal head" boxes in Figure 13. These are omitted for models only predicting video-level labels.

MILBoost

the MILBoost model is a weakly-supervised binary classification model in the Multiple Instance Learning model family.

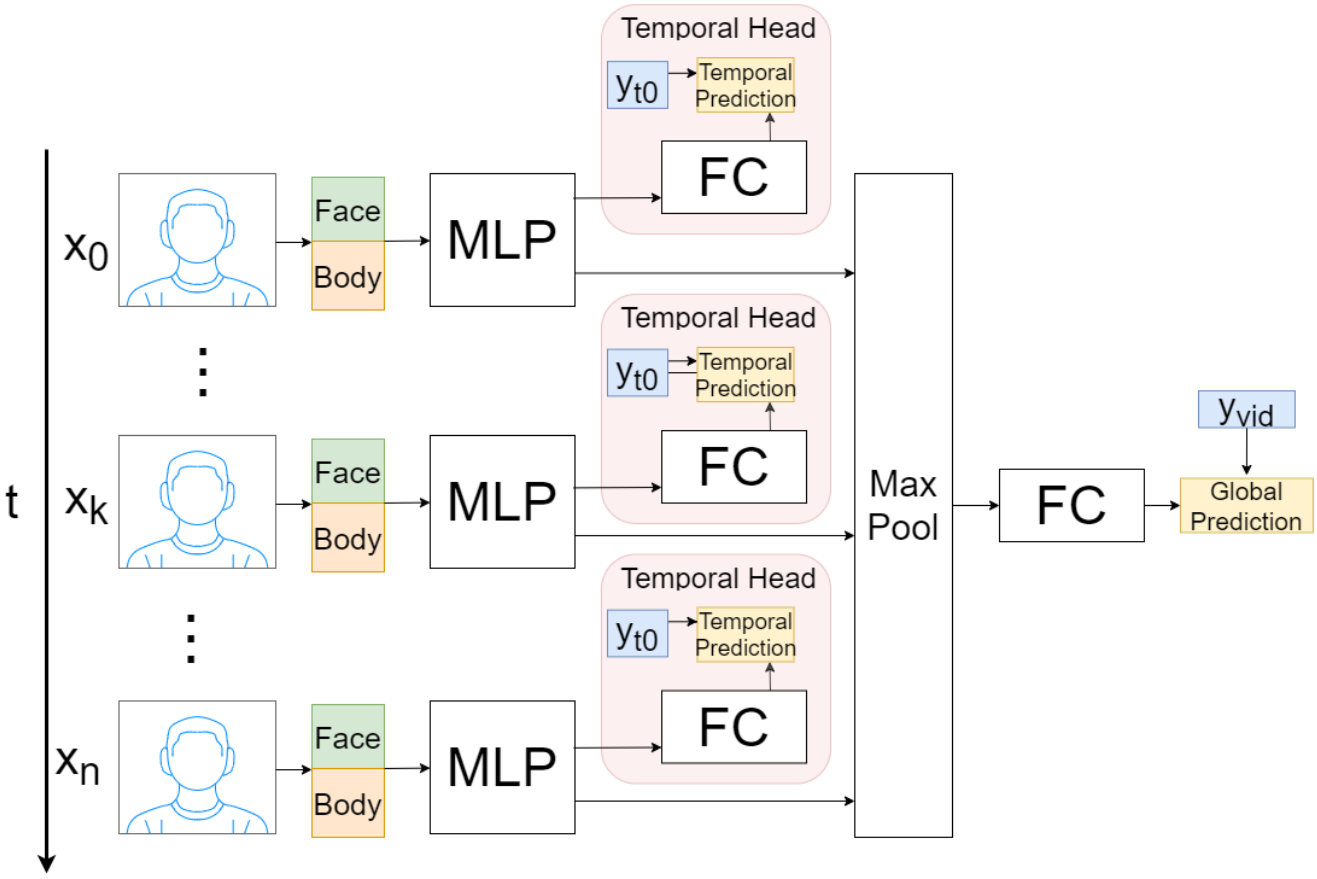


Figure 12: MLP model architecture showing both global and temporal output heads. Temporal heads denoted by red boxes are omitted from models predicting video-level labels.

The MILBoost model was introduced by Viola et al. [103] for image object detection. This model combines the weakly supervised cost functions found in Multiple Instance Learning models with an AnyBoost[63] boosting scheme. This allows the model to predict continuous annotations while being trained on video-level labels. Due to this desirable property, these types of models have been applied in object localisation tasks[14, 103], pain localisation[92] and protein function classification[109].

Multiple instance learning models often define two key concepts: bags and instances. A bag is a collection of instances with an associated label. Instances represent independent model inputs. As an example, in an object localisation task from videos, each instance could represent a video frame, and each bag contains frames associated to a certain video. Positive bags must contain at least one positive instance, whereas negative bags can contain no positive instances.

Mathematical Definitions. As mentioned above, the MILboost model predicts an output for each instance in the bag and assigns bags as positive if it contains one ore more positive instances. During training the model obtains a decision boundary by minimising the Negative log-likelihood loss function, as defined in Equation 12:

$$\begin{aligned} \mathcal{L} &= -\sum_{i}^{N} t_i log(p_i) + (1 - t_i) log(1 - p_i) \\ p_i &= \max_j (p_{ij}) \\ p_{ij} &= \frac{1}{1 + exp(C(x_{ij}))} \end{aligned} \tag{12}$$

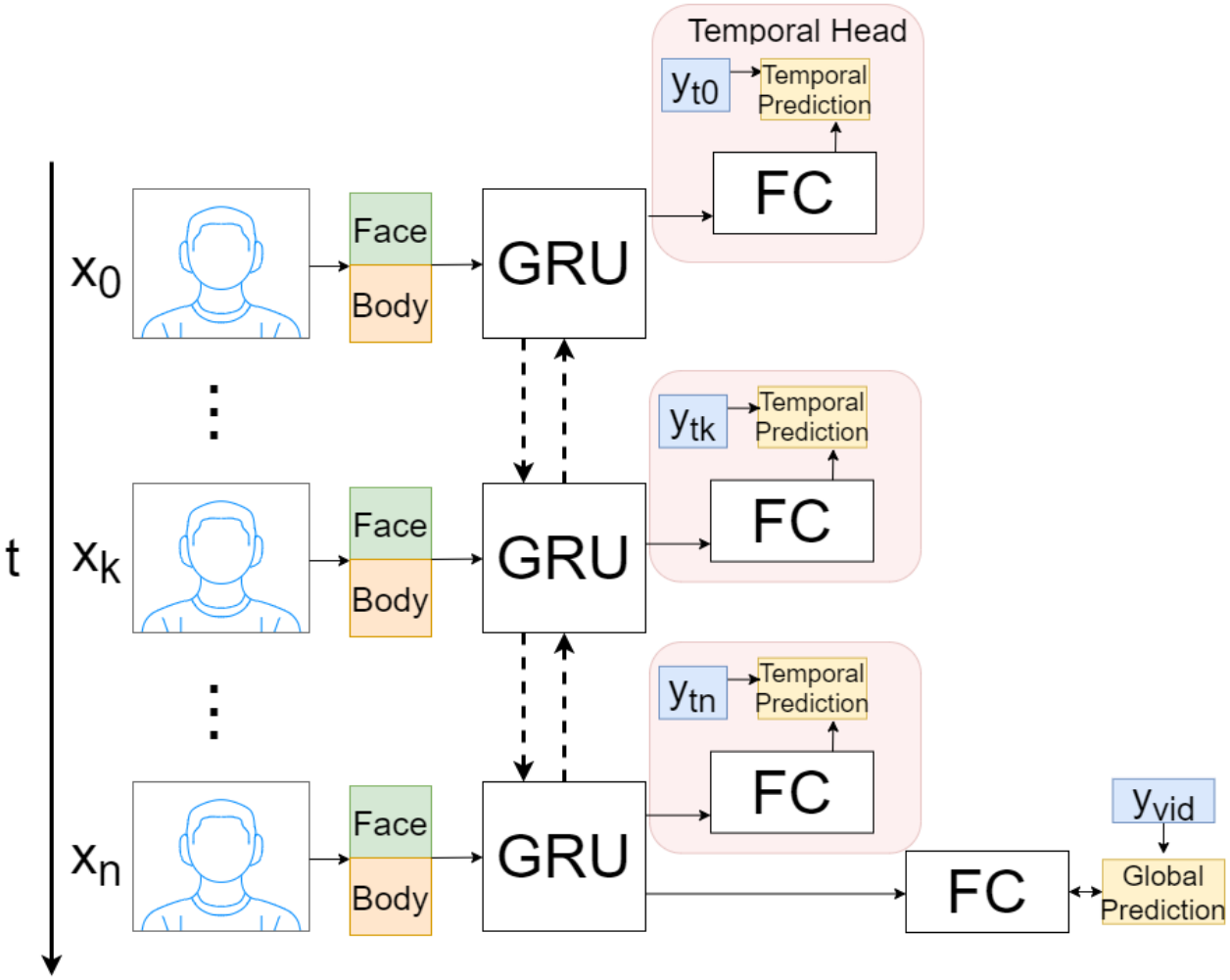


Figure 13: GRU model architecture showing both global and temporal output heads. Temporal heads denoted by red boxes are omitted from models predicting video-level labels.

In these equations, N is the number of bags or videos, t_i is the global video label and p_{ij} is the probability of instance j belonging to bag i, which is determined by the output of classifier C given input x_{ij}.

Using this loss function, we can obtain weights for each instance in a bag using the derivative of the loss with respect to the predicted instance label:

$$w_{ij} = \frac{\partial \mathcal{L}(C)}{\partial C(x_{ij})} \tag{13}$$

Using these weights we can perform a boosting iteration. In each iteration of boosting, a weak decision stump is added to the model ensemble. This decision stump is determined by the classifier that maximises the following equation:

$$c_t = \max_c \left[\sum_{ij} c(x_{ij}) w_{ij} \right] \tag{14}$$

The classifier's weight λ is determined by a line search maximising Equation 15.

$$log\mathcal{L}(C + \lambda_t c_t) \tag{15}$$

Applied model. In our models we use an ensemble of MIL-Boost models to obtain multi-class classification. These models are trained in an one-versus-all fashion, yielding three separate models for the three respective levels of Arousal. After training each model as described above, the bag-level

label is determined by the model with the highest class probability, that is:

$$p_i = \max_m \max_j (p_{ijm}) \tag{16}$$

where m is a single model in the ensemble and p_{ij} is the probability of instance i belonging to bag j predicted by model m. Similarly, temporal instance labels are assigned by taking the maximum positive instance probability across models:

$$p_{ij} = \max_m p_{ijm} \tag{17}$$

Models were trained using Decision Tree classifiers as the weak decision stump and a boosting scheme of 50 iterations. A visual representation of the MILBoost model can be found in Figure 14.

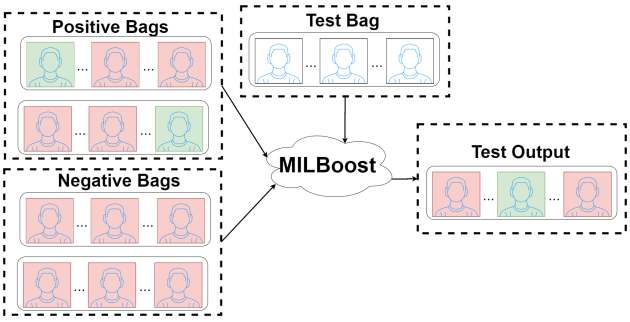


Figure 14: Visual description of the MILBoost model. Positive bags contain at least on positive instance, negative bags contain no positive instances..

Action Localisation Net

The Action Localisation model is another weakly supervised model, but applies a deep learning approach instead of using Multiple Instance Learning techniques. Similar to the MLP and GRU networks, these models are based on connected simple computational neurons. First introduced by Wang et al. [106], these models attempt to learn to capture temporal relevant segments from a single global label. The proposed model architecture in Wang et al. [106] heavily resembles the TSN network's architecture as proposed in [107], but does not require temporal annotations during training. This Action Localisation model computes an activation map over the spatial or temporal dimension, effectively assigning a relevance label for each instance in that dimension. This activation map is combined with the output of a classification head to form the predicted global label for the video or image. These models form a video-level classification based on the weighted sum of all instances in the relevant dimension, contrary to MIL models where only the entry with the maximum activation is considered.

Mathematical Definitions. As the applied model is based on a combination of GRU and MLP layers, the math behind these models has been omitted from this subsection. Similar to these models, training the Action Localisation model is performed by computing weight gradients using partial derivatives of the loss function with respect to a specific layer.

Applied model. The utilised Action Localisation model is an adapted version from the model proposed in [106]. Specifically, in their model the authors differentiate between multiple phases: sampling, feature extraction, classification and selection. In our model, we adapted the feature extraction phase to work with the output of the pretrained DeepFace and OpenPose model outputs instead of learning from raw frames. Furthermore, we modified the classification phase to support training as a regression problem instead. Figure 16 shows the architecture of our proposed model.

Clip Sampling This model relies on the sampling of clips during training of the model. This means that the model only sees small snippets of the complete video during training. To obtain these snippets, we first split our videos into evenly sized segments of 10 seconds. Then, for each batch a random subset of segments ($nr\_segments = 7$ in our case) are sampled from the total pool of segments. For each of these segments, a predefined number of shots ($nr\_shots = 15$) are sampled at random start times. These shots contain small sequences of features stemming from five consecutive frames. Doing so leaves us with an input dimension of $nr\_segments \times nr\_shots \times seq\_size \times feature\_dim$. Figure 15 shows an example of how clip sampling is performed during training.
During testing, we sample a shot every second in the input video, filling the segments with consecutive shots. Similar to [106], the recognition scores of segments are aggregated with a weighted sum to yield the final video-level prediction. The weight of each segment for final prediction is determined by multiplying its attention score with the output of the regression head.

Feature Extraction Instead of the temporal and spatial CNNs that Wang et al. [106] apply for feature extraction for each shot, we apply an LSTM over the sequence of features within a shot to obtain a representation for that sequence. These features are the outputs of the DeepFace and OpenPose models stemming from the selected frames in a shot. To obtain a representation per shot, we perform an Average Pooling operation over the sequence dimension. This representation is then fed to the regression and selection heads.

Model prediction Given this representation for each shot, the network performs two operations to obtain video-level

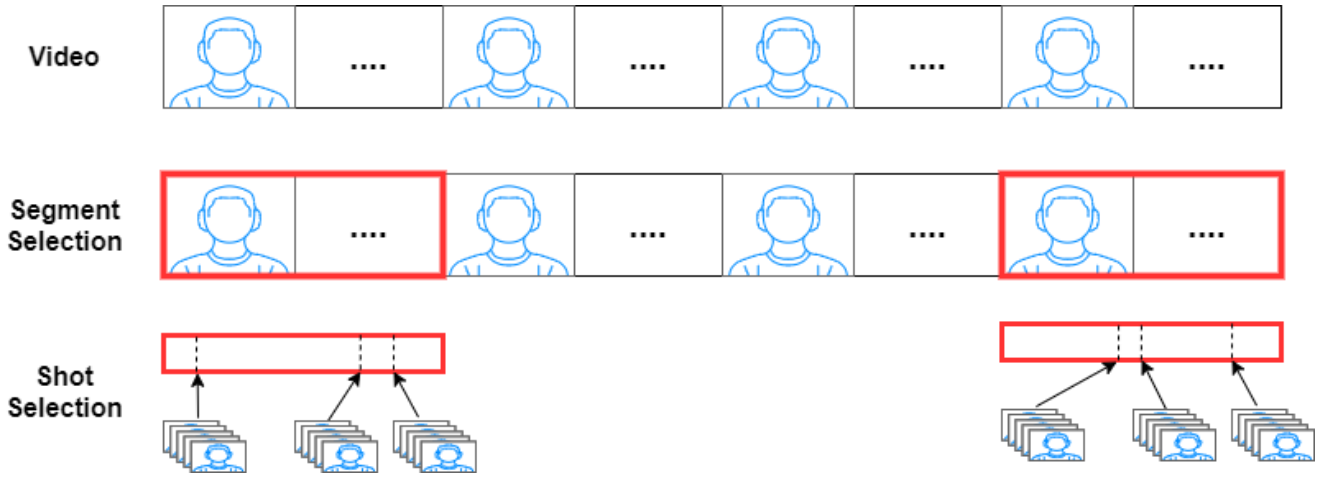


Figure 15: Shot sampling procedure during training of the Action Localisation model. An input video is segmented into evenly sized segments. A random subset of these segments is selected in each batch. From each segment, shots with random start times are selected.

predictions. First, the model computes a regression score per shot by applying a deep layer to each shot. These regression scores per shot are averaged to obtain a segment level regression score. Second, the model computes an attention or relevance score per segment by applying a different deep layer to all individual shots. Again, the scores for each shot are averaged to obtain segment-level scores. However, in this model head, a *softmax* operation over the segment dimension is applied to weigh each segment relative to other segments in the video. This weight can be seen as a relative importance per segment in the video. Finally, to obtain video-level predictions, the regression scores per segment and their relative importance are multiplied to obtain a weighted score per segment. These are then summed to obtain video-level classification scores. To obtain temporal predictions, the weighted score per segment is scaled to obtain values in the same value domain as the video-level prediction:

$$y_i = \left[\frac{y_i - \max \hat{yt}}{\sigma_{\hat{yt}}} * \sigma_{yt_{train}} \right] + \hat{y}_{vid} \tag{18}$$

In this equation, y_i is the predicted value at time i, $\hat{yt}$ is the prediction sequence for all timesteps, $\sigma_{\hat{yt}}$ denotes the standard deviation for the full prediction sequence. $\sigma_{yt_{train}}$ denotes the average standard deviation of the true annotation sequences in the training set and $\hat{y}_{vid}$ denotes the video-level prediction of the model. This transformation can be seen as a normalisation procedure, where we scale the maximum value of the prediction sequence to the predicted video-level label. We multiply by the standard deviation of the training annotation sequences to account for the difference in scales, as we expect that the standard deviations of test sequences will roughly approximate the same standard deviation.

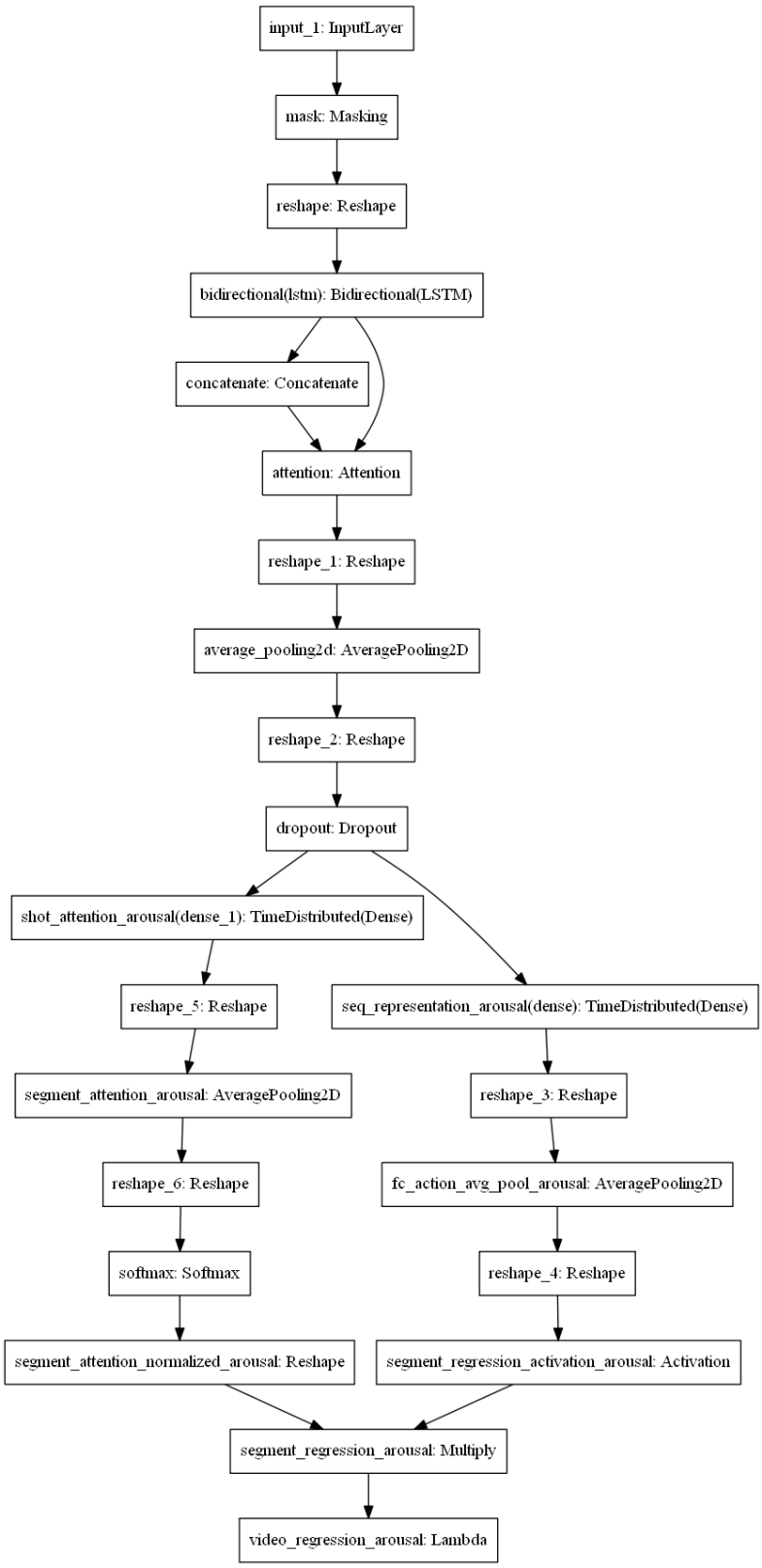


Figure 16: Model architecture of the Action Localisation network.

B DATA LOADING ON EXTERNAL MACHINES

Data containing recordings of human behaviour is privacy sensitive, and does often require users of that data to sign an End-user Licence Agreement (EULA). In this EULA, one specifies who will be allowed access to the data and on what terms data can be used. With the rise of deep learning and increasingly complex models, there is often a need for training in a distributed setting as training on a single host would be infeasible due to time or compute constraints. However, these compute clusters are inherently shared among users, therefore requiring additional thought on how to prevent unauthorised data access when processing sensitive data on shared networks. To this end, we propose a containerised system capable of running in a distributed setting, which minimises the risk of unauthorised data access.

We have posed the following requirements for this system:

(1) Data can not be accessed by other users of the cluster
(2) A data point will only reside on the shared compute node during the processing of that point
(3) Data will be inaccessible upon completion or failure of the processing software
(4) System does not require elevated privileges

With these requirements in mind, we propose the following method for data processing in shared compute environments. The system will be based on Singularity containers which will be deployed on each requested compute node in the cluster. These Singularity containers contain the processing code as well as a startup script that will create a mount to a protected data server. Once these containers are deployed on the node, it essentially creates an overlay of its internal file system on the host machine, and shares the users privileges. By doing so, the mount point is bound to system memory, which could allow for potential access from external users using that compute node. For this reason, we use SSHFS to create a mount point as this allows us to bind a specific user to that mount point, making it inaccessible for all others, including users with administrator rights[5]. This approach fulfils our first requirement.

After data is mounted on the host, data files are not yet pulled to the compute node; Only a list of files is available inside the Singularity container. Once the code starts processing a a data point, that particular data point is fetched from the external server and placed in the containers cache. Once the code progresses to a new data point, the cache is cleared, removing the old data point from the machine's memory. In this way data only resides on the machine during the actual processing of that data point, fulfilling requirement 2.

Upon completion of the processing code, the last data is cleared from cache, the mount point released and the container is terminated. However, to prevent data remaining in the machine's cache upon crashes or other unexpected termination of the processing code, we bind a handler to the data processing code, responding to various system signals, such as EXIT, SIGQUIT, SIGABRT and SIGTERM signals. These signals are raised by the system whenever the code exits, quits, aborts or terminates. Whenever one of such signals is received, the same procedure as upon successful completion is applied. Therefore, crashes of the code inside of the container will not result in remaining data on the host. Another possibility is a full crash or unexpected termination of the Singularity container. This could for example occur when cluster schedulers kill the container for trying to use more resources than requested. However, due to the containerised nature of our approach and the fact that the mount point is bound inside the container, crashes of the container will make the mount point unavailable and inaccessible. These precautions make sure that data will never be present on the compute nodes upon completion or failure of the processing code or container, as required by requirement 3.

Lastly, it is important that users do not require elevated privileges to use this data handling procedure, as this is generally not allowed by the compute cluster. Singularity and other containerised approaches are specifically designed with this requirement in mind and can therefore be ran with standard user permissions. Allowing for FUSE mounts within an unprivileged user name-space has been made possible since Linux kernel 4.18. This kernel dependency and the need for a Singularity installation are the only system requirement imposed by this approach on the compute cluster, there is no need for specialised user privileges, as specified by requirement 4.

This approach does therefore fulfil all set requirements to protect unauthorised data access when running on uncontrolled hosts. A diagram of the proposed method can be found in Figure 17. However, applying this method comes with some drawbacks that might hinder usability. First of all, due to the constraints on data availability on the compute node, data has to be fetched from a remote server using the SSH File Transfer Protocol (SFTP) protocol. Especially if one is working with a large number files with large sizes, this approach significantly slows down processing speeds, as downloading these files to the compute node can take some time. Second, batch processing comes at a cost of requiring more data to simultaneously exist on the compute node. However, especially for training deep models, such

[5]Documentation of FUSE access permissions: https://man7.org/linux/man-pages/man8/mount.fuse.8.html#OPTIONS

behaviour is desired to stabilise gradients and to optimise prediction speeds. Lastly, this approach can not utilise efficient cluster computations such as map-reduce, in which data is split and divided over multiple nodes, as the data is not present on a cluster level, only nodes can access the data from within their singularity containers. Summarising, this approach ensures data protection, but comes at the cost of reduced processing efficiency. Although some of these drawbacks might be optimised by relaxing some of the constraints of the system, this was out of scope for this project and has not been researched.

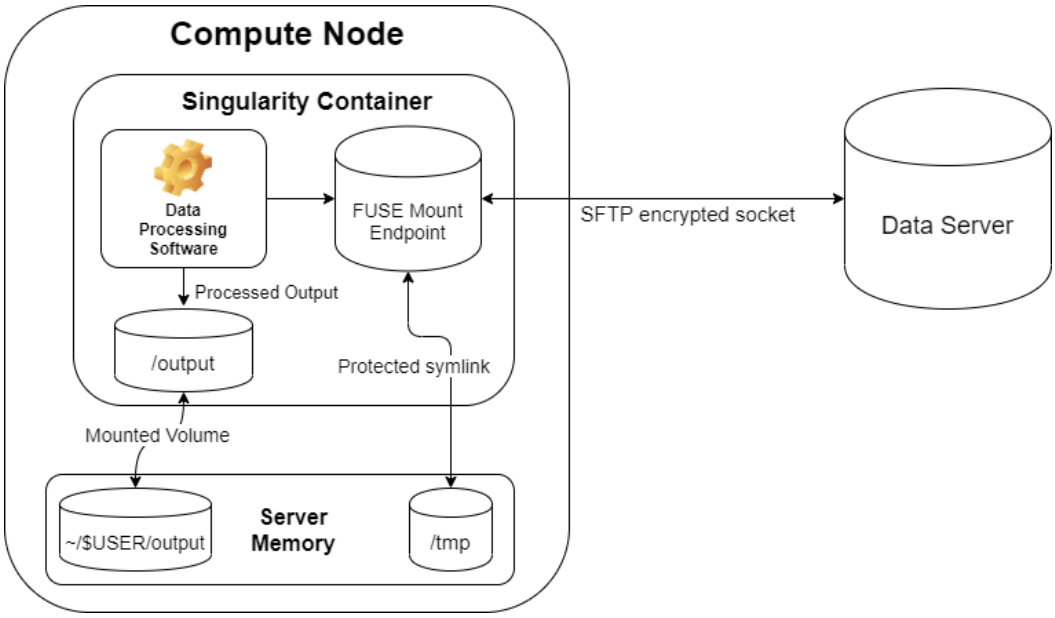


Figure 17: Data loading procedure.

C SELF-REPORT DISTRIBUTIONS

This section shows the data distributions for of Arousal and Valence for the RECOLA and Mementos datasets. We show the original distribution as well as the obtained distribution after our binning operation. Distributions for RECOLA can be found in Figure 18 and 19, distributions for Mementos in Figure 20 and 21.

RECOLA

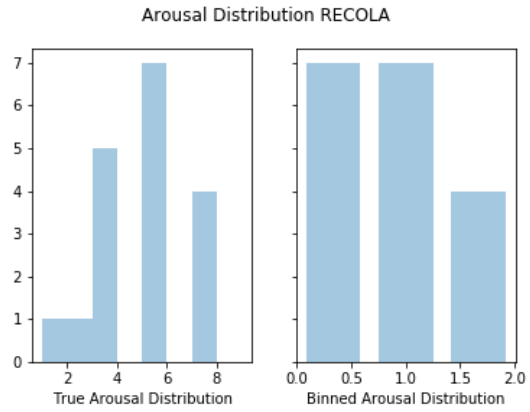


Figure 18: Distributions of self-reported Arousal of the RECOLA dataset before and after binning.

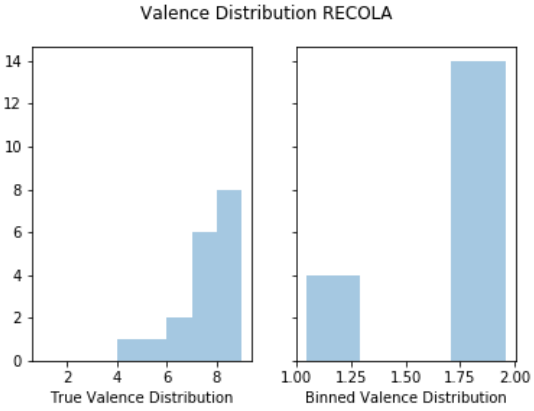


Figure 19: Distributions of self-reported Valence of the RECOLA dataset before and after binning.

As can be observed from the distribution of Valence, self-reports are severely skewed. As binning would result in bins containing 2, 0 and 16 samples respectively we decided to omit experiments for the prediction of Valence due to insufficient class variance.

Mementos

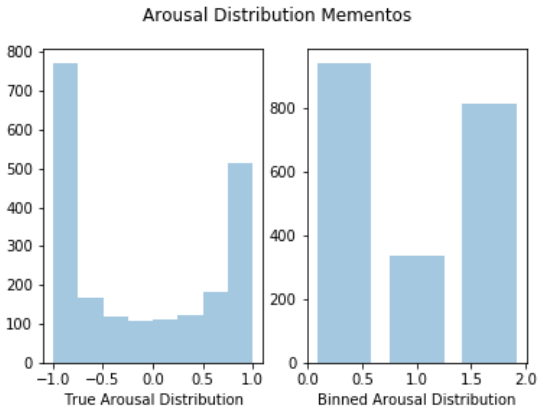


Figure 20: Distributions of self-reported Arousal of the Mementos dataset before and after binning.

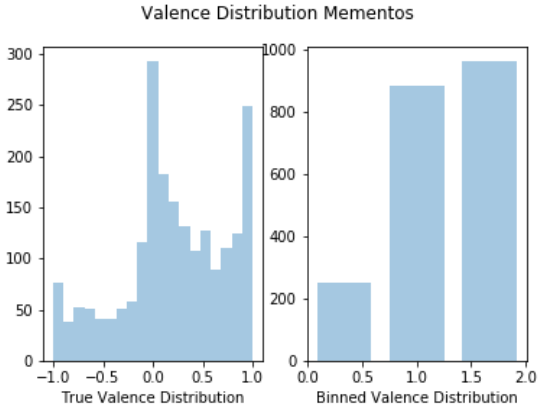


Figure 21: Distributions of self-reported Valence of the Mementos dataset before and after binning.

ACRONYMS

AU Action Unit. 3, 5, 7

CCC Concordance Correlation Coefficient. 19, 20
CNN Convolutional Neural Network. 5, 34

ECG Electro-Cardiogram. 6
EDA Electro-Dermal Activity. 6
EEG Electroencephalography. 30
EULA End-user Licence Agreement. 37

FACS Facial Action Coding System. 3, 5

GRU Gated Recurrent Unit. 10–12, 14–19, 21, 23, 30, 31, 33, 34

GSR Galvanic Skin Response. 5

LSTM Long short-term memory. 5, 11, 31, 34

MIL Multiple Instance Learning. 6, 11, 18, 34
MLP Multilayer Perceptron. 9–19, 29–32, 34
MSE Mean Squared Error. 11, 12

RBF Radial Basis Function. 29
ReLU Rectified Linear Unit. 29
RNN Recurrent Neural Network. 11, 30

SFTP SSH File Transfer Protocol. 37
SVM Support Vector Machine. 5, 9, 11, 12, 29

VA Valence-Arousal. 6