sensors

MDPI

Article

CorrNet: Fine-Grained Emotion Recognition for Video Watching Using Wearable Physiological Sensors

Tianyi Zhang [1,2,*], Abdallah El Ali [2], Chen Wang [3], Alan Hanjalic [1] and Pablo Cesar [1,2,*]

[1] Multimedia Computing Group, Delft University of Technology, 2600AA Delft, The Netherlands; A.Hanjalic@tudelft.nl
[2] Centrum Wiskunde & Informatica (CWI), 1098XG Amsterdam, The Netherlands; abdallah.el.ali@cwi.nl
[3] Future Media and Convergence Institute, Xinhuanet & State Key Laboratory of Media Convergence Production Technology and Systems, Xinhua News Agency, Beijing 100000, China; wangchen@news.cn
* Correspondence: T.Zhang-5@tudelft.nl (T.Z.); P.S.CesarGarcia@tudelft.nl (P.C.)

Abstract: Recognizing user emotions while they watch short-form videos anytime and anywhere is essential for facilitating video content customization and personalization. However, most works either classify a single emotion per video stimuli, or are restricted to static, desktop environments. To address this, we propose a correlation-based emotion recognition algorithm (CorrNet) to recognize the valence and arousal (V-A) of each instance (fine-grained segment of signals) using only wearable, physiological signals (e.g., electrodermal activity, heart rate). CorrNet takes advantage of features both inside each instance (intra-modality features) and between different instances for the same video stimuli (correlation-based features). We first test our approach on an indoor-desktop affect dataset (CASE), and thereafter on an outdoor-mobile affect dataset (MERCA) which we collected using a smart wristband and wearable eyetracker. Results show that for subject-independent binary classification (high-low), CorrNet yields promising recognition accuracies: 76.37% and 74.03% for V-A on CASE, and 70.29% and 68.15% for V-A on MERCA. Our findings show: (1) instance segment lengths between 1–4 s result in highest recognition accuracies (2) accuracies between laboratory-grade and wearable sensors are comparable, even under low sampling rates (≤64 Hz) (3) large amounts of neutral V-A labels, an artifact of continuous affect annotation, result in varied recognition performance.

Keywords: emotion recognition; video; physiological signals; machine learning

Citation: Zhang, T.; Ali, A.E.; Wang, C.; Hanjalic, A.; Cesar, P. CorrNet: Fine-Grained Emotion Recognition for Video Watching Using Wearable Physiological Sensors? *Sensors* **2021**, *21*, 52. https://dx.doi.org/10.3390/s21010052

Received: 3 December 2020
Accepted: 21 December 2020
Published: 24 December 2020

1. Introduction

Emotions play an important role in users' selection and consumption of video content [1]. Recognizing the emotions of users while they watch videos freely in indoor and outdoor environments can enable customization and personalization of video content [2,3]. Although previous work has focused on emotion recognition for video watching, they are typically restricted to static, desktop environments [1,4,5], and focus on recognizing one emotion per video stimuli [6–8]. For the latter case, such emotion recognition is temporally imprecise since it does not capture the time-varying nature of human emotions [9,10]: users can have and report multiple emotions while watching a single video. Here, we define fine-grained emotion recognition as recognizing the temporal moment-by-moment valence and arousal [11,12] states, typically in segments of 0.5 s to 4 s depending on the duration of an emotion [13,14]. This is in contrast to emotion recognition per video [8,15]. In this work, we draw on dimensional models of emotion (cf., Russell's Circumplex Model of Emotions [12]), which describe emotions using a multi-dimensional space. Compared with discrete models (e.g., Self-Assessment Manikin (SAM) [16]), these have a finer level of granularity by introducing continuous variables, namely valence and arousal, to describe emotions [6].

While there has been research on fine-grained, temporally precise emotion recognition (cf., FEELtrace [17], DARMA [18], CASE [19]), these methods either require users to wear

or attach obtrusive sensors [20–22] (e.g., Electroencephalograph (EEG)), or rely on facial expression sensing [20,21,23,24] for fine-grained emotion recognition. With respect to EEG, emotion recognition accuracies up to 80% have been achieved over the past decade [25]. However, high resolution EEG signals need to be captured under strict laboratory environments without any electromagnetic interference [26], which makes their use limited to outdoor settings. Furthermore, EEG sensors can be obtrusive since electrodes need to be attached to a user's head during acquistion. Camera-based sensing, while less obtrusive, is not always possible in different scenarios. For example, in mobile settings, the front camera may potentially be used to unobtrusively collect facial expressions. However, the front camera cannot always capture the whole face of the user [27]. In addition, constant streaming of facial images can bring privacy concerns for both the user who watches videos and other persons whose faces may be captured in the context environment [28,29].

Unlike facial expressions, physiological signals (e.g., Heart Rate (HR), Blood Volume Pulse (BVP), Skin Temperature (ST), and Electrodermal Activity (EDA)) are largely involuntarily activated (i.e., spontaneous and not controllable), which enable a more objective means to measure affective reactions (i.e., valence and arousal) [6]. Furthermore, physiological signals can be measured using wearable sensing devices. With the proliferation of wearable physiological sensing devices (e.g., smartwatches and wristbands) that can measure signals such as HR or EDA, they have become easily accessible and widespread in daily life use [30,31]. Given the foregoing, we focus on fine-grained emotion recognition using wearable physiological sensors. To this end, we collected the Mobile Emotion Recognition with Continuous Annotation (MERCA) dataset, where users annotate their valence and arousal states using a continuous mobile annotation input technique (cf., [32]) in real-time while watching short-form videos.

Fine-grained emotion recognition needs to segment continuous signals into smaller (fine-grained) instances and recognize the emotions they represent. A major challenge for recognition is that the information inside each segment of the signals (i.e., instances) may not be sufficient for recognizing emotions. In previous works [21,33,34], sequence learning methods such as Long Short-Term Memory (LSTM) [35] networks have been used to extract the temporal information between different samples or instances as additional features for recognition. However, the temporal information extracted by sequence learning methods is based on the fine-grained emotion self-report annotated by users. Such reports may not be precise enough, be misaligned temporally to the actual state at which they were experienced, or be altogether inaccurate. If the network is trained with these labels, training error could accumulate and affect the recognition result for other instances within the same signal [36,37].

To address this challenge, this paper presents a fine-grained emotion recognition algorithm, CorrNet, which uses unsupervised learning to learn the features both inside and between different instances, and a supervised classifier to recognize the emotions for each of them. CorrNet takes advantages of the features both inside and between instances by extracting correlation-based features for all instances for the same video stimuli. Our work offers two primary contributions:

(1) We propose a novel emotion recognition algorithm to classify the valence and arousal in finer granularity using wearable physiological sensors. The proposed algorithm is tested both on an indoor-desktop dataset (CASE [19]), and on an outdoor-mobile dataset (MERCA), which we collected using wearable physiological sensors while users watched short-form (<10 min) [38] mobile videos. Results show good performance for binary valence-arousal (V-A) classification on both datasets (76.37% and 74.03% of V-A on CASE; 70.29% and 68.15% for V-A on MERCA), respectively. Our results outperform other state-of-the-art baseline methods for emotion recognition, including classic ML-based support vector machines (SVMs) and sequential learning approaches such as Long Short-Term Memory (LSTM) networks.
(2) We compare the performance of CorrNet through testing experiments with different parameters (e.g., different lengths of instances and different sampling rates) and

discuss how they could affect the recognition results. The discussion provides insight into how to design a fine-grained emotion recognition algorithm using segmented physiological signals. Our discussion also shows high recognition accuracy can be achieved using wearable physiological signals with low sampling rate ($\leq$64 Hz), which means lower power consumption and easier sensor deployment (e.g., do not need to stick electrodes on users' skin) compared with laboratory-grade sensors with higher sampling rate ($\geq$1000 Hz).

2. Related Work

In this section, we first introduce the existing models to quantify emotions. Then, we review the wearable physiological signals and existing algorithms for recognizing emotions and narrow our scope into specific techniques for recognizing fine-grained emotions.

2.1. Discrete vs. Dimensional Emotion Models

Emotions have been widely studied in psychology and neuroscience [39]. A variety of models have been proposed to measure and quantify emotions, which can be divided into two categories [6]: categorical and dimensional emotion models. Categorical emotion models divide emotions into different categories and describe them using emotion keywords. For example, the classic six-basic-emotion model by Ekman [40] summarized happy, sad, anger, fear, surprise, and disgust as six basic emotions, and viewed other emotions as combinations of these basic ones. Researchers also use categorical emotion models to quantify specific emotions such as frustration [41], stress [42,43], social anxiety, and depression [44]. Dimensional emotion models by contrast quantify emotions using a multi-dimensional space. Compared with categorical emotion models, dimensional emotion models can describe emotion on a finer level of granularity by using continuous values to model emotions. These models, typically Russell's Circumplex Model of Emotions [12] which describe emotions using valence and arousal, are widely used for fine-grained or continuous emotion recognition [17,19].

Our work aims to recognize emotions in fine granularity. The emotion model we use should be able to show the dynamic and continuous changes of users' emotion, therefore in this work we use dimensional models (i.e., valence and arousal) to model emotions.

2.2. Wearable Physiological Sensing for Emotion Recognition

Physiological signals collected from wearable sensors are widely used for recognizing emotions outside a laboratory environment [45–49]. For example, Costa et al. [45] developed an ambient intelligent system to recognize valence and arousal using Electrocardiogram (ECG), Blood Volume Pulse (BVP) and Electrodermal Activity (EDA) from iGenda, a smart wristband. Alexandros et al. [46] proposed a recognition system, HealthyOffice, to recognize stress, anxiety and depression in the workplace using ECG and BVP using a wristband and a mobile phone. Compared with signals which indicate the cognitive activities from the Central Nervous System (CNS), the signals which interpret the physiological behaviors in the Autonomic Nervous System (ANS) are easier to obtain using wearable sensors. For example, many commercialized smart watches and wristbands (e.g., Empatica E4 wristband and Toshiba W110 wristbands [50]) have integrated photoplethysmogram (PPG) and skin conductance (SC) sensors to measure Heart Rate (HR) and EDA. Recent studies have drawn on these signals to ubiquitously measure user experience, such as user engagement of mobile games [51], synchrony between presenters and audience members [49], and students' emotional engagement during lectures [52]. However, the signals measuring signals in the ANS (normally single channel) are less information rich than Electroencephalogram (EEG) signals (normally 16–32 channels). This brings up challenges of how to design algorithms that ensure robust and accurate emotion recognition.

Our work aims to develop emotion recognition algorithms for video watching that are not limited to laboratory and indoor environments. Following prior work [45,46,49,51,52],

we narrow our scope on using physiological signals such as ECG, BVP, EDA and HR from wearable sensors.

2.3. Emotion Recognition Algorithms Using Physiological Signals

Algorithms for recognizing emotions using physiological signals can be divided into two major categories: model specific methods and model free methods [6]. Model specific methods require carefully hand-crafted features to classify emotions from physiological signals. In general, statistical features from the time-domain (e.g., mean, standard deviation, first differential [53–55] of the signal) and frequency-domain (e.g., mean of amplitude, mean of absolute value [52,56], or signal FFT [51]) are commonly used. Features are selected or designed by researchers thus they do not depend on the emotion ground truth labeled by users. However, there is no consensus of which features are the most reliable for recognizing emotions [6,57]. Therefore researchers have to carefully design features according to the data they collected, limiting the generalizability of their algorithms. The extracted features are then input into classifiers such as Support Vector Machine (SVM) [58], K-Nearest Neighbor (KNN) [59], or Random Forest (RF) [60] to classify emotions. Since the model specific methods require researchers to select features based on empirical experiments, it is costly with respect to time and does not guarantee that selected features are optimized [6,7].

Model free methods on the other hand use neural networks to learn the inherent structure behind the data and automatically extract features for recognition. Deep learning networks such as convolutional neural networks (CNNs) [61,62] and Long Short-Term Memory (LSTM) networks [33,34] are commonly used and achieve high accuracy. For example, Ma et al. implemented [33] a multimodal residual LSTM network to classify valence and arousal and obtained a classification accuracy of 92.87% and 92.30% for arousal and valence, respectively. According to the research from Suhara et al. [63], LSTM networks could outperform classic machine learning algorithms such as Support Vector Machines (SVMs) for forecasting emotion states. Although model free methods achieve high recognition accuracy, they easily overfit on the training data when the ground truth labels are not accurate [64]. This appears to be a common phenomenon when users label their emotions [19,65].

Our work attempts to draw on the advantages of both model specific and model free methods by using unsupervised learning techniques to automatically extract features and supervised learning techniques to classify emotions.

2.4. Fine-Grained Emotion Recognition

While there exists many algorithms that are designed for recognizing emotions based on physiological signals, techniques for fine-grained emotion recognition are still in their infancy [22]. Fine-grained emotion recognition requires algorithms to output multiple emotion states by relying on signals within one certain time interval. For temporal signals, this is normally done using two kinds of methods:

The first kind of methods views the target emotion states as a continuous sequence and directly calculate the mapping (regression) from input signals to output emotion sequences. These methods include sequential learning approaches such as LSTM [33,34], and temporal regression such as support vector regression (SVR) [66,67] and polynomial regression [68]. While previous work has shown that regression approaches, especially sequential learning using recurrent structures can achieve high accuracy [10,20,69], these methods are sensitive to the accuracy of the ground truth. Since the recurrent structure is trained from the beginning to the end of the signal, the regression error from the first few samples could be accumulated and affect the results of the whole sequence.

The second kind of methods segments continuous signals into different fine-grained instances and classifies the emotion of each instance independently. Therefore, the recognition result of different instances will not affect each other. For example, Romeo et al. [70] designed an SVM-based multi-instance learning algorithm to recognize valence and arousal for each fine-grained instance and achieves 68% of accuracy on high arousal. These kinds

of methods are also widely used for fine-grained emotion recognition with different data modalities such as facial expressions [71] and vocal features [72] (e.g., pitch and loudness). The main challenge for this kind of methods is to extract and fuse both the features inside and between instances, as the information which resides only within instances may not be enough to determine which emotion it represents. Previous works [70,73] use the joint loss [74] of instances and bags (instances under one video stimuli) to fuse the features inside and between instances. However, it could lead to temporal ambiguity of emotions as instances are not directly trained by their emotion labels (and instead trained by the label of bags) [70].

In our work, we draw on the second kind of methods (due to imprecision of fine-grained emotion ground truth from self-reports), and aim to extract and fuse the information within and between instances without compromising the link between instances and their emotion labels.

3. Methodology

In this section, a correlation-based emotion recognition algorithm (CorrNet) is proposed to classify fine-grained emotion states (i.e., valence and arousal (V-A)) from physiological signals. The procedure of the proposed algorithm is illustrated in Figure 1. CorrNet contains three stages: (1) Intra-modality feature learning: the obtained physiological signals are firstly grouped into two modalities (signals from two different nerve systems, e.g., oculomotor nerve system and autonomic nervous system). At the first stage, original signals are projected into a low dimensional latency space where intra-modal features are learned using a convolutional auto-encoder. After that, the feature vectors from the latency space are grouped according to the video stimulus the users watched. (2) Correlation-based feature extraction: In the second stage, the cross-modal features are obtained through correlation-based feature extraction. (3) Broad Learning System classification: At the last stage, the extracted features are inputted into a broad learning system (BLS) to classify valence and arousal for each instance. Each stage is discussed below, and the pseudocode of CorrNet is shown in Algorithm 1.

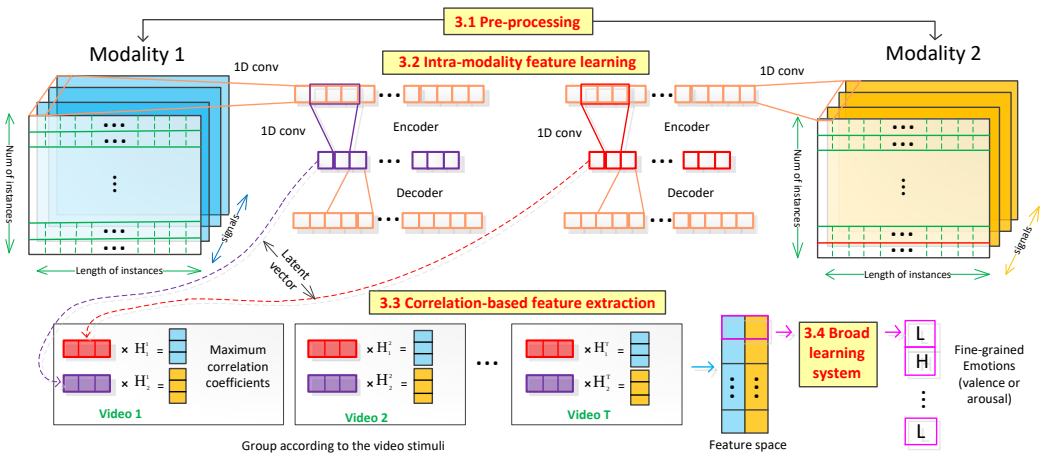

Figure 1. The procedure of proposed CorrNet.

Algorithm 1 CorrNet

Input: Training set with n instances in modality 1: $X_1 = \{x_1^i\}_{i=1}^n, x_1^i \in R^{L \times C_1}$ and modality 2: $X_2 = \{x_2^i\}_{i=1}^n, x_2^i \in R^{L \times C_2}$

Output: Fine-grained emotion labels (i.e., valence:$V_a = \{v_i\}_{i=1}^n$ and arousal $A_r = \{a_i\}_{i=1}^n$)

for j = 1 and 2 **do**
  Encoder $\rightarrow \phi_j = X_j \otimes \psi(\omega, c)$
  Decoder $\rightarrow \eta_j = \phi \bar{\otimes} \psi'(C_j, c)$
end for
Group instances according to video stimulus:
for t in **T** = number of video stimulus **do**
  $(H_1^t, H_2^t) = \mathbf{CCA}(\psi_1^t, \psi_2^t)$
  $F^t = [\psi_1^t \cdot H_1^t, \psi_2^t \cdot H_2^t]$
end for
$F = \{F^t\}_{t=1}^T, F \in R^{n \times k}$
$(a_i, v_i)_{i=1}^n = \mathbf{BLS}(F)$

3.1. Pre-Processing

Suppose $S = \{s_c\}_{c=1}^C$ is the set of obtained physiological signals, where C is the number (channels) of physiological signals. The signals are firstly segmented into multiple instances with a fixed length L. After the segmentation, the input of the algorithm become $X = \{x^i\}_{i=1}^n$, where $x^i \in R^{L \times C}$. The starting and the ending points of an instance are the starting and ending timestamps of the segmentation, respectively. The goal of CorrNet is to classify the V-A for each instance. For that, input X is divided into two modalities $X_1 = \{x_1^i\}_{i=1}^n, x_1^i \in R^{L \times C_1}$, and $X_2 = \{x_2^i\}_{i=1}^n, x_2^i \in R^{L \times C_2}$ ($C_1 + C_2 = C$) based on the information these physiological signals represent. For example, the two modalities could be oculomotor nerve system (ONS) and autonomic nervous system (ANS), where the signals from ONS (pupil dilation [75] and saccadic eye movement [76]) and from ANS (skin conductance [77] and skin temperature [78]) are grouped together, respectively.

3.2. Intra-Modality Feature Learning

The purpose of intra-modality feature learning is to (a) fuse the information from different signal channels within a modality and (b) learn local features within each instance. To achieve this target, a two-layer convolutional auto-encoder [79] is implemented. We use just shallow structure (two layers) instead of deep to avoid overfitting since each instance does not contain much information.

Suppose that $\phi_1 = \{\varphi_1^i\}_{i=1}^n, \varphi_1^i \in R^\omega$ is the latent vector of X_i in modality 1, where ω is the dimension of the latent space, the ϕ_1 can be obtained by 1D convolution:

$$\phi_1 = X_1 \otimes \psi(\omega, c) = X_1 \dot{\otimes} \psi_{11}(1,1) \bar{\otimes} \psi_{12}(\omega, c) \tag{1}$$

where $\dot{\otimes}$ and $\bar{\otimes}$ are the convolution operations on the dimension of channels and length of instances, respectively. $\psi_{11} \in R^{1 \times C_1}$ and $\psi_{12} \in R^{c \times 1}$ are the convolution kernels for two layers, where c is the size of the convolution kernel. The first convolution layer fuses information from different channels while the second layer extracts local features between different time samples inside each instance. The latent vectors are then reconstructed using a convolutional decoder:

$$\eta_1 = \phi_1 \bar{\otimes} \psi'(C_1, c) \tag{2}$$

where $\psi' \in R^{c\times 1}$ is the convolution kernel for the decoder. The auto-encoder-decoder is trained by minimizing the binary cross entropy [80,81]:

$$H = -\frac{1}{n \cdot L}\sum_{i=1}^{n}\sum_{j=1}^{L} x_1^{ij} \cdot log(\eta_1^{ij})) + (1 - x_1^{ij}) \cdot (1 - log(\eta_1^{ij}) \tag{3}$$

where x_1^{ij} and η_1^{ij} are the j sample point in the instance of x_1^i and η_1^i, respectively. The latent vector ϕ_1 learned from the auto-encoder is the intra-modality features we want to obtain. The latent vector ϕ_2 for modality 2 can be calculated using the same method.

3.3. Correlation-Based Feature Extraction

In this stage, intra-modality features ϕ_1 and ϕ_2 are fused using a correlation-based feature extraction method [82]. The purpose of correlation-based feature extraction is to extract features which (a) maximize the correlation coefficient between two modalities and (b) fuse the features between different instances. The precise classification for each instance needs to take advantage of both local information within each instance and global information between different instances, as the change of signals are sometimes not synchronized with the change of emotions. Here, we hypothesize that the same video stimuli will trigger relatively similar valence and arousal across physiological responses among different subjects. Thus, the features from instances under the same stimuli are fused with the features from the other modality by maximizing the correlation between two modalities. The transformation which maps signals to features is a weak constraint because it is a linear mapping which does not bring new linearly independent features. If we use audio-visual features (which would be the same for all subjects for one video) from video content, it will bring strong constraints to all instances for subjects watching one video. In the extreme case, the classifier could rely only on the content-based features and discard the information from physiological signals. The linear transformation however extracts features that differ across subjects, so we do not have the same features for all subjects. Here, we use linear transformation instead of other complex transformations (e.g., deep structure [83]) to lower the computational cost and avoid overfitting (where a strong constraint can make the two modalities have a correlation coefficient of ≈1).

To extract correlation-based features, we first calculate the covariance (S_{11} and S_{22}) and cross-covariance (S_{12}) of the two modalities:

$$S_{11} = \frac{(\phi_1^t).^T\phi_1^t}{D^t - 1} + I^{\omega\times\omega}, S_{12} = \frac{(\phi_2^t)^T\phi_1^t}{D^t - 1}, S_{22} = \frac{(\phi_2^t)^T\phi_2^t}{D^t - 1} + I^{\omega\times\omega} \tag{4}$$

where I is the unit matrix and ω is the dimension of the latent space, D^t is the dimension of ϕ_1^t. Then, we implement the Singular Value Decomposition (SVD) on the equation below:

$$[U, D, V] = \mathrm{SVD}(\mathrm{V_1D_1V_1^T} \cdot \mathrm{S_{12}} \cdot \mathrm{V_2D_2V_2^T}) \tag{5}$$

where D_1 and D_2 are diagonal matrices whose diagonal elements are the k biggest non-zero eigenvalues of S_{11} and S_{22}, respectively, where $D_1 = \mathrm{diag}(\frac{1}{\sqrt{D_{11}}}, \frac{1}{\sqrt{D_{12}}}, \ldots, \frac{1}{\sqrt{D_{1k}}})$ and D_2 have the same format). $V_1 = [V_{11}, V_{12}, \ldots, V_{1k}]$ is composed of the k corresponding eigenvectors of $[D_{11}, D_{12}, \ldots, D_{1k}]$, respectively, where V_2 is calculated using the same method. Now, the two linear projections (H_1^t, H_2^t) can be calculated by:

$$H_1 = V_1D_1V_1^T \cdot U', \quad H_2 = V_2D_2V_2^T \cdot V' \tag{6}$$

where U' and V' consist of the first K columns of U, V, respectively. At last, the correlation-based features of ϕ_1^t and ϕ_2^t can be obtained by: $F^t = [\phi_1^t \cdot H_1^t, \phi_2^t \cdot H_2^t]$. We then implement the above procedure among all the T stimuli and get the correlation based features $F \in R^{n\times 2K}$ for all n instances.

3.4. Broad Learning System For Classification

While the previous two stages focus on unsupervised feature extraction, the last stage (Figure 1) focuses on a supervised classifier. Here, a Broad Learning System (BLS) [84] is used to map the extracted features to valence and arousal. Compared with deep learning systems such as Deep Belief Networks (DBNs) [85] and Convolutional Neural Networks (CNNs) [86], BLS is less time-consuming because it does not need to use gradient descent to train the network with multiple epochs. BLS maps the original training data into two high dimensional nodes (i.e., feature nodes and enhance nodes). Instead of using backpropagation to calculate the weights between the nodes and labels, BLS calculates the weights through pseudo-inverse, which makes the classification process faster and lowers likelihood of avoid overfitting [87].

Suppose $F' \in R^{n' \times 2K}$ is the training set selected from the features $F \in R^{n \times 2K}$. We first normalize F' to have mean of 0 and standard deviation of 1 using z-score normalization [88]. Then, the first feature node A_1 can be calculated by:

$$A_1 = F'' \cdot W_{A_1} \tag{7}$$

where $F'' = [F'|1]$ is the augmented matrix of F'. $W_{A1} \in R^{2K \times N_1}$ is the sparse autoencoder [89] of a random matrix W' whose element $w'_{ij} \in [-1, 1]$ are random numbers. BLS use random matrices as transformation matrices to map training data into high dimensional space. Although this method is fast, the nature of randomness suffers from unpredictability [84]. That is why an autoencoder is used to to slightly fine-tune the random nodes to a set of sparse and compact nodes. Generally, the sparse autoencoder can be obtain by solving a optimization problem [89]:

$$W_{A1} = \arg\max ||W' \cdot W_{A1} - H''||_2^2 + \lambda ||W_{A1}||_1$$
$$W_{A1} \cdot H'' = W' \tag{8}$$

where $\lambda = 10^{-3}$ is the regulation parameter.

With the same method, we can generate all N_2 high-dimensional nodes $A = \{A_i\}_{i=1}^{N2}$. Then, we calculate the enhance nodes B by:

$$B = \text{tansig}[\frac{\mathrm{A}' \cdot \text{orth}(\mathrm{W}'') \cdot \mathrm{S}}{\max(\mathrm{A}' \cdot \text{orth}(\mathrm{W}''))}] \tag{9}$$

where $A' = [A|1]$ is the augmented matrix of A. orth(W'') stands for the ortho-normalization of the random matrix W'', whose element $w''_{ij} \in [-1, 1]$ are random numbers. $S = 1200$ is the shrinkage parameter of the enhanced nodes. tansig $= \frac{2}{1+e^{-2t}} - 1$ is the active function for the enhance nodes. After that, we can obtain the input nodes $E = [A, B]$ in the two high dimensional spaces.

The last step of BLS is to calculate the weights between the input nodes and labels. Suppose the network can be presented as $EW = y$, where the W is the connection weights between the input nodes E and output labels $y, y = A_r$ (arousal) or $y = V_a$ (valence), the weights can be obtained by $W = E^{-1}y$. Although the real inverse E^{-1} is hard to calculate, we can estimate W with pesudo-inverse [84]:

$$W = (E^T \cdot E + I^{n' \times n'} \cdot C)^{-1} E^T \cdot y \tag{10}$$

$C = 2^{-30}$ is the regularization parameter for sparse regularization. After this, the network has been established and all parameters are settled. If a new sample E_t comes, the output y_t can be obtained by $y_t = E_t \cdot W$.

4. Datasets

To evaluate the performance of CorrNet, we test it on two datasets: CASE and MERCA. To the best of our knowledge, Continuously Annotated Signals of Emotion (CASE) [19] is the only published dataset which has continuously self-annotated physiological signals. However, the CASE dataset is collected in an indoor, desktop environment. To verify the validity of CorrNet using wearable physiological sensors, we collected continuous self-annotated physiological signals. Here, users annotated their valence and arousal levels using a continuous mobile annotation technique (cf., [32]) in a controlled, outdoor environment. This data collection resulted in the Mobile Emotion Recognition with Continuous Annotation (MERCA) dataset, which we describe below in Section 4.2. Testing on MERCA allows us to additionally test performance across different application scenarios (i.e., CASE: indoor-desktop video watching; MERCA: outdoor-mobile video watching). Details on each dataset are shown below.

4.1. CASE Dataset

The CASE dataset [19] contains physiological recordings from 30 participants (15 m, 15 f), aged between 22–37. Valence and arousal are annotated by participants using a physical joystick (shown in Figure 2) while they watched eight video clips on a desktop screen. The data collection experiment for CASE is a 1 (task: watch videos and continuously annotate emotions) $\times$ 4 (video emotions: amusing vs. boring vs. relaxing vs. scary) within-subjects design, tested in an indoor laboratory environment. Eight video clips (two videos per emotion, duration M = 158.75 s and SD = 23.67 s) were selected to elicit the corresponding emotions. These videos are clips chosen from movies and documentaries. The emotional content of the videos used in CASE dataset was verified in a pre-study [19]. The authors first selected 20 video clips from previous works [90,91] and thereafter let 12 participants (no overlap with the participants of the data collection experiment) view and rate these videos. Then the eight videos that have the highest inter-annotator agreement were selected. Six sensors (ECG, BVP, EDA, RESP, TEMP, EMG (3 channels), shown in Table 1) were equipped to collect physiological signals. All sensors were synchronized and sampled at 1000 Hz (sample size: 2,451,650 samples $\times$ 8 signals $\times$ 30 participants). The V-A ratings (sample size: 49,033 samples $\times$ 2 annotations $\times$ 30 participants) were collected in 20 Hz according to the sampling rate of the physical joystick.

Figure 2. The experiment setup and annotation interface for CASE [19].

4.2. MERCA Dataset

4.2.1. Experiment Setup

In total, 20 participants (12 m, 8 f) aged between 22 and 32 participated in the data collection experiment of MERCA. The number of participants in MERCA dataset is similar to some of the widely used emotion recognition datasets (e.g., CASE [19], K-EmoCon [92], DECAF [93]) with continuously annotated physiological signals. Participants were recruited from different institutions with diverse backgrounds, education levels and nationalities. All were familiar with watching videos on smartphones, and none reported visual, auditory or motor impairments. Our experiment strictly followed human data collection guidelines through our institute's ethics and data protection committee, where informed consent was obtained from all participants. As in CASE, the data collection experiment for MERCA followed a 1 (task: watch videos and continuously annotate emotions) $\times$ 4 (video emotions: joy vs. fear vs. sad vs. neutral) within-subjects design. As shown in Figure 3,

the experiment was conducted in the outdoor campus of our institute. Participants could walk or stand freely while watching videos. Participants were told to watch the videos as they normally would in such settings. To prevent participants from running into obstacles, traffic, or other people, the experimenter always accompanied the participant from a distance to guarantee their safety. The experiment setting parallels watching mobile videos while walking or waiting for a bus or train, which is a common phenomenon in mobile video consumption [94–96]. Figure 4 illustrates how our experiment setting parallels the application scenario of evaluating the user experience when watching mobile videos. When watching mobile videos, users would be equipped with wearable sensors to measure their physiological signals ubiquitously (with their consent). The signals will then be sent to the servers of the video provider to recognize the emotions of users in fine granularity. Lastly, the obtained emotions will be aligned with the video content for the video providers to analyze the relationship between video content and user emotions.

Figure 3. The experiment environment of MERCA. Participant photos shown with permission.

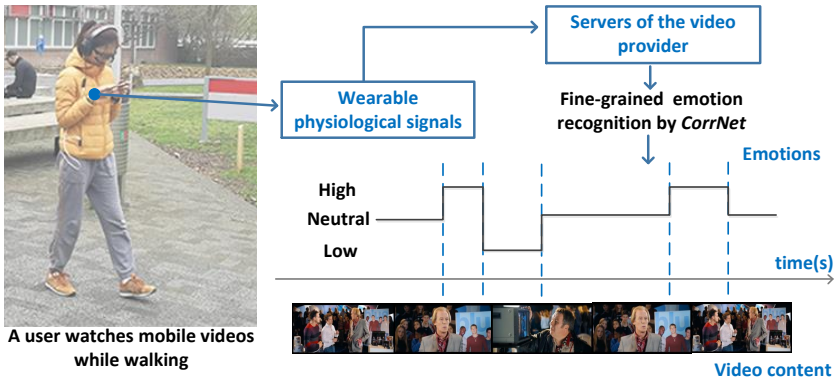

Figure 4. The illustration of CorrNet for evaluating mobile video watching user experience.

4.2.2. Video Stimuli

In total, 12 video clips (three videos per emotion, duration M = 81.4 s and SD = 22.5 s) were selected to elicit the corresponding emotions. Ten-second black screens were added before and after each video to decrease the effects of emotions overlapping among different videos. We chose the 12 videos according to 2D emotion annotations from the self-reports in MAHNOB dataset [97]. We use the videos in MAHNOB dataset because it is a widely used dataset [98,99] with emotion self-reports from more than 30 reviewers. We selected more videos compared with CASE because we aim to collect more samples for each emotion.

4.2.3. Software Setup

Emotions (as V-A) are annotated by participants using a real-time, continuous emotion annotation (RCEA) mobile application [32]. Participants can input their valence and arousal using a virtual joystick (shown in Figure 5) on the screen of the mobile device which they use for video watching. The virtual joystick is designed based on Russell's Circumplex model [100]. The x and y axes of the joystick represent valence and arousal, respectively. Four colors are selected for four quadrants of the joystick base on Itten's color system [101] to give users feedback on which emotion users are currently annotating. A gradual transparency from the origin (0% transparency) to the edge (100% transparency) of the joystick is designed to minimize the overlapping area between the video player and the virtual joystick. The transparency is also an indication of the transition of V-A intensity. We also map the frame colors to each corresponding V-A quadrant for additional peripheral feedback of which emotion users are currently annotating. Before the experiment, a 15-minute tutorial was given to familiarize participants with the operation of annotating.

Figure 5. The real-time and continuous V-A annotation interface (cf., [32]) used for MERCA.

4.2.4. Data Collection

We used the Pupil Core wearable and Empatica E4 wristband to collect signals from Autonomic Nerve System (ANS) and Oculomotor Nerve System (ONS), respectively. We chose these two devices because they are wearable, which are suitable for collecing signals in outdoor environments and have been used by previous studies [30,102,103]. We placed the Empatica E4 tightly on users' wrist to avoid movement of the electrodes and that was checked by the experimenter whenever the experiment started. The experimenter also checked whether the electrodes are in the right position and the recording device could get stable signals instead of noise. We waited approximately three minutes before the start of the experiment to make sure the signal collection is stable.

From Empatica E4, we collected HR (1326×20) (sample size, samples $\times$ participants), BVP (84864×20), EDA (5304×20) and TEMP (5304×20) (shown in Table 1). From the wearable eyetracker, we collected pupil dilation (13260×20), saccadic amplitude (13260×20) and saccadic velocity (13260×20). Data from these two sensors were stored on one mobile device (the recording device). As shown in Figure 6, the eye tracker and E4 wristband were connected to the recording device through a USB-C cable and low-power bluetooth, respectively. The data from the two devices do not interfere with each other because they are connected to the recording device using different ports. Another mobile device (the displaying device) was used for showing the videos and collecting annotations. A noise-cancelling headphone was connected to the displaying device via Bluetooth. Timestamps of both devices were set according to the clock of the recording device, where all data is synchronized via an NTP server. The V-A ratings (sample size: 13,260 samples $\times$ 2 annotations $\times$ 20 participants) were collected in 10 Hz according to the sampling rate of the virtual joystick. The annotations on video level (post-stimuli) consist of 52.28% of all the annotations for the entire video watching. In total, 85.41% of annotations in one video watching are distributed across different VA planes, which demonstrates that different emotions can occur within one video watching.

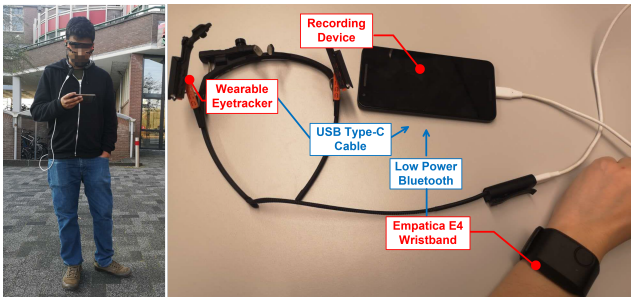

Figure 6. The hardware setup of MERCA. Image of study participant shown with permission.

Table 1. Technical and physiological specifications of sensors used in CASE and MERCA dataset.

Dataset	Signals	Sensor	Sampling Rate	Physiological System
CASE	ECG	SA9306	1000 Hz	Autonomic Nervous System
	BVP	SA9308M	1000 Hz	
	EDA	SA9309M	1000 Hz	
	RESP	SA9311M	1000 Hz	
	TEMP	SA9310M	1000 Hz	
	EMG	SA9401M-50	1000 Hz	Facial Nerve System
MERCA	HR	Empatica E4	1 Hz	Autonomic Nervous System
	BVP	Empatica E4	64 Hz	
	EDA	Empatica E4	4 Hz	
	TEMP	Empatica E4	4 Hz	
	Pupil dilation	Pupil Core	10 Hz	Oculomotor Nerve System
	Saccadic amplitude	Pupil Core	10 Hz	
	Saccadic velocity	Pupil Core	10 Hz	

Electrocardiogram (ECG), Blood Volume Pulse (BVP), Electrodermal activity (EDA), Respiration (RESP), Skin Temperature (TEMP), Electromyography (EMG), Heart Rate (HR).

5. Experiment and Results

In this section, we first introduce the implementation details of CorrNet for the CASE and MERCA datasets. We then evaluate the performance of CorrNet by both subject-dependent (SD) and subject-independent (SI) models, and compare with state-of-the-art approaches. Then, we conduct an ablation study to analyze the impact of different components in CorrNet. Lastly, we discuss about the computational complexity of the CorrNet.

5.1. Implementation Details

To decrease measurement bias in different trials, all signals (both CASE and MERCA) are normalized to $[0, 1]$ using Min-Max scaling normalization:

$$S_n = \frac{S - \min(S)}{\max(S) - \min(S)} \tag{11}$$

Normalization is implemented on each subject under each video stimuli (trial). Since signals in MERCA have different sampling rates, they are interpreted to the 32 Hz using linear interpretation [104]. Since the sampling rates of V-A and signals are 20 and 1000 Hz respectively, we down-sampled all the signals to 50 Hz by decimation down-sampling [105] (the choice of down-sampling rates is discussed in Section 6.2). The EDA signals were first filtered using a low pass filter with a 2 Hz cutoff frequency to remove noise [106]. For the BVP signal, we pre-processed it with a four-order butterworth bandpass filter with cutoff frequencies [30, 200] Hz to eliminate the bursts [107]. An elliptic band-pass filter

with cutoff frequencies [0.005, 0.1] was used to filter the ST signal [108]. We followed the standard filtering procedure widely used in previous works [6,106–108] to pre-process the physiological signals. Then the filtered signals are segmented into 2-second (sample size: 100 for CASE, 64 for MERCA) instances (the different choice of the segmentation length is discussed in Section 6.1). The intra-modality features are trained using adadelta optimizer [109] since it can automatically adapt learning rate. We used the Early-Stopping [110] technique to terminate training intra-modality features if there is no improvement on the validation loss for five epochs. The choice of other hyperparameters is listed in Table 2.

Table 2. The value of hyper-parameters in CorrNet.

Hyper-Parameter	Meaning	Value	
		CASE	MERCA
L	Length of each instance	2 s (100)	2 s (64)
ω	Dimension of latent space	2× L (200)	2 × L (128)
c	Size of conv-kernel	L/4 (25)	L/4 (16)
K	Dimension of corr features	L/2 (50)	L/2 (32)

We set ω to L/4 = 0.5 s because 0.5 is the smallest duration of emotions [13,14]. The dimensions of latent space and output of the correlation-based features are selected based on parameter optimization. If we increase ω and K, the latent vector and correlation-based features will start to contain redundant information (repeated values for all latent vectors and zeros for all correlation-based features). Our model is implemented using Keras. All our experiments are performed on a desktop with NVIDIA RTX 2080Ti GPU with 16 GB RAM.

5.2. Evaluation Protocol and Baselines

5.2.1. Classification Tasks

Three classification tasks were tested across both datatsets: (1) binary classification for low/high level of arousal and valence, (2) 3-class classification for low/neutral/high level of arousal and valence, (3) 4-class classification for the four quadrants of V-A space. We use the mean V-A of each instance as labels for classification. The mapping from continuous values of V-A to discretized categories is listed in Table 3.

Table 3. The mapping of V-A values and discretized classes.

Class	V-A Ratings (Binary)	V-A Ratings (3-Class)
Low	[1, 5)	[1, 3)
Neutral	-	[3, 6)
High	[5, 9]	[6, 9]
4-Class	**valence ratings**	**arousal ratings**
High-High (HH)	[9, 5)	[9, 5)
High-Low (HL)	[9, 5)	[5, 1)
Low-Low (LL)	[5, 1)	[5, 1)
Low-High (LH)	[5, 1)	[9, 5)

5.2.2. Evaluation Metrics

Three evaluation metrics are chosen to evaluate the performance of CorrNet:

- **Accuracy:** the percentage of correct predictions;
- **Confusion matrix:** the square matrix that shows the type of error in a supervised paradigm [70];
- **Weighted F1-score (W-F1):** the harmonic mean of precision and recall for each label (weighted averaged by the number of true instances for each label) [111].

These three metrics are widely used in evaluating machine learning algorithms [112]. We use weighted F1-score instead of macro and binary F1-score to take into account label imbalance.

5.2.3. Evaluation Method

We train and test the proposed method using both subject-dependent (SD) and subject-independent (SI) models. Subject-dependent model was tested using 10-fold cross validation. For each subject, their data are divided into 10 folds. We train CorrNet using nine folds and tested on the remaining fold. The subject-independent model is tested using Leave-one-subject-out cross validation (LOSOCV). Data from each subject are separated as testing data and the remaining data from other subjects are used for training. The results we show are the mean accuracy and W-F1 of each fold/subject used as testing data.

5.2.4. Baseline Comparison

Since there are no existing baseline methods, we compare the performance of CorrNet with both deep learning (DL) methods and classic machine learning (ML). For DL methods, we compare with 1D-CNN [113] with two and four convolutional layers. We tested 1D-CNN with a different number of convolutional layers to test whether the accuracy could be increased by making the network deeper. We also compare the performance with sequential learning approaches including LSTM [33,35] and Bidirectional LSTM (BiLSTM) [34,114] because they are widely used for the classification of time series. We train the 1D-CNN, LSTM and BiLSTM with the adadelta optimizer [109], which is the same as we used for training the intra-modality features. For ML methods, we compare with Support Vector Machine (SVM) [58], K-Nearest Neighbor (KNN) [59], Random Forest (RF) [60] and Gaussian Naive Bayes (GaussianNB) [115]. These methods are commonly used as baseline methods in datasets [62,116] and review [6,7,117] papers for affective computing. To train these ML models, we first pre-processed the signals using the same method we described in Section 5.1. We then select the mean, standard variance, average root mean square, mean of the absolute values, maximum amplitude and average amplitude for the original, first and second differential of all physiological signals. These are widely-used features for physiological signals in the task of emotion recognition [6].

5.3. Experiment Results

Performance of CorrNet on CASE and MERCA is shown in Table 4. In general, the subject-dependent (SD) model achieves higher accuracy and W-F1 than the subject-independent (SI) model, especially for the 3-class classification on MERCA. The accuracy of 4-class classification (four quadrants of V-A space) is lower than binary but higher than 3-class classification. Although the number of classes is higher, 4-class classification does not include testing between neutral and high/low (only two classes on V-A, respectively). Thus, the 3-class testing (high/neutral/low) on V-A independently is more challenging than four quadrants. To summarize, the overall performance on CASE is better than the performance on MERCA, which means a controlled, mobile environment can bring more challenges for emotion recognition. However, the performance on both datasets is comparable, both achieving more than 70% accuracy on binary classification and more than 60% accuracy on 3-class classification using a subject-dependent model. The results show good generalizability among different physiological signals and testing environments (desktop-indoor and mobile-outdoor).

Table 4. Validation results for CASE and MERCA.

	10-Fold (SD)				LOSOCV (SI)			
	CASE		MERCA		CASE		MERCA	
	acc	f1	acc	f1	acc	f1	acc	f1
valence-2 [1]	77.01%	0.74	75.88%	0.75	76.37%	0.76	70.29%	0.70
arousal-2 [1]	80.11%	0.79	74.98%	0.74	74.03%	0.72	68.15%	0.67
valence-3 [2]	61.83%	0.61	63.89%	0.63	60.15%	0.53	53.88%	0.53
arousal-3 [2]	62.03%	0.61	66.04%	0.65	58.22%	0.55	46.21%	0.42
4-class	69.36%	0.67	72.16%	0.70	55.08%	0.53	51.51%	0.50

[1] Binary classification. [2] 3-class classification.

5.4. Comparison with DL and ML Methods

The comparison of DL and ML methods with CorrNet using a subject-independent model is shown in Table 5. Compared with subject-dependent models, the subject-independent model is more challenging for training, which lead to less subject-bias (overfitting on specific subjects and resulting in high accuracy). Thus, we use subject-independent models to compare the performance of different methods. As shown in Table 5, for the 1D-CNN, deepening the network does not result in better performance. In fact, if we keeping increasing the number of convolution layers, the network will overfit on the training set. Here we can speculate that the information inside each instance is limited and insufficient to train a deep discriminative model. The performance of LSTM and BiLSTM is similar to 1D CNN, which means the recurrent structure does not help to increase the recognition accuracy. In general, CorrNet outperforms both ML and DL methods since it takes advantage of information across both modalities and their correlation. The only exception is that DL methods achieve higher accuracy (but lower W-F1) compared with CorrNet in 3-class classification of arousal on MERCA. High accuracy and low W-F1 means that the algorithm performs well only on a specific class (i.e., neutral arousal), which is a result of overfitting on that class. Thus, compared with DL methods, CorrNet has better performance of generalization among different classes.

5.5. Ablation Study

As stated, CorrNet contains three major components: intra-modality feature learning (IFL), correlation-based feature extraction (CFE), and broad learning system (BLS) for classification. We conduct an ablation study to verify the effectiveness of each component. We begin with only using the classifier on the raw signals. Then we test the performance of combining IFL and CFE with BLS independently. The results of binary classification trained using LOSOCV is shown in Table 6.

From the results, we draw the following observations: (1) Simply combining IFL and BLS does not improve classification performance when using only BLS on the raw data. IFL is a step of fusing signals from different channels and extracts local features within each instance. This is a step of information compression, thus it does not provide additional information other than what is provided from raw signals. However, it compresses the information within each instance and helps improve accuracy while combining with CFE. (2) The combination of CFE and BLS improves accuracy compared with using only BLS, however it is still lower than combining all three components. The results demonstrate the significance of fusing features between two modalities based on their correlation. (3) All components contribute to the classification task. The proposed CorrNet algorithm that jointly combines features within and between instances performs the best. These observations demonstrate the effectiveness of the proposed algorithm.

Table 5. Comparison between ML, DL methods and CorrNet using LOSOCV (accuracy (W-F1)).

	Deep Learning Methods				
	1D-CNN-2 [5]	**1D-CNN-4** [6]	**LSTM**	**BiLSTM**	**CorrNet**
valence-2 [1]	58.26% (0.53)	58.00% (0.52)	48.58% (0.40)	48.81% (0.41)	**76.37% (0.76)**
arousal-2 [1]	51.38% (0.44)	56.04% (0.48)	51.29% (0.38)	54.19% (0.42)	**74.03% (0.72)**
valence-3 [2]	50.51% (0.38)	49.31% (0.35)	50.44% (0.35)	51.58% (0.36)	**60.15% (0.53)**
arousal-3 [2]	45.89% (0.31)	47.11% (0.31)	40.52% (0.31)	42.12% (0.33)	**58.22% (0.55)**
valence-2 [3]	58.13% (0.49)	56.98% (0.48)	56.01% (0.46)	59.21% (0.46)	**70.29% (0.70)**
arousal-2 [3]	58.11% (0.54)	56.79% (0.53)	51.37% (0.49)	51.90% (0.50)	**68.15% (0.67)**
valence-3 [4]	45.23% (0.32)	43.50% (0.32)	46.62% (0.31)	46.56% (0.31)	**53.88% (0.53)**
arousal-3 [4]	45.41% (0.32)	46.56% (0.33)	**47.75% (0.32)**	47.70% (0.32)	46.21% (0.42)
	Classic Machine Learning Methods				
	SVM	**KNN**	**RF**	**GaussianNB**	**CorrNet**
valence-2 [1]	49.02% (0.42)	50.76% (0.50)	48.83% (0.48)	50.99% (0.39)	**76.37% (0.76)**
arousal-2 [1]	51.22% (0.42)	51.13% (0.51)	50.46% (0.49)	52.08% (0.41)	**74.03% (0.72)**
valence-3 [2]	42.52% (0.30)	38.95% (0.37)	37.62% (0.35)	43.26% (0.31)	**60.15% (0.53)**
arousal-3 [2]	50.18% (0.35)	43.38% (0.40)	42.29% (0.39)	27.98% (0.15)	**58.22% (0.55)**
valence-2 [3]	50.92% (0.39)	51.27% (0.51)	50.78% (0.50)	48.34% (0.38)	**70.29% (0.70)**
arousal-2 [3]	57.16% (0.45)	51.34% (0.51)	49.85% (0.49)	52.59% (0.42)	**68.15% (0.67)**
valence-3 [4]	44.89% (0.30)	37.89% (0.36)	38.48% (0.37)	24.91% (0.15)	**53.88% (0.53)**
arousal-3 [4]	44.49% (0.32)	37.52% (0.37)	38.44% (0.37)	34.68% (0.24)	**46.21% (0.42)**

[1] Binary classification on CASE. [2] 3-class classification on CASE. [3] Binary classification on MERCA. [4] 3-class classification on MERCA. [5] 1D-CNN with 2 convolutional layers. [6] 1D-CNN with 4 convolutional layers.

Table 6. Ablation study of different components in CorrNet (accuracy (W-F1)).

	CASE		**MERCA**	
	Valence	**Arousal**	**Valence**	**Arousal**
BLS	52.68% (0.50)	56.53% (0.56)	57.26% (0.57)	57.88% (0.49)
IFL + BLS	53.79% (0.46)	57.80% (0.57)	57.96% (0.56)	58.78% (0.45)
CFE + BLS	69.80% (0.68)	66.41% (0.63)	65.43% (0.65)	63.82% (0.63)
IFL + CFE + BLS	**76.37% (0.76)**	**74.03% (0.72)**	**70.29% (0.70)**	**68.15% (0.67)**

5.6. Computational Cost

The time complexity of CorrNet is $O((\omega^2 + L\omega)n^2) + (2c+1)\omega Ln + \omega K)$ for training and $O((c + K + 1)\omega n)$ for testing. The computational cost of CorrNet is not high due to (a) the simple (2-layer) structure for intra-modality feature learning, (b) the linear mapping (instead of other complex transformation) in correlation-based feature extraction, and (c) the use of pesudo-inverse (instead of gradient descent) in broad learning. The average training time on our testing machine (desktop with NVIDIA RTX 2080Ti GPU with 16 GB RAM), is 65.56 s and 24.67 s for CASE and MERCA, respectively (sampling rate = 50 Hz). The average detection time for each fine-grained instance is 29.01 ms, which means to recognize 2 s emotions, the algorithm only spends less than 30 ms after the network is trained.

6. Discussion

6.1. Towards More Precise Emotion Recognition: How Fine-Grained Should It Be?

The length of an instance is one of the key parameters which needs to be selected carefully when designing fine-grained emotion recognition algorithms. The shorter the lengths are, the finer the granularity of an emotion that could be recognized. However,

since emotion states are classified based on the information from each instance, this could entail that without sufficient information and the classification task becomes a random guess using irrelevant numbers.

To find the appropriate length of an instance, we conduct an experiment by testing CorrNet using different segmentation lengths. As shown in Figure 7, the W-F1 tested on CASE drops significantly after reducing the length to 0.5 s while the dropping threshold for MERCA is 0.25 s. This finding is in line with the finding from Paul et al. [13] that the duration of an emotion typically spans 0.5–4 s. The W-F1 also decreases after increasing the length to 8 s. Here we can speculate that overly high length instances could result in an inaccurate ground truth (more than one emotion in each instance) for classification. We find that the decrease of W-F1 on MERCA is more dramatic than the decrease on CASE, which indicates that for indoor-desktop environments, the emotion changes more slowly compared with outdoor-mobile environments (more instances with a longer length contain only one emotion). These results show that the segmentation length between 1–4 s can result in good performance (high W-F1), which can serve as an appropriate length to classify emotions using fine-grained emotion labels.

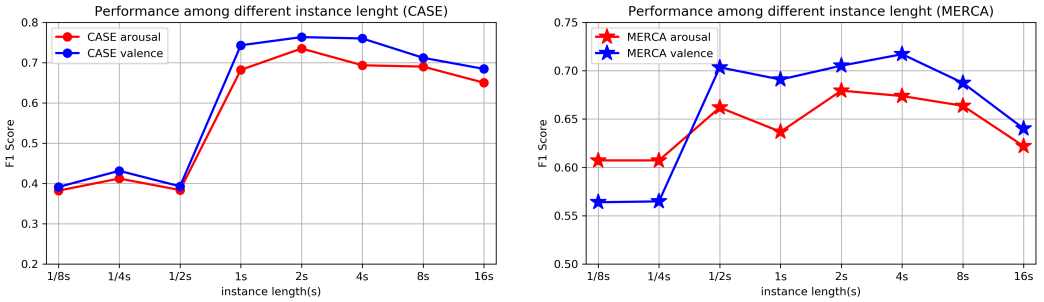

Figure 7. Comparison of the performance among different instance lengths: W-F1 of binary classification (LOSOCV).

6.2. Emotion Recognition Using Wearable Physiological Sensing: Do Higher Sampling Rates Result in Higher Accuracies?

Traditionally, physiological sensors designed for laboratory environments often have high sampling rates ($\geq$1000 Hz). Ideally, a higher sampling rate means better recovery of the original signal. However, a high sampling rate can also result in high power consumption and high-frequency noise, which can pose problems for usage of wearable sensors (i.e., the battery of wearable sensors is limited) in ubiquitous environments (i.e., more signal noise can occur compared with indoor laboratory environments). As our work focuses on fine-grained emotion recognition using wearable physiological sensors, it is worthwhile to investigate the influence of different sampling rates on CorrNet.

As the original sampling rate of CASE is 1000 Hz, we gradually down-sample the signals from CASE to 1 Hz and test the performance of CorrNet under different sampling rates. Although CASE was collected in a desktop environment, including it as an additional dataset helps us compare the results between laboratory-grade and wearable sensors. The down-sampling is implemented by decimating the last sampling point of every down-sampling segment. The decimate down-sampling we use is a simulation of collecting signals using wearable sensors with low sampling rate. The decimate down-sampling drops sampling points of signals in a fixed temporal interval to simulate that the A/D converter measures a continuous signal with lower frequency. Suppose the original signal $S = [s_1, s_2, \ldots, s_N]$ and the signal after down-sampling X is:

$$X = [x_{1M}, x_{2M}, \ldots, x_{kM}] \tag{12}$$

where $M = \frac{F1}{F2}$, $K = \frac{N}{M}$. $F1$ and $F2$ are the sampling rates before and after down-sampling, respectively.

Figure 8 shows the weighed F1 score and detection time among different sampling rates. As shown in Figure 8 (left), down-sampling to 50 Hz does not significantly decrease the W-F1 score. However, the detection time for each fine-grained instance increases dramatically if we raise the sampling rate to greater than 50 Hz. This result helps explain why for most of the wearable devices (e.g., Empatica E4 wristband, BITalino Kit, the highest sampling rate of physiological sensors is less than 64 Hz (e.g., 32 Hz for Empatica E4 and 40 Hz for BITalino Kit). The comparable recognition accuracy testing on the CASE and MERCA datasets also shows low sampling rates (32 Hz) do not significantly affect the performance of emotion recognition algorithms. Our result is consistent with the findings of Martin et al. [30], where the recognition accuracy is similar between the data collected using laboratory and wearable sensors. The take away message of this experiment is that physiological signals collected from wearable devices with a low sampling rate can also be used for precise recognition of emotions (i.e., valence and arousal) for evaluating affective states during short-form video watching.

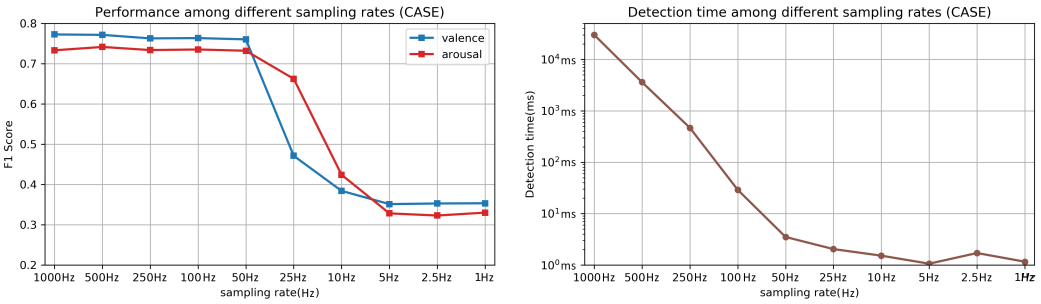

Figure 8. Comparison of the performance among different sampling rates: W-F1 of binary classification (LOSOCV, **left**) and detection time (**right**).

6.3. Data Imbalance and Overfitting in Fine-Grained Emotion Recognition

As shown in Figure 9 (down, LOSOCV)), there is an accuracy imbalance among different classes for 3-class classification (for binary classification we did not omit neutral labels but discrete them according to Table 3). We can see that the accuracy of class high and low (for both arousal and valence) is low, which does not occur when using the subject-dependent model. The test results on CASE are similar (instances with label of high (48%) and low (47%) are classified as neutral). Compared with the subject-independent model, the subject-dependent model is less sensitive to data imbalance, while there is still overfitting (about 30% of samples from high and low) on neutral category. We found that this can be a problem due to data imbalance when recognizing emotions using fine-grained emotion labels.

As shown in Figure 10, more than 60% of samples from CASE and 50% of samples from MERCA belong to the neutral class. The resulting high amounts of neutral V-A ratings cannot be attributed to the mobile aspect of MERCA's data collection, given that users spent most of their time (up to 73.2%) standing while watching and annotating [32]. We instead attribute this phenomenon to the act of annotating continuously, irrespective of environment (static vs. mobile). When users continuously annotate their emotions, they tend to annotate them as neutral by default (releasing virtual joystick) and non-neutral (actively annotating) only for specific scenes (e.g., kissing scenes for happy). These scenes only last for a short duration (users are not 100% of the time aroused), and for the remainder of the video clip users annotate their emotions as neutral.

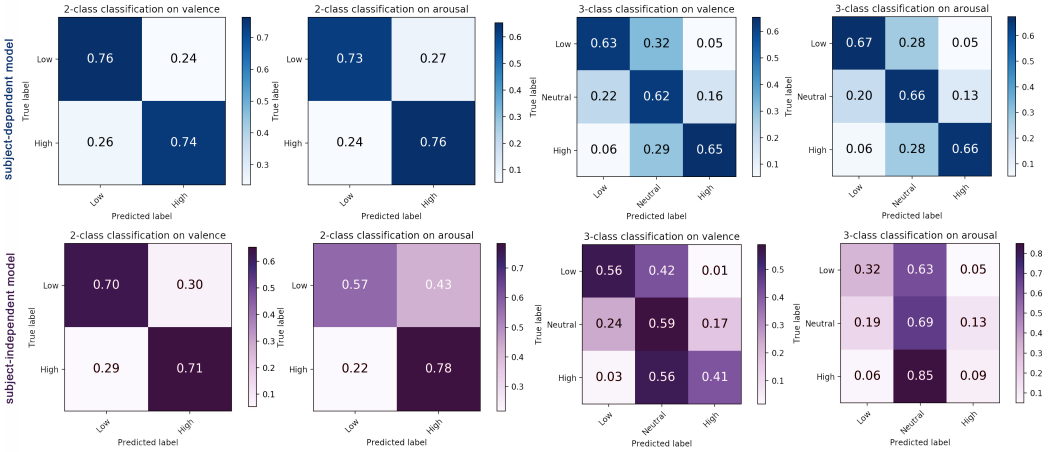

Figure 9. The result of 10-fold cross validation (subject-dependent model, **up**) and leave-one-subject-out cross validation (subject-independent model, **down**) on MERCA.

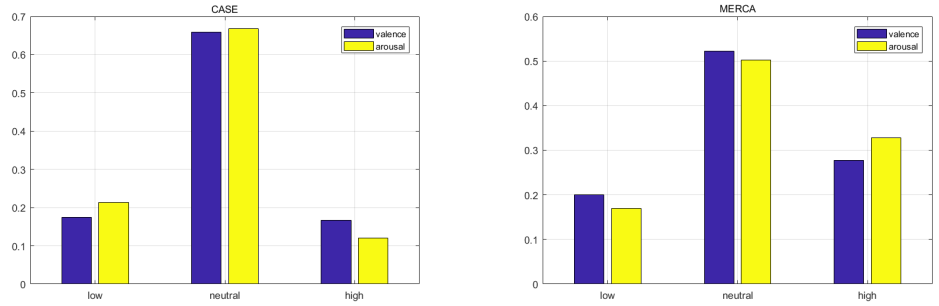

Figure 10. Sample percentage in each class of V-A.

The data imbalance can explain why the sequence learning techniques like LSTM do not perform well for such fine-grained emotion recognition. If most of the ground truth labels are neutral, the recurrent structure of sequence learning can easily overfit to output all classification results as neutral. The LOSOCV result shows the training accuracy of LSTM is 20.23% and 18.17% higher than the testing accuracy on CASE and MERCA respectively (averaged between V-A, 3-class classification). However, since CorrNet does not use the recurrent structure and learns the instance-label relationship independently, it does not suffer from the problem of overfitting: the training accuracy of CorrNet is only 1.01% and 4.82% higher than the testing accuracy on CASE and MERCA respectively (averaged between V-A, 3-class classification).

In addition, individuals differ in interoception levels, where self-reports of how they feel do not always correspond to their physiological response [118]. This is reflected in our observed patterns of physiological responses and continuous annotations. Thus, it also brings challenging for developing the subject-independent fine-grained emotion recognition algorithm. In general, the discussion above underscores the importance of carefully treating data imbalance and the problem of overfitting when designing any fine-grained emotion recognition algorithm.

7. Limitations and Future Work

Given the challenges of designing for fine-grained emotion recognition, there were naturally limitations to our work. First, although the performance of the subject-dependent model is relatively balanced among classes, the performance of the subject-independent model can still be improved if data imbalance is addressed. One promising approach is using the collected data to train a generative model (e.g., Generative Adversarial Networks [119]) to extend the size of the data for specific emotion categories (e.g., high arousal) by artificially generating more samples. Second, it is also essential for us to compare the performance of CorrNet on more datasets to further test its generalizability. However,

the number of datasets with continuously annotated physiological signals is to date limited. Additionally, there are no benchmark classification results for CASE, which is the only existing dataset with continuously annotated physiological signals. Thus, it is difficult to make comparisons with more advanced learning methods. Furthermore, although the computational time is short, CorrNet was not designed to predict valence and arousal in real-time. CorrNet requires the signals (in their entirety) under one stimulus as input to extract the correlation-based features. Such prediction of emotion can help users to avoid potential negative emotions such as fatigue while driving [120], or getting distracted during lectures [52].

At last, we only consider physiological signals and do not use other modalities such as facial expressions and EEG which contain more abundant information for emotion recognition [6]. CorrNet is designed to extract the correlation-based features from signals between two modalities. Thus, it is possible to extend CorrNet to other modalities such as EEG for better recognition accuracy. In this paper, we only test it using wearable physiological signals to maximize the generalizability of it towards different potential application scenarios (e.g., mobile video watching). Facial expressions, for example, are not always possible to capture when users are on the move [27], wearing a mask [121] and Head-Mounted Display (HMD) [122], or under the conditions with inadequate light [123]. In the future, we will extend CorrNet to use signals in other modalities and investigate whether the recognition accuracy can be further improved.

8. Conclusions

Physiological signals from different modalities contain different aspects of human emotions. In this work, we proposed CorrNet, a fine-grained emotion recognition algorithm to classify the fine-grained valence and arousal of users using wearable physiological signals while they watch videos. CorrNet takes advantage of the information both inside each instance (segmentation of signals) and between different instances under the same video stimuli. Our algorithm achieves good performance (more than 70% of accuracy on binary classification) on two datasets that differ in setting (indoor-desktop and outdoor-mobile), and outperforms both state-of-the-art DL and classic ML methods. Our experiments on different parameters of algorithms shows fine-grained emotion recognition, typically in 1–4 s, can be achieved with high accuracy and low computational cost using wearable physiological even under low sampling rates.

Author Contributions: Conceptualization: T.Z., A.E.A. and P.C.; Funding Acquisition: C.W.; Investigation: T.Z. and A.E.A.; Methodology: T.Z. and A.E.A.; Project Administration: P.C.; Supervision: A.E.A., C.W., A.H. and P.C.; Writing—Original Draft Preparation: T.Z.; Writing—Review & Editing: T.Z., A.E.A., A.H. and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work is fund by the Joint PhD Program between Xinhuanet and Centrum Wiskunde & Informatica.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Centrum Wiskunde & Informatica (protocol code P2542 and 27-04-2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the privacy of the participants.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Soleymani, M.; Pantic, M.; Pun, T. Multimodal emotion recognition in response to videos. *IEEE Trans. Affect. Comput.* **2011**, *3*, 211–223. [CrossRef]
2. Niu, J.; Zhao, X.; Zhu, L.; Li, H. Affivir: An affect-based Internet video recommendation system. *Neurocomputing* **2013**, *120*, 422–433. [CrossRef]
3. Tripathi, A.; Ashwin, T.; Guddeti, R.M.R. EmoWare: A Context-Aware Framework for Personalized Video Recommendation Using Affective Video Sequences. *IEEE Access* **2019**, *7*, 51185–51200. [CrossRef]
4. Yazdani, A.; Lee, J.S.; Vesin, J.M.; Ebrahimi, T. Affect recognition based on physiological changes during the watching of music videos. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2012**, *2*, 1–26. [CrossRef]
5. Ali, M.; Al Machot, F.; Haj Mosa, A.; Jdeed, M.; Al Machot, E.; Kyamakya, K. A globally generalized emotion recognition system involving different physiological signals. *Sensors* **2018**, *18*, 1905. [CrossRef] [PubMed]
6. Shu, L.; Xie, J.; Yang, M.; Li, Z.; Li, Z.; Liao, D.; Xu, X.; Yang, X. A Review of Emotion Recognition Using Physiological Signals. *Sensors* **2018**, *18*, 2074. [CrossRef] [PubMed]
7. Jerritta, S.; Murugappan, M.; Nagarajan, R.; Wan, K. Physiological signals based human emotion recognition: A review. In Proceedings of the 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, Penang, Malaysia, 4–6 March 2011; pp. 410–415.
8. Maria, E.; Matthias, L.; Sten, H. Emotion recognition from physiological signal analysis: A review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55.
9. Nagel, F.; Kopiez, R.; Grewe, O.; Altenmüller, E. EMuJoy: Software for continuous measurement of perceived emotions in music. *Behav. Res. Methods* **2007**, *39*, 283–290. [CrossRef]
10. Soleymani, M.; Asghari-Esfeden, S.; Fu, Y.; Pantic, M. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Trans. Affect. Comput.* **2015**, *7*, 17–28. [CrossRef]
11. Lang, P.J. The emotion probe: Studies of motivation and attention. *Am. Psychol.* **1995**, *50*, 372. [CrossRef]
12. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
13. Paul, E. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*; OWL Books: New York, NY, USA, 2007.
14. Levenson, R.W. Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. *Soc. Psychophysiol. Theory Clin. Appl.* **1988**, 17–42.
15. Domínguez-Jiménez, J.; Campo-Landines, K.; Martínez-Santos, J.; Delahoz, E.; Contreras-Ortiz, S. A machine learning model for emotion recognition from physiological signals. *Biomed. Signal Process. Control* **2020**, *55*, 101646. [CrossRef]
16. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
17. Cowie, R.; Douglas-Cowie, E.; Savvidou, S.; McMahon, E.; Sawey, M.; Schröder, M. 'FEELTRACE': An instrument for recording perceived emotion in real time. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, UK, 5–7 September 2000.
18. Girard, J.M.; Wright, A.G. DARMA: Software for dual axis rating and media annotation. *Behav. Res. Methods* **2018**, *50*, 902–909. [CrossRef]
19. Sharma, K.; Castellini, C.; van den Broek, E.L.; Albu-Schaeffer, A.; Schwenker, F. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Sci. Data* **2019**, *6*, 1–13. [CrossRef]
20. Soleymani, M.; Asghari-Esfeden, S.; Pantic, M.; Fu, Y. Continuous emotion detection using EEG signals and facial expressions. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 14–18 July 2014; pp. 1–6.
21. Haripriyadharshini, S.; Gnanasaravanan, S. EEG Based Human Facial Emotion Recognition System Using LSTMRNN. *Asian J. Appl. Sci. Technol. (AJAST)* **2018**, 2, 264-269
22. Hasanzadeh, F.; Annabestani, M.; Moghimi, S. Continuous Emotion Recognition during Music Listening Using EEG Signals: A Fuzzy Parallel Cascades Model. *arXiv* **2019**, arXiv:1910.10489.
23. Wu, S.; Du, Z.; Li, W.; Huang, D.; Wang, Y. Continuous Emotion Recognition in Videos by Fusing Facial Expression, Head Pose and Eye Gaze. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 40–48.
24. Zhao, S.; Yao, H.; Jiang, X. Predicting continuous probability distribution of image emotions in valence-arousal space. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 879–882.
25. Craik, A.; He, Y.; Contreras-Vidal, J.L. Deep learning for electroencephalogram (EEG) classification tasks: A review. *J. Neural Eng.* **2019**, *16*, 031001. [CrossRef]
26. Casson, A.J. Wearable EEG and beyond. *Biomed. Eng. Lett.* **2019**, *9*, 53–71. [CrossRef]
27. Khamis, M.; Baier, A.; Henze, N.; Alt, F.; Bulling, A. Understanding Face and Eye Visibility in Front-Facing Cameras of Smartphones Used in the Wild. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18), Montreal, QC, Canada, 21–26 April 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–12. [CrossRef]
28. Friedman, B.; Kahn, P.H., Jr.; Hagman, J.; Severson, R.L.; Gill, B. The watcher and the watched: Social judgments about privacy in a public place. *Hum. Comput. Interact.* **2006**, *21*, 235–272. [CrossRef]

29. Stanko, T.L.; Beckman, C.M. Watching you watching me: Boundary control and capturing attention in the context of ubiquitous technology use. *Acad. Manag. J.* **2015**, *58*, 712–738. [CrossRef]
30. Ragot, M.; Martin, N.; Em, S.; Pallamin, N.; Diverrez, J.M. Emotion recognition using physiological signals: Laboratory vs. wearable sensors. In *International Conference on Applied Human Factors and Ergonomics*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 15–22.
31. Gashi, S.; Di Lascio, E.; Stancu, B.; Swain, V.D.; Mishra, V.; Gjoreski, M.; Santini, S. Detection of Artifacts in Ambulatory Electrodermal Activity Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 1–31. [CrossRef]
32. Zhang, T.; El Ali, A.; Wang, C.; Hanjalic, A.; Cesar, P. RCEA: Real-Time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20), Honolulu, HI, USA, 26 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–15. [CrossRef]
33. Ma, J.; Tang, H.; Zheng, W.L.; Lu, B.L. Emotion Recognition using Multimodal Residual LSTM Network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 176–183.
34. Zhong, S.H.; Fares, A.; Jiang, J. An Attentional-LSTM for Improved Classification of Brain Activities Evoked by Images. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1295–1303.
35. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef] [PubMed]
36. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
37. Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; Jiang, H.; Inkpen, D. Enhanced lstm for natural language inference. *arXiv* **2016**, arXiv:1609.06038.
38. Bentley, F.; Lottridge, D. Understanding Mass-Market Mobile TV Behaviors in the Streaming Era. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19), Glasgow, UK, 4–9 May 2019; ACM: New York, NY, USA, 2019; pp. 261:1–261:11. [CrossRef]
39. Zhang, X.; Li, W.; Chen, X.; Lu, S. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 1–30. [CrossRef]
40. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
41. Taylor, B.; Dey, A.; Siewiorek, D.; Smailagic, A. Using physiological sensors to detect levels of user frustration induced by system delays. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 7–11 September 2015; pp. 517–528.
42. Kyriakou, K.; Resch, B.; Sagl, G.; Petutschnig, A.; Werner, C.; Niederseer, D.; Liedlgruber, M.; Wilhelm, F.H.; Osborne, T.; Pykett, J. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors* **2019**, *19*, 3805. [CrossRef]
43. Sethi, K.; Ramya, T.; Singh, H.P.; Dutta, R. Stress detection and relief using wearable physiological sensors. *Telkomnika* **2019**, *17*, 1139–1146. [CrossRef]
44. Salekin, A.; Eberle, J.W.; Glenn, J.J.; Teachman, B.A.; Stankovic, J.A. A weakly supervised learning framework for detecting social anxiety and depression. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–26. [CrossRef] [PubMed]
45. Costa, A.; Rincon, J.A.; Carrascosa, C.; Julian, V.; Novais, P. Emotions detection on an ambient intelligent system using wearable devices. *Future Gener. Comput. Syst.* **2019**, *92*, 479–489. [CrossRef]
46. Zenonos, A.; Khan, A.; Kalogridis, G.; Vatsikas, S.; Lewis, T.; Sooriyabandara, M. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In Proceedings of the 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), Sydney, Australia, 14–18 March 2016; pp. 1–6.
47. Ayata, D.; Yaslan, Y.; Kamasak, M.E. Emotion based music recommendation system using wearable physiological sensors. *IEEE Trans. Consum. Electron.* **2018**, *64*, 196–203. [CrossRef]
48. Yao, L.; Liu, Y.; Li, W.; Zhou, L.; Ge, Y.; Chai, J.; Sun, X. Using physiological measures to evaluate user experience of mobile applications. In *International Conference on Engineering Psychology and Cognitive Ergonomics*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 301–310.
49. Gashi, S.; Di Lascio, E.; Santini, S. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 1–19. [CrossRef]
50. Puke, S.; Suzuki, T.; Nakayama, K.; Tanaka, H.; Minami, S. Blood pressure estimation from pulse wave velocity measured on the chest. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 6107–6110.
51. Huynh, S.; Kim, S.; Ko, J.; Balan, R.K.; Lee, Y. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–27. [CrossRef]
52. Di Lascio, E.; Gashi, S.; Santini, S. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–21. [CrossRef]
53. Yang, W.; Rifqi, M.; Marsala, C.; Pinna, A. Towards Better Understanding of Player's Game Experience. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 442–449.
54. Wioleta, S. Using physiological signals for emotion recognition. In Proceedings of the 2013 6th International Conference on Human System Interactions (HSI), Sopot, Poland, 6–8 June 2013; pp. 556–561.

55. Niu, X.; Chen, L.; Xie, H.; Chen, Q.; Li, H. Emotion pattern recognition using physiological signals. *Sens. Transducers* **2014**, *172*, 147.
56. Zecca, M.; Micera, S.; Carrozza, M.C.; Dario, P. Control of multifunctional prosthetic hands by processing the electromyographic signal. *Crit. Rev. Biomed. Eng.* **2002**, *30*, 459–485. [CrossRef]
57. Calvo, R.A.; D'Mello, S. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* **2010**, *1*, 18–37. [CrossRef]
58. He, C.; Yao, Y.J.; Ye, X.S. An emotion recognition system based on physiological signals obtained by wearable sensors. In *Wearable Sensors and Robots*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 15–25.
59. Chen, L.; Li, M.; Su, W.; Wu, M.; Hirota, K.; Pedrycz, W. Adaptive Feature Selection-Based AdaBoost-KNN With Direct Optimization for Dynamic Emotion Recognition in Human–Robot Interaction. *IEEE Trans. Emerg. Top. Comput. Intell.* **2019**. [CrossRef]
60. Rigas, G.; Katsis, C.D.; Ganiatsas, G.; Fotiadis, D.I. A user independent, biosignal based, emotion recognition method. In *International Conference on User Modeling*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 314–318.
61. Ali, M.; Al Machot, F.; Mosa, A.H.; Kyamakya, K. Cnn based subject-independent driver emotion recognition system involving physiological signals for adas. In *Advanced Microsystems for Automotive Applications 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 125–138.
62. Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; Arunkumar, N. Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *IEEE Access* **2018**, *7*, 57–67. [CrossRef]
63. Suhara, Y.; Xu, Y.; Pentland, A. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 715–724.
64. Zhang, T. Multi-modal Fusion Methods for Robust Emotion Recognition using Body-worn Physiological Sensors in Mobile Environments. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 463–467.
65. Tkalcic, M.; Odic, A.; Kosir, A.; Tasic, J. Affective labeling in a content-based recommender system for images. *IEEE Trans. Multimed.* **2012**, *15*, 391–400. [CrossRef]
66. Chang, C.Y.; Zheng, J.Y.; Wang, C.J. Based on support vector regression for emotion recognition using physiological signals. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–7.
67. Hassanien, A.E.; Kilany, M.; Houssein, E.H.; AlQaheri, H. Intelligent human emotion recognition based on elephant herding optimization tuned support vector regression. *Biomed. Signal Process. Control* **2018**, *45*, 182–191. [CrossRef]
68. Wei, J.; Chen, T.; Liu, G.; Yang, J. Higher-order multivariable polynomial regression to estimate human affective states. *Sci. Rep.* **2016**, *6*, 23384. [CrossRef] [PubMed]
69. Nicolaou, M.A.; Gunes, H.; Pantic, M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.* **2011**, *2*, 92–105. [CrossRef]
70. Romeo, L.; Cavallo, A.; Pepa, L.; Berthouze, N.; Pontil, M. Multiple Instance Learning for Emotion Recognition using Physiological Signals. *IEEE Trans. Affect. Comput.* **2019**. [CrossRef]
71. Gibson, J.; Katsamanis, A.; Romero, F.; Xiao, B.; Georgiou, P.; Narayanan, S. Multiple instance learning for behavioral coding. *IEEE Trans. Affect. Comput.* **2015**, *8*, 81–94. [CrossRef]
72. Lee, C.C.; Katsamanis, A.; Black, M.P.; Baucom, B.R.; Georgiou, P.G.; Narayanan, S.S. Affective state recognition in married couples' interactions using PCA-based vocal entrainment measures with multiple instance learning. In *International Conference on Affective Computing and Intelligent Interaction*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 31–41.
73. Wu, B.; Zhong, E.; Horner, A.; Yang, Q. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In Proceedings of the 22nd ACM International Conference on Multimedia, Mountain View, CA, USA, 18–19 June 2014; pp. 117–126.
74. Maron, O.; Lozano-Pérez, T. A framework for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **1997**, *10*, 570–576.
75. Fernandez, E.; Gangitano, C.; Del Fà, A.; Sangiacomo, C.O.; Talamonti, G.; Draicchio, F.; Sbriccoli, A. Oculomotor nerve regeneration in rats: Functional, histological, and neuroanatomical studies. *J. Neurosurg.* **1987**, *67*, 428–437. [CrossRef]
76. Ibbotson, M.R.; Crowder, N.A.; Cloherty, S.L.; Price, N.S.; Mustari, M.J. Saccadic modulation of neural responses: Possible roles in saccadic suppression, enhancement, and time compression. *J. Neurosci.* **2008**, *28*, 10952–10960. [CrossRef]
77. Picard, R.W. Future affective technology for autism and emotion communication. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 3575–3584. [CrossRef] [PubMed]
78. Greaney, J.L.; Kenney, W.L.; Alexander, L.M. Sympathetic regulation during thermal stress in human aging and disease. *Auton. Neurosci.* **2016**, *196*, 81–90. [CrossRef]
79. Chen, M.; Shi, X.; Zhang, Y.; Wu, D.; Guizani, M. Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Trans. Big Data* **2017**. [CrossRef]
80. Creswell, A.; Arulkumaran, K.; Bharath, A.A. On denoising autoencoders trained to minimise binary cross-entropy. *arXiv* **2017**, arXiv:1708.08487.
81. Ap, S.C.; Lauly, S.; Larochelle, H.; Khapra, M.; Ravindran, B.; Raykar, V.C.; Saha, A. An autoencoder approach to learning bilingual word representations. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1853–1861.

82. Zhang, T.; El Ali, A.; Wang, C.; Zhu, X.; Cesar, P. CorrFeat: Correlation-based Feature Extraction Algorithm using Skin Conductance and Pupil Diameter for Emotion Recognition. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 404–408.
83. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; pp. 1247–1255.
84. Chen, C.P.; Liu, Z. Broad learning system: An effective and efficient incremental learning system without the need for deep architecture. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 10–24. [CrossRef]
85. Movahedi, F.; Coyle, J.L.; Sejdić, E. Deep belief networks for electroencephalography: A review of recent contributions and future outlooks. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 642–652. [CrossRef]
86. Liu, C.; Tang, T.; Lv, K.; Wang, M. Multi-Feature Based Emotion Recognition for Video Clips. In Proceedings of the ACM 2018 on International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 630–634.
87. Chen, H.; Jiang, B.; Ding, S.X. A Broad Learning Aided Data-Driven Framework of Fast Fault Diagnosis for High-Speed Trains. *IEEE Intell. Transp. Syst. Mag.* **2020**. [CrossRef]
88. Jain, A.; Nandakumar, K.; Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognit.* **2005**, *38*, 2270–2285. [CrossRef]
89. Olshausen, B.A.; Field, D.J. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vis. Res.* **1997**, *37*, 3311–3325. [CrossRef]
90. Hewig, J.; Hagemann, D.; Seifert, J.; Gollwitzer, M.; Naumann, E.; Bartussek, D. A revised film set for the induction of basic emotions. *Cogn. Emot.* **2005**, *19*, 1095. [CrossRef]
91. Bartolini, E.E. *Eliciting Emotion with Film: Development of a Stimulus Set*; Wesleyan University: Middletown, CT, USA, 2011.
92. Park, C.Y.; Cha, N.; Kang, S.; Kim, A.; Khandoker, A.H.; Hadjileontiadis, L.; Oh, A.; Jeong, Y.; Lee, U. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *arXiv* **2020**, arXiv:2005.04120.
93. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N. DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Trans. Affect. Comput.* **2015**, *6*, 209–222. [CrossRef]
94. Lin, T.T.; Chiu, C. Investigating adopter categories and determinants affecting the adoption of mobile television in China. *China Media Res.* **2014**, *10*, 74–87.
95. McNally, J.; Harrington, B. How Millennials and Teens Consume Mobile Video. In Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17), Hilversum, The Netherlands, 14–16 June 2017; ACM: New York, NY, USA, 2017; pp. 31–39. [CrossRef]
96. O'Hara, K.; Mitchell, A.S.; Vorbau, A. Consuming Video on Mobile Devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07), San Jose, CA, USA, 28 April–3 May 2007; ACM: New York, NY, USA, 2007; pp. 857–866. [CrossRef]
97. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **2012**, *3*, 42–55. [CrossRef]
98. Ferdinando, H.; Seppänen, T.; Alasaarela, E. Enhancing Emotion Recognition from ECG Signals using Supervised Dimensionality Reduction. In Proceedings of the ICPRAM, Porto, Portugal, 24–26 February 2017; pp. 112–118.
99. Gui, D.; Zhong, S.H.; Ming, Z. Implicit Affective Video Tagging Using Pupillary Response. In *International Conference on Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 165–176.
100. Olson, D.H.; Russell, C.S.; Sprenkle, D.H. *Circumplex Model: Systemic Assessment and Treatment of Families*; Psychology Press: Hove, UK, 1989.
101. Itten, J. *Mein Vorkurs am Bauhaus*; Otto Maier Verlag: Ravensburg, Germany, 1963.
102. Schmidt, P.; Reiss, A.; Dürichen, R.; Van Laerhoven, K. Labelling Affective States "in the Wild" Practical Guidelines and Lessons Learned. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018; pp. 654–659.
103. Zhao, B.; Wang, Z.; Yu, Z.; Guo, B. EmotionSense: Emotion recognition based on wearable wristband. In Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 8–12 October 2018; pp. 346–355.
104. Meijering, E. A chronology of interpolation: From ancient astronomy to modern signal and image processing. *Proc. IEEE* **2002**, *90*, 319–342. [CrossRef]
105. Daniels, R.W. *Approximation Methods for Electronic Filter Design: With Applications to Passive, Active, and Digital Networks*; McGraw-Hill: New York, NY, USA, 1974.
106. Fleureau, J.; Guillotel, P.; Orlac, I. Affective benchmarking of movies based on the physiological responses of a real audience. In Proceedings of the IEEE 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 73–78.
107. Chu, Y.; Zhao, X.; Han, J.; Su, Y. Physiological signal-based method for measurement of pain intensity. *Front. Neurosci.* **2017**, *11*, 279. [CrossRef] [PubMed]
108. Karthikeyan, P.; Murugappan, M.; Yaacob, S. Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. *J. Phys. Ther. Sci.* **2012**, *24*, 1341–1344. [CrossRef]

109. Zeiler, M.D. Adadelta: An adaptive learning rate method. *arXiv* **2012**, arXiv:1212.5701.
110. Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69.
111. Chinchor, N. MUC-3 evaluation metrics. In *Proceedings of the 3rd Conference on Message Understanding*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1991; pp. 17–24.
112. Fatourechi, M.; Ward, R.K.; Mason, S.G.; Huggins, J.; Schlögl, A.; Birch, G.E. Comparison of evaluation metrics in classification applications with imbalanced datasets. In Proceedings of the IEEE 2008 Seventh International Conference on Machine Learning and Applications, San Diego, CA, USA, 11–13 December 2008; pp. 777–782.
113. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
114. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
115. Wickramasuriya, D.S.; Faghih, R.T. Online and offline anger detection via electromyography analysis. In Proceedings of the 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), Bethesda, MD, USA, 6–8 November 2017; pp. 52–55.
116. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [CrossRef]
117. Kukolja, D.; Popović, S.; Horvat, M.; Kovač, B.; Ćosić, K. Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *Int. J. Hum. Comput. Stud.* **2014**, *72*, 717–727. [CrossRef]
118. Critchley, H.D.; Garfinkel, S.N. Interoception and emotion. *Curr. Opin. Psychol.* **2017**, *17*, 7–14. [CrossRef] [PubMed]
119. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
120. Ed-doughmi, Y.; Idrissi, N. Driver fatigue detection using recurrent neural networks. In Proceedings of the 2nd International Conference on Networking, Information Systems & Security, Sochi, Russia, 12–15 September 2019; pp. 1–6.
121. Lou, J.; Wang, Y.; Nduka, C.; Hamedi, M.; Mavridou, I.; Wang, F.Y.; Yu, H. Realistic facial expression reconstruction for VR HMD users. *IEEE Trans. Multimed.* **2019**, *22*, 730–743. [CrossRef]
122. Genç, Ç.; Colley, A.; Löchtefeld, M.; Häkkilä, J. Face mask design to mitigate facial expression occlusion. In Proceedings of the 2020 International Symposium on Wearable Computers, Cancun, Mexico, 14–17 September 2020; pp. 40–44.
123. Oulefki, A.; Aouache, M.; Bengherabi, M. Low-Light Face Image Enhancement Based on Dynamic Face Part Selection. In *Iberian Conference on Pattern Recognition and Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 86–97.