# ON FINE-GRAINED TEMPORAL EMOTION RECOGNITION IN VIDEO

## HOW TO TRADE OFF RECOGNITION ACCURACY WITH ANNOTATION COMPLEXITY?

# ON FINE-GRAINED TEMPORAL EMOTION RECOGNITION IN VIDEO

## HOW TO TRADE OFF RECOGNITION ACCURACY WITH ANNOTATION COMPLEXITY?

## Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. T. H. J. J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
maandag 3 oktober 2022 om 12:30 uur

door

**Tianyi ZHANG**

Master of Science in Control Engineering,
Nanjing University of Aeronautics and Astronautics, Nanjing, China,
geboren te Nanjing, China.

Dit proefschrift is goedgekeurd door de promotoren.
promotor: prof. dr. P. Cesar
promotor: prof. dr. A. Hanjalic

Samenstelling promotiecommissie:

| | |
|---|---|
| Rector Magnificus, | voorzitter<br>Technische Universiteit Delft |
| *Onafhankelijke leden:* | |
| Prof. dr. G. J. P. M. Houben, | Technische Universiteit Delft |
| Prof. dr. A. Bozzon, | Technische Universiteit Delft |
| Prof. dr. M. Worring, | Universiteit van Amsterdam |
| Prof. dr. A. A. Salah, | Universiteit Utrecht |
| Dr. H. S. Hung, | Technische Universiteit Delft |

Dr. A. El Ali heeft in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

*Keywords:* Emotion Recognition, Physiological Signals, Machine Learning, Video Watching

*Printed by:* Gildeprint

ISBN 978-94-6384-376-8

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

*This thesis is dedicated to my parents and grandmother*
*who always believe I can make it during my PhD period*

# CONTENTS

# Summary

Fine-grained emotion recognition is the process of automatically identifying the emotions of users at a fine granularity level, typically in the time intervals of 0.5s to 4s according to the expected duration of emotions. Previous work mainly focused on developing algorithms to recognize only one emotion for a video based on the user feedback after watching the video. These methods are known as post-stimuli emotion recognition. Compared to post-stimuli emotion recognition, fine-grained emotion recognition can provide segment-by-segment prediction results, making it possible to capture the temporal dynamics of users' emotions when watching videos. The recognition result it provides can be aligned with the video content and tell us which specific content in the video evokes which emotions. Most of the previous works on fine-grained emotion recognition require fine-grained emotion labels to train the recognition algorithm. However, the experiments to collect these fine-grained emotion labels are usually costly and time-consuming. Thus, this thesis focuses on investigating whether we can accurately predict the emotions of users at a fine granularity level with only a limited amount of emotion ground truth labels for training.

We start our technical contribution in Chapter 3 by building up the baseline methods which are trained using fine-grained emotion labels. This can help us understand how accurate the recognition can be if we take advantage of the fine-grained emotion labels. 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 physiological signals. *CorrNet* extracts features both inside each fine-grained signal segment (instance) and between different instances for the same video stimuli (correlation-based features). We found out that, compared to sequential learning, correlation-based instance learning offers advantages of higher recognition accuracy, less overfitting and less computational complexity.

Compared to collecting fine-grained emotion labels, it is easier to collect only one emotion label after the user watched that stimulus (i.e., the post-stimuli emotion labels). Therefore, in the second technical chapter (Chapter 4) of the thesis, we investigate whether the emotions can be recognized at a fine granularity level by training with only post-stimuli emotion labels (i.e., labels users annotated after watching videos), and propose an Emotion recognition algorithm based on Deep Multiple Instance Learning (*EDMIL*). *EDMIL* recognizes fine- grained valence and arousal (V-A) labels by identifying which instances represent the post-stimuli V-A annotated by users after watching the videos. Instead of fully-supervised training, the instances are weakly-supervised by the post-stimuli labels in the training stage. Our experiments show that weakly supervised learning can reduce overfitting caused by the temporal mismatch between fine-grained annotations and input signals.

Although the weakly-supervised learning algorithm developed in Chapter 4 can obtain accurate recognition results with only few annotations, it can only identify the annotated (post-stimuli) emotion from the baseline emotion (e.g., neutral) because only post-stimuli labels are used for training. The non-annotated emotions are all categorized as part of the baseline. To overcome this, in Chapter 5, we propose an Emotion recognition algorithm based on Deep Siamese Networks (EmoDSN). *EmoDSN* recognizes fine-grained valence and arousal (V-A) labels by maximizing the distance metric between signal segments with different V-A labels. According to the experiments we run in this chapter, *EmoDSN* achieves promising results by using only 5 shots (5 samples in each emotion category) of training data.

Reflecting on the achievements reported in this thesis, we conclude that the fully-supervised algorithm (Chapter 3) can result in more accurate fine-grained emotion recognition results if the annotation quantity is sufficient. The weakly-supervised learning method (Chapter 4) can result in better recognition results at the instance level compared to fully-supervised methods. We also found that the weakly-supervised learning methods can perform the best if users annotate their most salient, but short emotions or their overall and longer-duration (i.e., persisting) emotions. The few-shot learning method (Chapter 5) can obtain more emotion categories (more than the weakly-supervised learning) by using less amount of samples for training (better than the fully-supervised learning). However, the limitation of it is that accurate recognition results can only be achieved by a subject-dependent model.

# 1

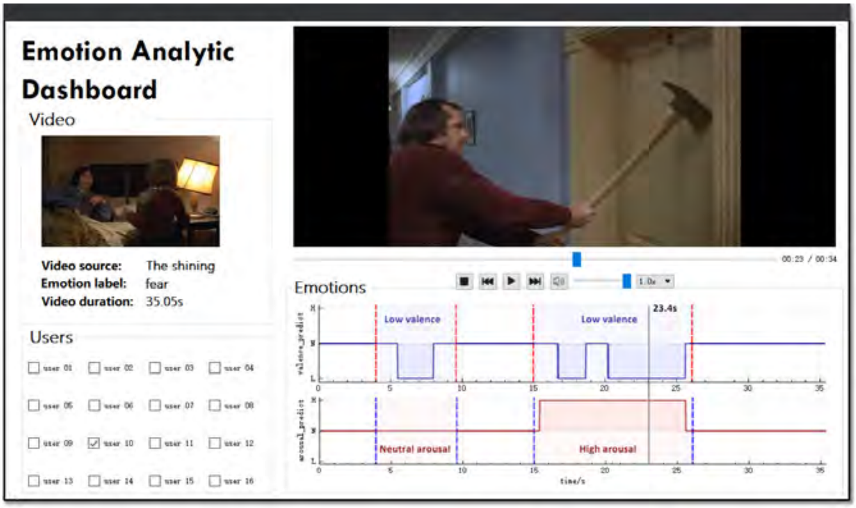
# Introduction

## 1.1. BACKGROUND AND MOTIVATION

Emotions are manifested in all segments of our lives. It has been shown that emotions influence various cognitive processes, including perception and organization of memory [1], goal generation, evaluation, and decision-making [2]. Emotions also play an important role in users' selection and consumption of video content [3]. According to the *gratification theory* [4], emotions influence the video selection either directly, by being gratifying experiences per se, or indirectly, by contributing to cognitive and social needs [5]. It was found by Zillmann [6] that users prefer to select movies and videos that are expected to maintain or maximize their pleasure states. Other researchers [4, 7] found that affective states corresponding to negative feeling states such as sadness can also gratify users because the sad scenes can induce empathy towards characters. Thus, different users actively seek out different media content eliciting different affective states to satisfy their emotional needs.

To help video providers (e.g., movie or advertising producers, online tutors) better understand their users' emotions towards the video content they make, an emotion *analytics dashboard* could be useful. For online tutors, a dashboard which shows students' emotions towards their online courses could help them to identify which parts of their courses are interesting or boring for students. Based on that, they could adjust the content of online courses to promote the engagement and learning experience of students [8]. Similarly, by pinpointing the moments which trigger negative emotions (e.g., discomfortable ), the emotion analytics dashboard could help filmmakers to improve the story arch of the narrative for their films and delete or adjust the scenes that make the audience distracted or confused. To realize that, video providers could invite users that represent their target customers to their lab for a test screening when making a new video (e.g., movie, advertisement, online course). Users would be equipped with sensors to measure their physiological signals (e.g., Blood Volume Pulse (BVP), Skin Temperature, Electrodermal Activity (EDA)). These signals would then be sent to the servers of the video provider and used to predict the emotions of the users by the content of different video segments. At last, the obtained emotions would be aligned with the video content and visualized on the emotion dashboard for the video providers to analyze the relationship between video content and the emotions of users and help them inform their subsequent video production steps.

Building the emotion analytics dashboard requires acquiring the emotions of users. Most previous works just recognize one overarching (average, dominant) emotion per video stimulus (i.e., post-stimuli emotion recognition) [9, 10], not capturing the time-varying nature of human emotions [11, 12]: users could reveal multiple emotions over time while watching a single video. This is illustrated by an example of the emotion *analytics dashboard* shown in Figure 1.1. Although the overall rating for this video is fear, the user watching the video reveals that emotion only in the time interval from 15.2s to 26.7s, as indicated by low valence and high arousal. The video providers (e.g., movie and ads producers) need to obtain the emotions of users aligned with the specific video content to analyze the relationship between video content and users' emotions. Fine-grained emotion recognition is defined as recognizing the emotions of users at a fine granularity level, typically in the intervals of 0.5s to 4s according to the expected duration of emotions [13, 14]. Fine-grained emotion recognition can therefore provide



Emotion Analytic Dashboard

Video

Video source: The shining
Emotion label: fear
Video duration: 35.05s

Users

□ user 01 □ user 02 □ user 03 □ user 04
□ user 05 □ user 06 □ user 07 □ user 08
□ user 09 ☑ user 10 □ user 11 □ user 12
□ user 13 □ user 14 □ user 15 □ user 16

Figure 1.1: An emotion dashboard used for evaluating the personalized watching experience of a user: his or her emotion is at low *valence* and high *arousal* when the man in the movie is smashing the door with an ax

more precise recognition results for video providers to better understand and analyze the relationship between users' emotions and video content over time.

Machine learning algorithms, such as artificial neural networks, can automatically learn the non-linear and high dimensional mapping between the input signals and ground truth labels, which makes them powerful tools for emotion recognition [9, 15]. Most of the previous works [12, 16, 17] on fine-grained emotion recognition using machine learning methods rely on large amounts of annotated signals to train an accurate recognition model. This means that the input signals are required to also be annotated at a fine granularity level [18]. To collect such fine-grained emotion labels (e.g., valence and arousal), researchers either ask users themselves to label their emotions in real-time while watching videos [19, 20] or invite professional annotators (e.g., trained clinical psychologists) to label users' emotions segment-by-segment (e.g., using videos of users' facial expressions [21]) after watching videos [21–23]. However, both approaches come with significant challenges. Asking users to momentarily self-report their emotions can incur unacceptable mental workload and result in user fatigue. In the second approach, at least three external annotators are usually required to get a meaningful annotator agreement (e.g., high kappa score) [21, 24, 25]. Thus, the experiments to collect these ground-truth labels are usually costly and time-consuming and therefore a large extra burden and investment for video providers.

In view of the above, the main challenge for the thesis is to find a trade-off between emotion recognition accuracy at a fine level of granularity and the amount of annotation used for training the algorithms. To overcome this challenge, this thesis proposes new scientific concepts demonstrated by machine-learning algorithms that use weakly-supervised learning and few-shot learning to obtain fine-grained emotions of users. This will directly feed into an emotion analytics dashboard helping video providers understand the users' affective response to the media content they provide.

## 1.2. BASIC CONCEPTS

To recognize emotions at a fine-level of granularity using machine learning techniques, three basic concepts need to be defined: 1) the ***emotion model*** to define emotions, 2) the ***input modalities/signals*** to train the machine learning models and 3) the ***ground truth labels*** to validate the designed models. The *emotion model* and *ground truth labels* determine which kind of prediction this thesis could present to help the video providers to better understand their users. The input modalities/signals determine whether these output/predictions can be obtained in a *reliable* manner. Below, we describe and discuss each of them in more detail.

### 1.2.1. EMOTION MODEL

Common affective concepts typically include affect, emotion, mood and feeling. According to Gross et at. [26], affect can be viewed as the superordinate category which contains emotion, feeling and mood. In neuroscience [27], emotion often refers to a mental state that arises spontaneously. It is often accompanied by physical and physiological changes that are relevant to the human organs and tissues such as heart, skin, blood flow and muscle. Feeling and mood however, last longer than emotions and are often fueled by a mix of emotions. One of the biggest differences between feeling, mood and emotion is that the memories (which differ across the individuals) play a significant role in the formation of feeling and mood [9], while emotions have stronger links with specific stimuli.

To define or model emotions, researchers primarily use two kinds of models [9]: *categorical* and *dimensional* emotion models. Categorical emotion models divide emotions into discrete categories such as happy, sad, anger, fear, surprise, and disgust. For example, Ekman [28] summarizes *six basic emotions*: happy, sad, anger, fear, surprise, and disgust, and viewed the other emotions as the production of reactions and combinations of these basic emotions. Dimensional models, also known as continuous emotion models, describe emotions using a multi-dimensional space. For example, the classic *Russell's Circumplex Model of Emotions* [29] describes emotions in a two-dimensional circular space: valence and arousal. An additional dimension of *dominance* was later introduced to allow differentiation of the emotional states showing the same levels of arousal and valence based on the degree of control exerted by a stimulus (e.g., rage versus grief) [30].

Compared to discrete models, the dimensional models introduce continuous variables to describe emotions, which enables them to model emotions at a finer level of granularity [31, 32]. Although *dominance* is an important dimension to describe emotions, some researchers [33, 34] argue that it can be considered redundant and seen as a consequence of the *core affect* (valence and arousal, defined by *Russell* [35]). Previous work on affective video content analysis [31, 36] also show that the effect of the dominance dimension becomes visible only at points with distinctly high absolute valence values. Consequently, dominance plays only a limited role in characterizing various emotional states. We therefore focus on valence and arousal to describe emotions in this thesis.

### 1.2.2. INPUT MODALITIES

Derived from traditional emotion theories, researchers summarize the input to recognize emotions into two categories [37]: expressions and embodiments. The expressions used for recognizing emotions include facial expressions, voice, body language and gestures. The embodiment theories view emotions as physiological behaviors of the human neural systems [9]. According to *James-Lange* theory [38] and *Cannon's* theory [39], we feel emotions and experience physiological reactions, such as sweating, trembling, and muscle tension, simultaneously. Unlike the visual and vocal expressions, physiological signals (e.g, Blood Volume Pulse (BVP), Skin Temperature (SKT), Electrodermal Activity (EDA)) measured from the human neural systems are largely involuntarily activated, which makes it more objective to recognize emotions [9]. We choose physiological signals as input modalities due to the following reasons:

- Previous works on neuroscience and emotions showed that physiological signals and emotions have strong correlations [40–42]. For example, when the arousal of the user increases, the heart rate (HR) and EDA of the user will also increase. When the valence of the user changes from positive to negative (or from high to low, depending on the model), the heart rate variability (HRV) decreases [9, 41].

- Compared to the expressions, physiological signals are better observable and measurable by non-intrusive and wearable devices. For example, while a camera is needed for capturing facial expressions, the viewport capture from the front-facing camera does not always capture the full face in an outdoor, mobile environment when a user is on the move [43].

### 1.2.3. GROUND TRUTH LABELS

For acquiring the ground truth labels used to validate fine-grained emotion recognition algorithms, the common practice in the previous work was to either ask external annotators to continuously annotate users' emotions [24], or to use the continuous self-reports from users themselves [44]. Compared to self-reports, external annotators can misidentify some emotions. For example, according to the experiments of Song et al. [45] and Abdic et al. [46], negative valence is often misidentified by external annotators as positive when users smile because of sarcasm and frustration. According to the *appraisal theory* [47], our emotions can be represented by our appraisals or estimations of events or stimuli. The appraisals, which can be collected by self-reports from users, are directly linked with the subjective experience of individuals towards specific events or stimuli [40]. This thesis focuses on recognizing the personalized experience of users watching videos. Therefore, we used the momentary self-reports of valence and arousal from users as the emotion ground truth labels to validate the machine learning models for emotion recognition we designed.

## 1.3. THESIS OBJECTIVE AND SCOPE

The *objective* of this thesis is to propose new scientific concepts and their algorithmic implementation to find a trade-off between emotion recognition accuracy at a fine level of granularity and the amount of annotation used for training the algorithms. The *scope* of the thesis is limited to personalized fine-grained emotion recognition for video watching, by relying on physiological signals and machine learning algorithms. While the proposed concepts and algorithms are expected to feed into the development of emotion analytics dashboards for video providers, the research of how to adapt these algorithms for an emotion analytics dashboard is beyond the scope of this thesis. The research of emotion recognition for other media content (e.g., music) or using different modalities (e.g., facial expressions, video content) is also beyond the scope of the thesis.

## 1.4. RESEARCH QUESTIONS

Fine-grained emotion recognition requires algorithms to predict emotion states by relying on signals within a specific time interval. This is typically done using two kinds of methods: regression and classification. Regression methods view the target emotion states as a continuous sequence and directly calculate the mapping (regression) from input signals to output emotion sequences. Classification methods on the other hand first divide the entire signals into multiple segmentations (instances) and classify the emotion for each fine-grained instance. Both the regression and classification methods need fine-grained emotion labels to train the recognition algorithm. However, the experiments to collect these fine-grained emotion labels are usually costly and time-consuming. To collect these labels, researchers developed momentary emotion annotation tools, such as *FEELTrace* [48], *CASE* [49], *RCEA* [19] and *RCEA-360VR* [50] which allow users to input their emotions (e.g., valence and arousal) in real-time. However, momentary annotation requires users to multi-task (e.g., watch videos and annotate at the same time), which poses risks in increasing user mental workload [19, 51] and is not always feasible if users watch longer videos [52].

Therefore, the main challenge of the thesis is to explore whether we can accurately predict the emotions of users at a fine granularity level with minimum labels (known as *the problem of small data* for emotion recognition [53–55]). To answer this question, we need to first understand how accurate the recognition can be if we take advantage of the fine-grained emotion labels (**Chapter 3**). Although previous works have explored this question using deep learning models, these methods still have the problem of overfitting on the training data when there is temporal resolution mismatch (due to the different interoception levels across individuals) between physiological signals and fine-grained self-reports [56]. This leads to our first research question:

**Research question 1: How can we recognize emotions at a fine granularity level by taking advantage of the fine-grained emotion labels?**

**RQ 1.1:** *Can correlation-based instance learning offer advantages for fine-grained emotion recognition compared to sequence learning?*

Although previous works [12, 16, 57] provide useful insights on fine-grained emotion recognition using fine-grained labels, there are still two challenges which lead to low recognition accuracy. The first challenge is that the information inside each segment of the signals (instances) may not be sufficient for recognizing emotions. To address that challenge, sequence learning methods such as Long Short Term Memory (LSTM) networks have been deployed to extract the temporal information between different samples or instances as additional features for recognition [58–60]. However, the use of sequence learning methods leads to another challenge: the temporal resolution mismatch between physiological signals and fine-grained labels could lead to the problem of overfitting for sequence learning [61, 62]. The recurrent learning structure used by sequential learning approaches could cause the accumulation of error when the network is trained with mismatched signals and labels.

Previous work [63, 64] found that the same video stimulus could trigger relatively similar valence and arousal among different users. Thus, the emotion self-reports of the users watching the same videos are correlated. Inspired by this idea, we explore whether the correlation-based instance learning offer advantages for reducing the problem of overfitting when fine-grained emotion labels are used for training. Specifically, we explore whether the *correlation-based instance learning* can overcome the two above-mentioned challenges by extracting features between signal segments using correlation-based features. To achieve this, we propose a correlation-based recognition algorithm (*CorrNet*) in **Chapter 3**, which uses unsupervised learning to learn the features both inside and between different instances, and a fully-supervised classifier to recognize the emotions for each of them. We compare the recognition results with state-of-the-art machine learning and deep learning (include sequential learning) methods to find out whether it can offer advantages for fine-grained emotion recognition by providing more accurate recognition results. Comparing the performance of correlation-based instance learning with other classic deep learning approaches could help us build baselines about how accurate the recognition can be by taking advantage of the fine-grained emotion labels.

**RQ 1.2:** *Does correlation-based instance learning provide imbalanced prediction accuracies for different emotion categories?*

The datasets which contain fine-grained emotion labels usually have an imbalanced distribution of samples for different emotion categories [20, 65]. When users annotate their emotions, they tend to annotate them as neutral by default and non-neutral only for specific scenes (e.g., kissing scenes for happy). Thus, the training data could contain a large number of samples labeled as neutral and only a small amount of data labeled as other emotions. We investigate whether the sample imbalance will influence the prediction accuracies for different emotion categories in **Chapter 3**. The study of this research question helps us to understand the limitation of recognition algorithms trained with fine-grained labels.

After building up the baselines, we start with investigating whether the emotions can be recognized at a fine granularity level by training with only post-stimuli emotion labels (**Chapter 4**). Post-stimuli emotion labels are the labels users annotate after watching the videos. Training the recognition system using only post-stimuli labels is easy for both developers and users: developers do not need to collect a large amount of fine-grained emotion self-reports and users only need to input their emotions once after they watch the videos. However, when training machine learning algorithms to recognize fine-grained emotions using post-stimuli labels, the information on which fine-grained instances represent the emotion annotated by the users is missing. According to the *peak-end theory* [66], the post-stimuli labels represent only the most salient (peak) or recent (end) emotion while watching a video, rather than the naturally dynamic and subtle emotional changes that may occur within it. This is defined as the problem of *temporal ambiguity* of emotion recognition by Romeo et al. [52]. This problem can lead to overfitting [67, 68] if all the instances are fully-supervised by the post-stimuli labels. To address this, we define the following research question:

**Research question 2: How to utilize physiological signals with only post-stimuli emotion labels to train a fine-grained emotion recognition system?**

**RQ 2.1:** *Does weakly-supervised multiple instance learning solve the problem of time ambiguity if only post-stimuli emotion labels are available for training?*

Addressing the problem of time-ambiguity requires networks to learn the probability for instances to correlate the corresponding post-stimuli labels. In the paradigm of Multiple Instance Learning (MIL), the input is a set of *bags* which are composed of multiple *instances*. At the training stage, each bag has a corresponding label while each instance does not. MIL assigns each instance a matching score (instance gain) which maximizes the probability to predict the bag label [69]. Unlike fully-supervised learning which learns the precise mapping from input signal segments and segment labels, the MIL uses weakly-supervised learning to model the probability of each signal segment corresponding to the one label of the entire signal. Inspired by the idea of MIL, we want to find out whether the instance gains learned by MIL can represent the probability for instances to predict the post-stimuli emotion labels. To achieve this, we propose a Deep MIL based Emotion recognition algorithm (*EDMIL*) in **Chapter 4**. Instead of implementing fully-supervised training for all the instances using post-stimuli labels, the instances are weakly supervised by the post-stimuli labels. That will enable the recognition algorithm to estimate the emotion label for each instance without learning the prior relationship between fine-grained emotion labels and instances.

**RQ 2.2:** *Which feature extraction methods are the most effective for emotion recognition trained with only post-stimuli emotion labels?*

Theoretically, the feature extraction methods with end-to-end models should result in the best performance as the deep representation is trained to best recognize these labels. However, according to our findings when answering research question 1, if we train the network using fine-grained emotion labels and fully-supervised learning methods, the end-to-end model will overfit because of the temporal resolution mismatch between physiological signals and fine-grained emotion labels due to different interoception levels across individuals [56]. To answer this research question, we compare three kinds of feature extraction methods in **Chapter 4**: (1) an end-to-end feature extraction method using 1D-CNN, (2) an unsupervised feature extraction method developed in research question 1 and (3) a manual feature extraction method which is widely used by previous works as baselines for emotion recognition [9, 10, 70]. The comparison can help us find out whether the end-to-end, deep feature extraction still has the problem of overfitting for weakly-supervised learning.

**RQ 2.3:** *What are the advantages and disadvantages of recognizing emotions using weakly-supervised training by post-stimuli labels compared with fully-supervised training with fine-grained labels?*

Unlike the fully-supervised learning methods which learn the instance-label relationship by building the direct mapping between instances and fine-grained emotion labels, *EDMIL* estimates the emotion label for each instance by the matching scores of instances and post-stimuli labels. Thus, we compare the results between training with post-stimuli labels (weakly-supervised) and fine-grained labels (fully-supervised) in **Chapter 4**. The comparison can help us understand whether the fine-grained (i.e., instance-level) labels can improve or compromise the recognition performance.

To discuss the performance differences, we use both the recognition accuracy and dynamic time warping distance (DTW) [71] as evaluation metrics. Compared to accuracy, DTW is less sensitive to the time-shift of specific values in the sequence, which can better evaluate the performance of recognizing emotions of the whole video instead of individual instances. Answering this research question will help us understand the pros and cons of recognizing emotions using weakly-supervised training by post-stimuli labels compared to baseline methods in Chapter 3.

Although the recognition system trained with post-stimuli emotion labels requires less annotation, it can only identify the annotated (post-stimuli) emotion from the baseline emotion (e.g., neutral) because only post-stimuli labels are used for training. Thus, it can only recognize one emotion, except for the baseline emotion (e.g., neutral) for one video watching. If there is more than one emotion except for the baselines, *EDMIL* will fail to recognize it. The unlabeled emotions are all categorized as part of the baseline. In addition, developers still need to collect a large amount of data for training. Thus, we want to investigate whether we can achieve accurate prediction and recognize multiple emotion categories trained with only few annotated samples (e.g., less than 10 for each emotion category) (**Chapter 5**). This brings us to the following research question:

**Research question 3: What are the effective ways to recognize emotions at a fine granularity level when only few signal segments are annotated?**

**RQ 3.1:** *Do deep siamese networks enable learning useful representation from few annotated signal segments that are competitive with other few-shot learning counterparts for fine-grained emotion recognition?*

Recently, *Few-shot learning* (FSL) networks have been proposed to learn the representation models from only few annotated samples [72]. The uniqueness of FSL networks is that they can converge on a small amount of training data and provide relatively accurate prediction results. Siamese network is a kind of FSL algorithm which learns the difference between samples with different labels [73]. It learns to maximize the distance metric between samples with different labels, which enables the network to converge rapidly on a small number of training samples. To learn the representation of emotions using few annotated samples, researchers need to design different embedding networks for different data modalities. Although previous works provide useful insights on how to design embedding networks for FSL for other data modalities such as images [74], there is still limited amount of research on how to design embedding networks for physiological signals. Thus, we want to find out whether the representation (embeddings) learned by siamese network can discriminate signal segments from different emotion categories.

To achieve this, we propose a Deep Siamese Network based Emotion recognition algorithm (*EmoDSN*) in **Chapter 5**. Compared to other FSL algorithms, *EmoDSN* uses the pair-by-pair learning structure (learn the difference between two samples in two categories) instead of using the one-to-many learning structure (learn the difference between one sample and samples in other categories). By answering this research question, we want to find out whether the pair-by-pair learning structure is more suitable compared to the one-to-many learning structure for fine-grained emotion recognition using physiological signals.

**RQ 3.2:** *Do the temporal moments for selecting training samples affect the performance of deep siamese network based fine-grained emotion recognition?*

Since we use only few annotated samples to train the network, we investigate samples from which temporal moments can better represent the distribution for the whole video watching and result in better recognition results (in **Chapter 5**). Specifically, we choose the beginning, ending and the changing points of video watching and compare the results using signal segments from these temporal moments as training samples. We also want to investigate the amount of samples *EmoDSN* needs to obtain accurate recognition when selecting training samples at different temporal moments of video watching. Answering this research question enable the possibility of helping researchers maximize the recognition performance and minimize the amount of training samples by asking users to annotate at the most suitable temporal moments while watching a video.

## 1.5. THESIS OUTLINE AND CONTRIBUTIONS

The thesis is composed of six chapters. Following this introductory chapter, and **Chapter 2** presenting the related previous work and the datasets used in the thesis, we set out to answer the research questions 1, 2 and 3 in **Chapter 3, 4** and **5**, respectively. Figure 1.2 shows the relationships between the three research questions and corresponding chapters of the thesis. Below, we outline the main contribution per chapter and link it to the publications the presented work is based on.

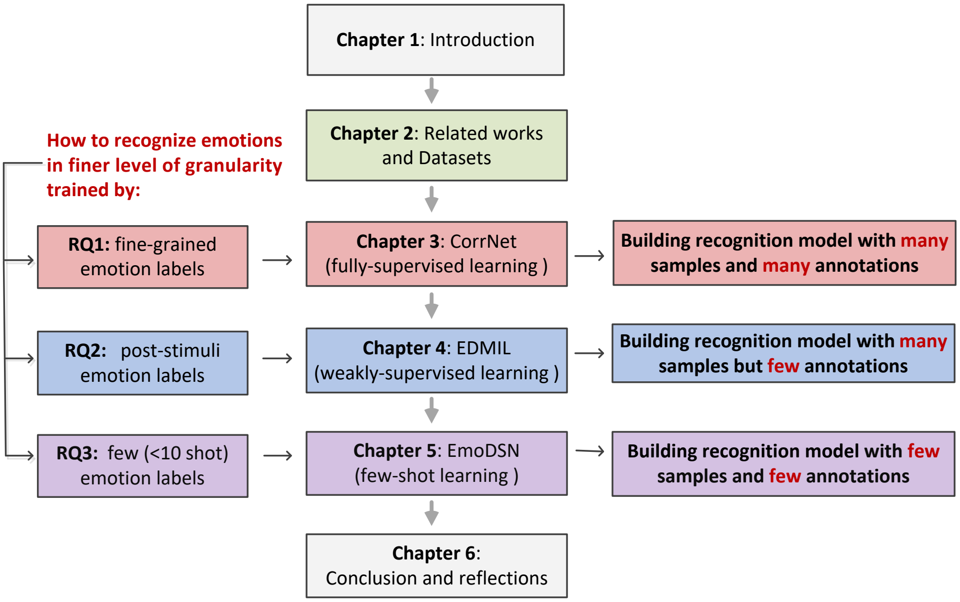

Chapter 1: Introduction
Chapter 2: Related works and Datasets
How to recognize emotions in finer level of granularity trained by:
RQ1: fine-grained emotion labels
Chapter 3: CorrNet (fully-supervised learning )
Building recognition model with many samples and many annotations
RQ2: post-stimuli emotion labels
Chapter 4: EDMIL (weakly-supervised learning )
Building recognition model with many samples but few annotations
RQ3: few (<10 shot) emotion labels
Chapter 5: EmoDSN (few-shot learning )
Building recognition model with few samples and few annotations
Chapter 6: Conclusion and reflections

Figure 1.2: The relationship between research questions and thesis structure

**Chapter 2 - Related works and datasets**: this chapter first reviews the related work for emotion recognition using physiological signals. According to the thesis scope, it focuses on the work on recognizing emotions at a fine granularity level. After that, we introduce the datasets we use for validating the algorithms proposed in chapters 3, 4 and 5.

**This chapter is based on:**

[1] ***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 2020 Apr 21 (pp. 1-15)*

[2] *Xue T, El Ali A,* ***Zhang T***, *Ding G, Cesar P. "RCEA-360VR: Real-time, Continuous Emotion Annotation in 360VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels". In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 2021 May 8 (pp. 1-15*)

[3] *Xue T, El Ali A,* ***Zhang T***, *Ding G, Cesar P. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos. IEEE Transactions on Multimedia. 2021 Nov 13.*

**Chapter 3 - Correlation-based fine-grained emotion recognition**: this chapter builds up the baseline for recognizing emotions at a fine granularity level by using fine-grained emotion labels (**RQ1**). By comparing the performance of correlation-based instance learning methods (proposed in this thesis) and other widely-used fully-supervised learning methods for fine-grained emotion recognition, this chapter provides a comprehensive study on the limitations of building up the recognition system using fine-grained emotion labels.

**The contributions of this chapter are:**

1) The design of a correlation-based instance learning (*CorrNet*) for fine-grained emotion recognition. The algorithm outperforms (higher recognition accuracy) the state-out-the-art methods, including sequential learning algorithms (RQ 1.1);
2) The evaluation of the performance of *CorrNet* on different emotion categories tested by both the subject-dependent and subject-independent models. We found the problem of overfitting (instances with labels of high (48%) and low (47%) valence and arousal are classified as neutral) when tested with the subject-independent model due to the imbalanced distribution of training samples on different emotion categories (RQ 1.2);
3) The performance comparison of *CorrNet* through testing experiments with different parameters (e.g, different lengths of instances and different sampling rates). Our findings show: (1) instance segment lengths between 1-4s result in highest recognition accuracies (2) accuracies between laboratory-grade and wearable sensors are comparable, even under low sampling rates ($\leq 64Hz$). The comparison provides insights into how to design a fine-grained emotion recognition algorithm using segmented physiological signals and fine-grained emotion labels (RQ 1.1).

**This chapter is based on:**

[4] ***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 2019 International Conference on Multimodal Interaction 2019 Oct 14 (pp. 404-408).*

[5] ***Zhang T***, *El Ali A, Wang C, Hanjalic A, Cesar P. Corrnet: Fine-grained emotion recognition for video watching using wearable physiological sensors. Sensors. 2021 Jan;21(1):52.*

**Chapter 4 - Estimating fine-grained emotions from post-stimuli emotion labels: a weakly-supervised learning approach for fine-grained emotion recognition**: this chapter explores the possibility of estimating fine-grained emotions using post-stimuli emotion labels and weakly-supervised learning algorithms (**RQ2**). We present the design and validation of a deep multiple instance learning algorithm (*EDMIL*) to solve the problem of temporal ambiguity caused by post-stimuli emotion labels. By comparing the predictions provided by fully-supervised and weakly-supervised learning algorithms, this chapter helps us understand the advantages and disadvantages of training the recognition algorithm using post-stimuli emotion labels compared to using fine-grained emotion labels.

**The contributions of this chapter are:**

1) The design of an end-to-end deep multiple instance learning framework (*EDMIL*) to identify which instances represent the post-stimuli emotion at a fine granularity level using physiological signals. Recognition results show good performance for subject-independent 3-class (high/neutral/low) classification (RQ 2.1).
2) The experiments on the performance of *EDMIL* towards different instance lengths and feature extraction methods. The experiment results show that instance segment lengths between 1–2s result in the highest recognition accuracies. The results also show that feature extraction using an end-to-end structure can promote the recognition accuracy compared to manual feature extraction and unsupervised learning feature extraction methods (RQ 2.2).
3) The results comparison between algorithms trained with fine-grained and post-stimuli emotion labels using fully-supervised and weakly-supervised learning respectively. We found that compared to fully-supervised learning, weakly supervised learning can reduce overfitting that results from the temporal mismatch between fine-grained annotations and input signals (RQ 2.3).

**This chapter is based on:**

[6] ***Zhang T***, *El Ali A, Wang C, Hanjalic A, Cesar P. Weakly-supervised Learning for Fine-grained Emotion Recognition using Physiological Signals. IEEE transaction on affective computing 2022.*

**Chapter 5 - Predicting emotions from few annotated samples: a few-shot learning approach for fine-grained emotion recognition**: this chapter investigates the possibility of predicting fine-grained emotions using only few annotated signal segments (**RQ3**). We present the design and validation of a deep siamese network (*EmoDSN*) to learn useful representations (embeddings) from few annotated signal segments for fine-grained emotion recognition. This chapter focuses on how the learning structure of the network and temporal positions of training samples would influence the recognition results when only few annotated signal segments are available.

**The contributions of this chapter are:**

1) The design and evaluation of an end-to-end deep siamese network based emotion recognition algorithm (*EmoDSN*) which can predict emotion at a fine granularity level (2s) trained by a small amount (< 10 shot) of physiological segments. Recognition results show good performance for both personalized binary (1D-2C) and 5-class (2D-5C) classification. The algorithm can help researchers to understand the personalized experience of users watching videos by collecting only a small amount of data for training (RQ 3.1).
2) The performance comparison between *EmoDSN* and state-of-the-art few-shot learning (FSL) networks. We find that the pair-by-pair learning structure used by our method results in better performance and less training time compared with other FSL methods using one-to-many structures. This finding could help researchers to design the learning structures of the FSL network for fine-grained emotion recognition by small data (RQ 3.2).

3) The experiments to identify training samples from which temporal moments of video watching (e.g., begin, end and changing points) can better represent the distribution of emotion labels and result in better recognition results. We find that the changing points of emotion annotation and the ending moments of video watching are better temporal moments for training samples (result in higher recognition accuracy) when only few annotated samples are available. Our answers to this research question enable the possibility of helping researchers minimize the amount of annotated training samples by asking users to annotate at the most suitable temporal moments while watching a video (RQ 3.2).

**This chapter is based on:**

[7] ***Zhang T***, *El Ali A, Hanjalic A, Cesar P. Few-shot Learning for Fine-grained Emotion Recognition using Physiological Signals. IEEE transaction on multimedia 2022.*

**Chapter 6- Conclusions and reflections**: this chapter concludes this thesis and points out possible future directions. This chapter starts with answering the research questions raised in chapter 1. After that, we discuss alternative methods which have the potential to recognize emotion at a fine granularity level when there is a limited amount of annotation. Based on the obtained insights from the discussion and observed limitations of the research reported in the thesis, we point out possible directions for future work .

**This chapter is based on:**

[8] ***Zhang T***. *Multi-modal Fusion Methods for Robust Emotion Recognition using Body-worn Physiological Sensors in Mobile Environments. In 2019 International Conference on Multimodal Interaction 2019 Oct 14 (pp. 463-467).*

[9] *Furdui A,* ***Zhang T***, *Worring M, Cesar P, El Ali A. AC-WGAN-GP: Augmenting ECG and GSR Signals using Conditional Generative Models for Arousal Classification. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers 2021 Sep 21 (pp. 21-22).*

# REFERENCES

[1] G. H. Bower, *Mood and memory.* American psychologist **36**, 129 (1981).

[2] A. Damasio, *Descartes' error: emotion, reason and the human brain. 1994 new york,* NY: Avon Books .

[3] M. Soleymani, M. Pantic, and T. Pun, *Multimodal emotion recognition in response to videos,* IEEE transactions on affective computing **3**, 211 (2011).

[4] A. Bartsch, *Emotional gratification in entertainment experience. why viewers of movies and television series find it rewarding to experience emotions,* Media Psychology **15**, 267 (2012).

[5] P. Arriaga, J. Alexandre, O. Postolache, M. J. Fonseca, T. Langlois, and T. Chambel, *Why do we watch? the role of emotion gratifications and individual differences in predicting rewatchability and movie recommendation,* Behavioral Sciences **10**, 8 (2020).

[6] D. Zillmann, *Mood management through communication choices,* American Behavioral Scientist **31**, 327 (1988).

[7] M. B. Oliver, *Exploring the paradox of the enjoyment of sad films,* Human Communication Research **19**, 315 (1993).

[8] R. GhasemAghaei, A. Arya, and R. Biddle, *A dashboard for affective e-learning: Data visualization for monitoring online learner emotions,* in *EdMedia+ Innovate Learning* (Association for the Advancement of Computing in Education (AACE), 2016) pp. 1536–1543.

[9] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, *A review of emotion recognition using physiological signals,* Sensors **18**, 2074 (2018).

[10] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, *Physiological signals based human emotion recognition: a review,* in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications* (IEEE, 2011) pp. 410–415.

[11] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, *Emujoy: Software for continuous measurement of perceived emotions in music,* Behavior Research Methods **39**, 283 (2007).

[12] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, *Analysis of eeg signals and facial expressions for continuous emotion detection,* IEEE Transactions on Affective Computing **7**, 17 (2015).

[13] E. Paul, *Emotions revealed: recognizing faces and feelings to improve communication and emotional life,* NY: OWL Books (2007).

[14] R. W. Levenson, *Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity,* Social psychophysiology: Theory and clinical applications (1988).

[15] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, *Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos),* IEEE Access **7**, 57 (2018).

[16] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, *Continuous emotion detection using eeg signals and facial expressions,* in *2014 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2014) pp. 1–6.

[17] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, *Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze,* in *2019 International Conference on Multimodal Interaction* (2019) pp. 40–48.

[18] G. Van Houdt, C. Mosquera, and G. Napoles, *A review on the long short-term memory model,* Artificial Intelligence Review **53**, 5929 (2020).

[19] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2020) pp. 1–15.

[20] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* Scientific data **6**, 1 (2019).

[21] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, *Introducing the recola multimodal corpus of remote collaborative and affective interactions,* in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (IEEE, 2013) pp. 1–8.

[22] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, *Avec 2015: The 5th international audio/visual emotion challenge and workshop,* in *Proceedings of the 23rd ACM international conference on Multimedia* (2015) pp. 1335–1336.

[23] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, *Avec 2016: Depression, mood, and emotion recognition workshop and challenge,* in *Proceedings of the 6th international workshop on audio/visual emotion challenge* (2016) pp. 3–10.

[24] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, *K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations,* arXiv preprint:2005.04120 (2020).

[25] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, *Decaf: Meg-based multimodal database for decoding affective physiological responses,* IEEE Transactions on Affective Computing **6**, 209 (2015).

[26] J. J. Gross, *The future's so bright, i gotta wear shades,* Emotion Review **2**, 212 (2010).

[27] K. Christoff, Z. C. Irving, K. C. Fox, R. N. Spreng, and J. R. Andrews-Hanna, *Mind-wandering as spontaneous thought: a dynamic framework,* Nature Reviews Neuroscience **17**, 718 (2016).

[28] P. Ekman, *An argument for basic emotions,* Cognition & emotion **6**, 169 (1992).

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

[30] A. Mehrabian, *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies,* (1980).

[31] A. Hanjalic and L.-Q. Xu, *Affective video content representation and modeling,* IEEE transactions on multimedia **7**, 143 (2005).

[32] A. Hanjalic and L.-Q. Xu, *User-oriented affective video content analysis,* in *Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL 2001)* (IEEE, 2001) pp. 50–57.

[33] T. M. Libkuman, H. Otani, R. Kern, S. G. Viger, and N. Novak, *Multidimensional normative ratings for the international affective picture system,* Behavior research methods **39**, 326 (2007).

[34] A. Betella and P. F. Verschure, *The affective slider: A digital self-assessment scale for the measurement of human emotions,* PloS one **11**, e0148037 (2016).

[35] J. A. Russell and L. F. Barrett, *Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.* Journal of personality and social psychology **76**, 805 (1999).

[36] R. Dietz and A. Lang, *Affective agents: Effects of agent affect on arousal, attention, liking and learning,* in *Proceedings of the Third International Cognitive Technology Conference, San Francisco* (1999).

[37] R. A. Calvo and S. D'Mello, *Affect detection: An interdisciplinary review of models, methods, and their applications,* IEEE Transactions on affective computing **1**, 18 (2010).

[38] W. James, *What is an emotion?* Mind **9**, 188 (1884).

[39] W. B. Cannon, *Again the james-lange and the thalamic theories of emotion.* Psychological Review **38**, 281 (1931).

[40] J. R. Loaiza, *Emotions and the problem of variability,* Review of Philosophy and Psychology , 1 (2020).

[41] S. D. Kreibig, *Autonomic nervous system activity in emotion: A review,* Biological psychology **84**, 394 (2010).

[42] P. Ekman, R. W. Levenson, and W. V. Friesen, *Autonomic nervous system activity distinguishes among emotions,* science **221**, 1208 (1983).

[43] M. Khamis, A. Baier, N. Henze, F. Alt, and A. Bulling, *Understanding face and eye visibility in front-facing cameras of smartphones used in the wild,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2018) pp. 1–12.

[44] K. Sharma, C. Castellini, E. L. Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* arXiv preprint:1812.02782 (2018).

[45] M. Song, Z. Yang, A. Baird, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, and B. Schuller, *Audiovisual analysis for recognising frustration during game-play: Introducing the multimodal game frustration database,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 517–523.

[46] I. Abdic, L. Fridman, D. McDuff, E. Marchi, B. Reimer, and B. Schuller, *Driver frustration detection from audio and video in the wild,* Proceedings of the KI , 237 (2016).

[47] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research* (Oxford University Press, 2001).

[48] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, *'feeltrace': An instrument for recording perceived emotion in real time,* in *ISCA tutorial and research workshop (ITRW) on speech and emotion* (2000).

[49] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek, *Continuous, real-time emotion annotation: A novel joystick-based analysis framework,* IEEE Transactions on Affective Computing (2017).

[50] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, *Rcea-360vr: Real-time, continuous emotion annotation in 360 vr videos for collecting precise viewport-dependent ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2021).

[51] D. Lottridge and M. Chignell, *Sliders rate valence but not arousal: Psychometrics of self-reported emotion assessment,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 54 (SAGE Publications Sage CA: Los Angeles, CA, 2010) pp. 1766–1770.

[52] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, *Multiple instance learning for emotion recognition using physiological signals,* IEEE Transactions on Affective Computing (2019).

[53] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, *Deep learning for emotion recognition on small datasets using transfer learning,* in *Proceedings of the ACM on international conference on multimodal interaction* (2015) pp. 443–449.

[54] J. Bang, T. Hur, D. Kim, J. Lee, Y. Han, O. Banos, J.-I. Kim, S. Lee, *et al.*, *Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments,* Sensors **18**, 3744 (2018).

[55] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille, *Regularizing face verification nets for pain intensity regression,* in *2017 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2017) pp. 1087–1091.

[56] H. D. Critchley and S. N. Garfinkel, *Interoception and emotion,* Current opinion in psychology **17**, 7 (2017).

[57] M. A. Nicolaou, H. Gunes, and M. Pantic, *Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,* IEEE Transactions on Affective Computing **2**, 92 (2011).

[58] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, *Emotion recognition using multimodal residual lstm network,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 176–183.

[59] S.-h. Zhong, A. Fares, and J. Jiang, *An attentional-lstm for improved classification of brain activities evoked by images,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 1295–1303.

[60] S. Haripriyadharshini and S. Gnanasaravanan, *Eeg based human facial emotion recognition system using lstmrnn,* .

[61] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks,* in *Advances in neural information processing systems* (2014) pp. 3104–3112.

[62] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, *Enhanced lstm for natural language inference,* arXiv preprint:1609.06038 (2016).

[63] E. A. Veltmeijer, C. Gerritsen, and K. Hindriks, *Automatic emotion recognition for groups: a review,* IEEE Transactions on Affective Computing (2021).

[64] A. V. Tarasov and A. V. Savchenko, *Emotion recognition of a group of people in video analytics using deep off-the-shelf image embeddings,* in *International Conference on Analysis of Images, Social Networks and Texts* (Springer, 2018) pp. 191–198.

[65] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, *Ceap-360vr: A continuous physiological and behavioral emotion annotation dataset for 360 vr videos,* IEEE Transactions on Multimedia (2021).

[66] B. L. Fredrickson and D. Kahneman, *Duration neglect in retrospective evaluations of affective episodes.* Journal of personality and social psychology **65**, 45 (1993).

[67] J. Wu, Z. Zhou, Y. Wang, Y. Li, X. Xu, and Y. Uchida, *Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction,* in *2019 International Conference on Multimodal Interaction* (2019) pp. 582–588.

[68] Z.-H. Zhou, *Ensemble methods: foundations and algorithms* (Chapman and Hall/CRC, 2012).

[69] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, *Multiple instance learning: A survey of problem characteristics and applications,* Pattern Recognition **77**, 329 (2018).

[70] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, *Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications,* International journal of human-computer studies **72**, 717 (2014).

[71] H. Sakoe and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition,* IEEE transactions on acoustics, speech, and signal processing **26**, 43 (1978).

[72] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, *Dense classification and implanting for few-shot learning,* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019) pp. 9258–9267.

[73] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, *Signature verification using a "siamese" time delay neural network,* International Journal of Pattern Recognition and Artificial Intelligence **7**, 669 (1993).

[74] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, *Zero-shot emotion recognition via affective structural embedding,* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019) pp. 1151–1160.

# 2

# Related works and Datasets

*In this chapter, we first review the most relevant previous work on emotion recognition using physiological signals. Then, we narrow our scope down to the related work on recognizing emotions at a fine granularity level. The review of this previous work helped us understand the advantages and limitations of the state-of-the-art methods, which motivated the development of the algorithms we proposed in the following three chapters of the thesis. At last, we describe here in detail the three datasets we used in this thesis: CASE [1], MERCA [2] and CEAP-360VR [3] which are collected in three environments: desktop, mobile and VR, respectively.*

This chapter is based on the following publications:

1. ***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, pp. 1-15, April 2020*
2. *Xue T, El Ali A,* ***Zhang T****, Ding G, Cesar P. "RCEA-360VR: Real-time, Continuous Emotion Annotation in 360VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels". In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-15, May 2021*
3. *Xue T, El Ali A,* ***Zhang T****, Ding G, Cesar P. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos. IEEE Transactions on Multimedia, November 2021*

## 2.1. Related works

### 2.1.1. Emotion recognition using physiological signals

To unravel the relationship between emotions and physiological signals, researchers have used different methods. Psychologists and physiologists use phenomenological methods [4, 5] to directly observe and summarize the relationship between physiological signals and emotions. Researchers often ask subjects to elicit emotions by themselves (e.g., to relive an experience that would elicit a given emotion [4]) and observe the changes in physiological signals. A disadvantage of this approach is the difficulty in coping with complex (i.e., non-linear) changes [6], making the conclusions on the relations between emotions and signals potentially inconsistent and contradictory (i.e., the *variable phenomena* [7, 8]).

Except for experimental observation, which can handle well linear correlation between physiological signals and emotions, researchers in the field of affective computing use computational methods [9, 10] to deal with non-linear mapping between them. Emotion recognition using computational methods can be divided into two major categories: model-specific methods and model-free methods [11]. Model-specific methods require pre-designed hand-crafted features to classify emotions from physiological signals. In general, statistical features from the time-domain (e.g., mean, standard deviation, first differential [12–14] of the signal) and frequency-domain (e.g., mean of amplitude, mean of absolute value [15, 16], signal Fast Fourier Transform (FFT) [17]) are commonly extracted for this purpose. The extracted features then serve as input into classifiers, such as Support Vector Machine (SVM) [18], K-Nearest Neighbor (KNN) [19], or Random Forest (RF) [20] to classify emotions. For example, Zhao et al. [21] extract 223 features from 4 physiological signals to recognize the valence and arousal of users. Their algorithm, which merges information from users' personalities using a hypergraph learning framework, achieves up to 70% accuracy on the *ASCERTAIN* [22] dataset. Jimenez et al. select 13 features from PPG (4 in the time domain, 9 in the frequency domain) and 14 features from GSR (all in the time domain) to recognize six basic emotions. A simple k-nearest neighbor (KNN) classifier is used by Aasim et al. [23] to classify valence and arousal on their newly collected MULtile SEnsorial media (MULSEmedia) dataset. Similar to the accuracy from the work of Zhao et al. [21], they achieve 85.18% and 76.54% accuracy for valence and arousal, respectively.

Although model-specific methods have been widely used by researchers for a long time [11], they require researchers to select features based on empirical experiments [11, 24]. However, there is no consensus on which features are the most reliable for recognizing emotions [10, 11]. Therefore researchers have to carefully design features according to the data they collected, limiting the generalizability of their algorithms. 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 [11, 24].

The model-free methods use artificial neural networks to learn the inherent structure between input signals and emotion labels. Thus, they can automatically extract features from physiological signals for recognition. Neural networks such as convolutional neural networks (CNNs) [25, 26] and Long Short-Term Memory (LSTM) networks [27, 28] are widely used for emotion recognition and achieve high accuracy. For example, a regularized deep fusion framework is designed by Zhang et al. [29] to learn task-specific representations for each physiological signal. Their experiments show that the method can improve the performance of subject-independent emotion recognition by 6% compared to other fusion methods such as single modal classifiers (i.e., SVM, Decision Tree, and Naive Bayes). Ma et al. implement [27] a multimodal residual LSTM network to classify valence and arousal and obtain a classification accuracy of 92.87% and 92.30% for arousal and valence respectively. According to the research from Suhara et al [30], 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 can easily overfit on the training data when using deep and sophisticated structures [31]. This appears to be a common phenomenon when users label their emotions [1, 32]. Thus, there are still challenges in designing model-free methods for better generalizability for emotion recognition.

The methods and algorithms proposed in this thesis attempt to draw on the advantages of both model-specific and model-free methods. The correlation-based fine-grained emotion recognition algorithm designed in **Chapter 3** uses unsupervised learning techniques to automatically extract features (model-specific) and supervised learning classifiers to classify emotions. The weakly-supervised and few-shot learning algorithms designed in **Chapter 4** and **5** use end-to-end learning structures (model free) which can automatically extract features from physiological signals.

### 2.1.2. FINE-GRAINED EMOTION RECOGNITION

While there exist many algorithms that are designed for recognizing emotions based on physiological signals, techniques for fine-grained emotion recognition are still in their infancy [33]. Fine-grained emotion recognition requires algorithms to output different emotion states along a video by relying on signals within short time intervals. For temporal signals, this is typically done using two kinds of methods: the regression methods and classification methods.

The regression methods view the target emotion states as a continuous sequence and directly calculate the mapping from input signals to output sequences of emotions. These methods include sequential learning approaches, such as LSTM [27, 28], and temporal regression, such as support vector regression (SVR) [34, 35] and polynomial regression [36]. While previous work has shown that regression approaches, especially sequential learning using recurrent structures, can achieve high accuracy [37–39], 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 [29, 40].

The classification methods segment continuous signals into different fine-grained signal segments and classify the emotion of each instance independently. Therefore, the recognition result of different signal segments will not affect each other. For example, Romeo et al. [41] designed an SVM-based multi-instance learning algorithm to recognize valence and arousal for each fine-grained instance achieving 68% of accuracy on high arousal. Awais et al. [42] designed an LSTM-based classification method to classify emotions every 5 seconds. Srinivasan et al. [43] implemented *decision trees* on a RaspberryPi device to classify the valence (positive/negative) of users every 10 seconds. This kind of method is also widely used for fine-grained emotion recognition with different data modalities, such as facial expressions [44] and vocal features [45] (e.g., pitch and loudness). The main challenge here is to extract and fuse both the features inside and between signal segments, as the information residing only within signal segments may not be enough to determine which emotion it represents. Previous work [41, 46] uses the joint loss [47] of signal segments and the entire signal to fuse the features inside and between signal segments. This, however, could lead to temporal ambiguity of emotions as signal segments are not directly trained by their emotion labels, but, instead, by the label of the entire signal [41].

Both the regression and classification methods need fine-grained emotion labels collected at the same or similar granularity to ensure that every sample has a corresponding ground truth for learning. For classification methods, the granularity of the required ground truth labels is the same as the classification results (e.g., 5s and 10s for [42] and [43], respectively). For regression methods, the granularity of the required labels is usually the same as the input signals [38, 39]. Since collecting these ground truth labels is costly and time-consuming (both the regression and classification methods in previous works), this thesis explores new machine learning methods which can obtain accurate recognition results with a limited amount of labels for training. The weakly-supervised and few-shot learning methods developed in **Chapter 4** and **5** require only post-stimuli labels (one emotion label for one video watching) and few annotated samples (<10 samples per emotion category) for training, respectively.

## 2.2. DATASETS

To evaluate the performance of the algorithms proposed in this thesis, we use three datasets: CASE [1], MERCA [48] and CEAP-360VR [49]. CASE is the only published dataset which has continuously self-annotated physiological signals. However, the CASE dataset is collected in an indoor, desktop environment. To evaluate the generalizability of the algorithms, we collected the MERCA and CEAP-360VR datasets with continuous self-annotated (valence and arousal, V-A) physiological signals in an outdoor-mobile and indoor-Virtual-Reality 360°-video watching setting, respectively. Testing on MERCA and CEAP-360VR allows us to additionally test the performance of our algorithms across different application scenarios. It also allows us to test our algorithms across signals collected using golden standard (CASE) and wearable (MERCA and CEAP-360VR) devices. Table 2.1 illustrates the scenarios and physiological sensors for collecting these three datasets. Below we describe each dataset in more detail.

### 2.2.1. CASE DATASET

The CASE dataset [1] contains physiological recordings from 30 participants (15m, 15f), aged between 22-37. Valence and arousal are annotated by the participants using a physical joystick (shown in Figure 2.1) 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) ×4 (video emotions: amusing vs. boring vs. relaxing vs. scary) within-subjects design, tested in an indoor laboratory environment. Eight video clips,

Table 2.1: The testing scenarios and used physiological sensors for CASE, MERCA and CEAP-360VR

| Dataset | Sensor | Scenario |
|---|---|---|
| CASE | Thought Technology SA series, golden standard device, high sampling-rate (1000Hz) | indoor, desktop video watching |
| MERCA | Empatica E4, wearable, consumer device, low sampling rate (≤64Hz) | outdoor, mobile video watching |
| CEAP 360VR | Empatica E4, wearable, consumer device, low sampling rate (≤64Hz) | indoor, HMD-based 360° VR video watching |

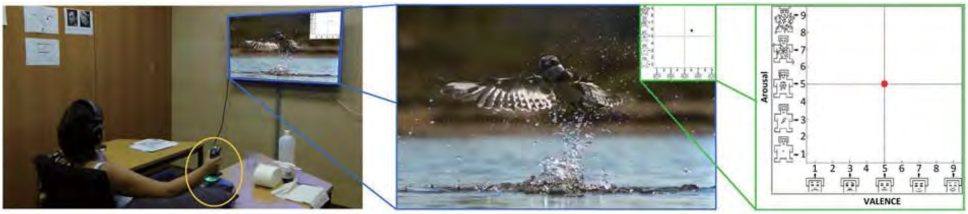


Figure 2.1: The experiment setup and annotation interface for CASE [1].

two videos per emotion and with the duration statistics M=158.75s, SD = 23.67s, were selected to elicit the corresponding emotions. These videos contains clips extracted from movies and documentaries. Emotion attributes of the selected videos are determined based on online ratings by 12 annotators. Six sensors (ECG, BVP, EDA, RESP, TEMP, EMG) were deployed to collect physiological signals. All sensors were synchronized and sampled at 1000Hz (sample size: 2451650 samples × 7 signals × 30 participants). The V-A ratings (sample size: 49033 samples × 2 annotations × 30 participants) were collected at the frequency of 20Hz according to the sampling rate of the physical joystick.

### 2.2.2. MERCA DATASET

#### EXPERIMENT SETUP

20 participants (12m, 8f) aged between 22-32 participated in the data collection experiment of MERCA. 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. As in CASE, the data collection experiment for MERCA followed a 1 (task: watch videos and continuously annotate emotions) × 4 (video emotions: joy vs. fear vs. sad vs. neutral) within-subjects design. As shown in Figure 2.2, the experiment was conducted in the outdoor campus of our institute. Participants could walk or stand freely while watching videos. 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 [50–52].

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.

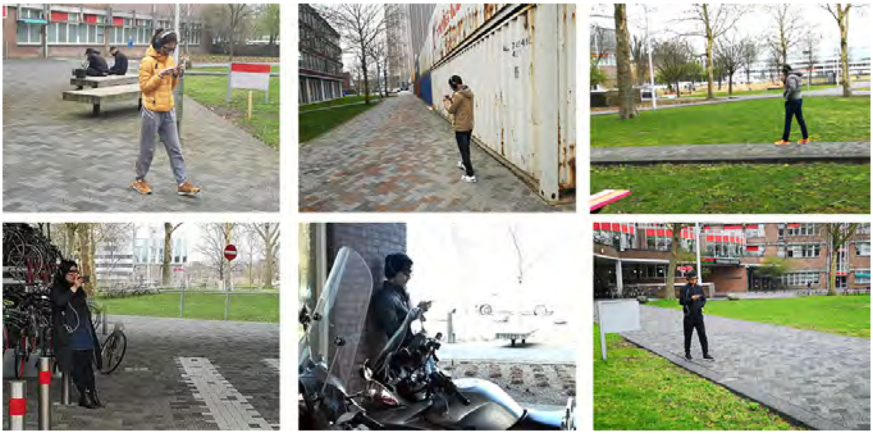


Figure 2.2: The experiment environment of MERCA. Participant photos shown with permission.

### Video stimuli

12 video clips (three videos per emotion, duration M = 81.4s and SD = 22.5s) were selected to elicit the corresponding emotions. 10s of black screens are added before and after each video to decrease the influence of the emotion elicited in one video on the one elicited by another video. We chose the 12 videos according to 2D emotion annotations from the self-reports in MAHNOB dataset [53]. We use the videos in MAHNOB dataset because it is a widely used dataset [54, 55] with emotion self-reports from more than 30 reviewers. We selected more videos compared to CASE, because we aim to collect more samples for each emotion.

### Data collection

Emotions (as V-A) are annotated by the participants using a real-time, continuous emotion annotation (RCEA) mobile application [56]. Participants can input their valence and arousal using a virtual joystick (shown in Figure 2.3) on the screen of the mobile device which they use for video watching. The virtual joystick is designed based on Russell's Circumplex model [57]. 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 [58] 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 the participants with the procedure.

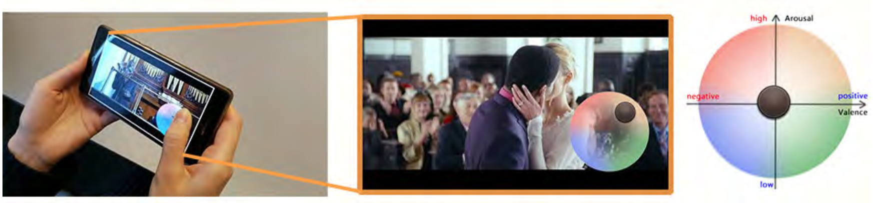


Figure 2.3: The real-time and continuous V-A annotation interface (cf., [56]) used for MERCA.

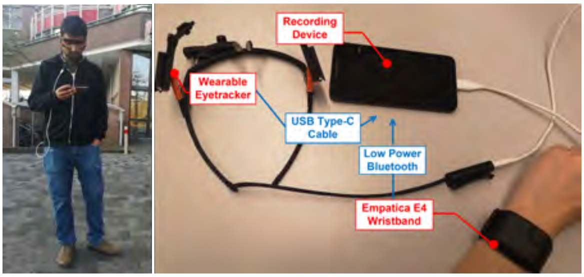


Figure 2.4: The hardware setup of MERCA. Image of study participant shown with permission.

We used the Pupil Core wearable eyetracker [1] and Empatica E4[2] wristband to collect signals from the Autonomic Nerve System (ANS) and the Oculomotor Nerve System (ONS), respectively. We chose these two devices because they are wearable and therefore suitable for collecting signals in outdoor environments, and because they have been used by previous studies [59–61]. 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 whether 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) [3], BVP (84864 × 20)[3], EDA (5304 × 20)[3] and TEMP (5304 × 20)[3]. From the wearable eyetracker, we collected pupil dilation (13260 × 20)[3], saccadic amplitude (13260 × 20)[3] and saccadic velocity (13260 × 20)[3]. Data from these two sensors were stored on one mobile device (the recording device). As shown in Figure 2.4, 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

[1] https://pupil-labs.com/products/core/

[2] https://www.empatica.com/en-eu/research/e4/

[3] sample size, samples × participants

server (`android.pool.ntp.org`).

### 2.2.3. CEAP-360VR dataset

#### Experiment setup

The *CEAP-360VR*[4] [49] (Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° Videos) dataset contains physiological signals for 32 participants (16m, 16f) aged between 18-33 (M=25, SD=4.0). They were recruited by posters from the universities nearby for the experimental setup that simulates the scenario in which users watch 360° videos using HMD-based VR devices. All participants reported normal or corrected-to-normal vision and not being color-blind. 50% of the participants are female and 27 participants have used VR devices less than five times before. As in CASE and MERCA, the data collection experiment in CEAP-360VR is a 1 (task: watching 360° videos and continuously annotate valence and arousal) × 4 (video emotions: high valence+high arousal vs. high valence+low arousal vs. low valence+low arousal vs. low valence+high arousal) within-subjects design. During the experiment, participants sat on a swivel chair and were free to look in any direction. Every participant was orally instructed to get familiar with the continuous emotion annotation method and the visualization feedback, as well as the 360° video viewing experience by moving the head and rotating the chair.

#### Video Stimuli

We selected two sample 360° videos (8 videos in total) to represent each emotion type. The eight videos are chosen based on the database provided by Li et al [62], which contains valence and arousal mean ratings (V-A mean ratings) from 95 subjects. We used youtube-dl[5] to download the contents from YouTube with 4K in resolution (3840 X 1920 pixels), equirectangular format. The videos come in different lengths and most are longer than 2 minutes, so we extracted a 60-s segment from each of them with no scene cuts.

#### Data collection

Similar to CASE and MERCA, a physical joystick (Joy-Con Controller, shown in Figure 2.5 (a) ) was used by the users to annotate their valence and arousal continuously while they were watching 360° video clips. The movement of the joystick head maps into a 2D Valence-Arousal space, in which the x axis indicates valence while the y axis indicates arousal, as shown in Figure. 2.5 (c).

The physiological signals of the participants were measured through the Empatica E4 wristband. For Empatica E4, they collect HR (1Hz, sample size: 360 samples × 32 users), BVP (64Hz, 11520 samples × 32 users), EDA (4Hz, 1440 samples × 32 users) and TEMP (4Hz, 1440 samples × 32 users). The collected signals were stored on a mobile device which was connected with E4 using low-power bluetooth. One laptop was used to play the 360° videos as well as to log the head and eye movement from HTC Vive Pro Eye (as shown in Figure 2.5 (b) ). Timestamps of the mobile device were set according to the clock of the laptop, synchronized via an NTP server. The V-A ratings (sample size = 3600 samples × 32 users) were collected at 10Hz according to the sampling rate of the physical

[4] https://www.dis.cwi.nl/ceap-360vr-dataset/

[5] https://github.com/ytdl-org/youtube-dl

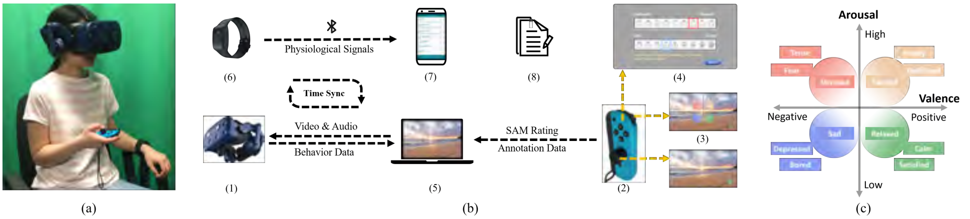


Figure 2.5: (a) A participant in our experiment watching a 360° video using the HTC VIVE Pro Eye HMD and annotating her emotional state using a Joy-Con controller, while wearing an Empatica E4 Wristband on the non-dominant hand. (b) The system schematic shows various aspects of the experiment set-up and data acquisition. (c) Valence-Arousal model space based on Russell's Circumplex model [63]. In our annotation system, four distinct colors are selected across quadrants (HEX values = #eecdac, #7fc087, #879af0, #f4978e for quadrants one to four clock-wise, respectively, which has been shown to be intuitive and easy for users to understand [64])

joystick. After each 360° video was played, users were asked to rate their post-stimuli V-A for that video using a within-VR SAM [65] rating scale.

# REFERENCES

[1] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* Scientific data **6**, 1 (2019).

[2] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2020) pp. 1–15.

[3] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, *Rcea-360vr: Real-time, continuous emotion annotation in 360 vr videos for collecting precise viewport-dependent ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2021).

[4] P. Ekman, R. W. Levenson, and W. V. Friesen, *Autonomic nervous system activity distinguishes among emotions,* science **221**, 1208 (1983).

[5] R. W. Levenson, P. Ekman, and W. V. Friesen, *Voluntary facial action generates emotion-specific autonomic nervous system activity,* Psychophysiology **27**, 363 (1990).

[6] A. Scarantino, *Basic emotions, psychological construction, and the problem of variability.* (2015).

[7] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, *The brain basis of emotion: a meta-analytic review,* The Behavioral and brain sciences **35**, 121 (2012).

[8] L. F. Barrett, *Are emotions natural kinds?* Perspectives on psychological science **1**, 28 (2006).

[9] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, *Basic emotions are associated with distinct patterns of cardiorespiratory activity,* International journal of psychophysiology **61**, 5 (2006).

[10] R. A. Calvo and S. D'Mello, *Affect detection: An interdisciplinary review of models, methods, and their applications,* IEEE Transactions on affective computing **1**, 18 (2010).

[11] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, *A review of emotion recognition using physiological signals,* Sensors **18**, 2074 (2018).

[12] W. Yang, M. Rifqi, C. Marsala, and A. Pinna, *Towards better understanding of player's game experience,* in *Proceedings of the ACM on International Conference on Multimedia Retrieval* (2018) pp. 442–449.

[13] S. Wioleta, *Using physiological signals for emotion recognition,* in *2013 6th International Conference on Human System Interactions (HSI)* (IEEE, 2013) pp. 556–561.

[14] X. Niu, L. Chen, H. Xie, Q. Chen, and H. Li, *Emotion pattern recognition using physiological signals,* Sensors & Transducers **172**, 147 (2014).

[15] E. Di Lascio, S. Gashi, and S. Santini, *Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors,* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1 (2018).

[16] M. Zecca, S. Micera, M. C. Carrozza, and P. Dario, *Control of multifunctional prosthetic hands by processing the electromyographic signal,* Critical Reviews™ in Biomedical Engineering **30** (2002).

[17] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, *Engagemon: Multi-modal engagement sensing for mobile games,* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1 (2018).

[18] C. He, Y.-j. Yao, and X.-s. Ye, *An emotion recognition system based on physiological signals obtained by wearable sensors,* in *Wearable sensors and robots* (Springer, 2017) pp. 15–25.

[19] L. Chen, M. Li, W. Su, M. Wu, K. Hirota, and W. Pedrycz, *Adaptive feature selection-based adaboost-knn with direct optimization for dynamic emotion recognition in human–robot interaction,* IEEE Transactions on Emerging Topics in Computational Intelligence (2019).

[20] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis, *A user independent, biosignal based, emotion recognition method,* in *International Conference on User Modeling* (Springer, 2007) pp. 314–318.

[21] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, *Personalized emotion recognition by personality-aware high-order learning of physiological signals,* ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **15**, 1 (2019).

[22] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, *Ascertain: Emotion and personality recognition using commercial sensors,* IEEE Transactions on Affective Computing **9**, 147 (2016).

[23] A. Raheel, M. Majid, and S. M. Anwar, *Dear-mulsemedia: Dataset for emotion analysis and recognition in response to multiple sensorial media,* Information Fusion **65**, 37 (2021).

[24] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, *Physiological signals based human emotion recognition: a review,* in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications* (IEEE, 2011) pp. 410–415.

[25] M. Ali, F. Al Machot, A. H. Mosa, and K. Kyamakya, *Cnn based subject-independent driver emotion recognition system involving physiological signals for adas,* in *Advanced Microsystems for Automotive Applications 2016* (Springer, 2016) pp. 125–138.

[26] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, *Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos),* IEEE Access **7**, 57 (2018).

[27] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, *Emotion recognition using multimodal residual lstm network,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 176–183.

[28] S.-h. Zhong, A. Fares, and J. Jiang, *An attentional-lstm for improved classification of brain activities evoked by images,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 1295–1303.

[29] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, and T. Zhang, *Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine,* IEEE transactions on cybernetics (2020).

[30] Y. Suhara, Y. Xu, and A. Pentland, *Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks,* in *Proceedings of the 26th International Conference on World Wide Web* (2017) pp. 715–724.

[31] T. Zhang, *Multi-modal fusion methods for robust emotion recognition using body-worn physiological sensors in mobile environments,* in *2019 International Conference on Multimodal Interaction* (2019) pp. 463–467.

[32] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, *Affective labeling in a content-based recommender system for images,* IEEE transactions on multimedia **15**, 391 (2012).

[33] F. Hasanzadeh, M. Annabestani, and S. Moghimi, *Continuous emotion recognition during music listening using eeg signals: A fuzzy parallel cascades model,* arXiv preprint:1910.10489 (2019).

[34] C.-Y. Chang, J.-Y. Zheng, and C.-J. Wang, *Based on support vector regression for emotion recognition using physiological signals,* in *The 2010 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2010) pp. 1–7.

[35] A. E. Hassanien, M. Kilany, E. H. Houssein, and H. AlQaheri, *Intelligent human emotion recognition based on elephant herding optimization tuned support vector regression,* Biomedical Signal Processing and Control **45**, 182 (2018).

[36] J. Wei, T. Chen, G. Liu, and J. Yang, *Higher-order multivariable polynomial regression to estimate human affective states,* Scientific reports **6**, 23384 (2016).

[37] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, *Continuous emotion detection using eeg signals and facial expressions,* in *2014 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2014) pp. 1–6.

[38] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, *Analysis of eeg signals and facial expressions for continuous emotion detection,* IEEE Transactions on Affective Computing **7**, 17 (2015).

[39] M. A. Nicolaou, H. Gunes, and M. Pantic, *Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,* IEEE Transactions on Affective Computing **2**, 92 (2011).

[40] T. Zhang, A. El Ali, C. Wang, X. Zhu, and P. Cesar, *Corrfeat: Correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition,* in *International Conference on Multimodal Interaction* (2019) pp. 404–408.

[41] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, *Multiple instance learning for emotion recognition using physiological signals,* IEEE Transactions on Affective Computing (2019).

[42] M. Awais, M. Raza, N. Singh, K. Bashir, U. Manzoor, S. ul Islam, and J. J. Rodrigues, *Lstm based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19,* IEEE Internet of Things Journal (2020).

[43] A. Srinivasan, S. Abirami, N. Divya, R. Akshya, and B. Sreeja, *Intelligent child safety system using machine learning in iot devices,* in *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (IEEE, 2020) pp. 1–6.

[44] J. Gibson, A. Katsamanis, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan, *Multiple instance learning for behavioral coding,* IEEE Transactions on Affective Computing **8**, 81 (2015).

[45] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. S. Narayanan, *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, 2011) pp. 31–41.

[46] B. Wu, E. Zhong, A. Horner, and Q. Yang, *Music emotion recognition by multi-label multi-layer multi-instance multi-view learning,* in *Proceedings of the 22nd ACM international conference on Multimedia* (2014) pp. 117–126.

[47] O. Maron and T. Lozano-Pérez, *A framework for multiple-instance learning,* in *Advances in neural information processing systems* (1998) pp. 570–576.

[48] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Corrnet: Fine-grained emotion recognition for video watching using wearable physiological sensors,* Sensors **21**, 52 (2021).

[49] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, *Ceap-360vr: A continuous physiological and behavioral emotion annotation dataset for 360 vr videos,* IEEE Transactions on Multimedia (2021).

[50] T. T. Lin and C. Chiu, *Investigating adopter categories and determinants affecting the adoption of mobile television in china,* China Media Research **10**, 74 (2014).

[51] J. McNally and B. Harrington, *How millennials and teens consume mobile video,* in *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '17 (ACM, New York, NY, USA, 2017) pp. 31–39.

[52] K. O'Hara, A. S. Mitchell, and A. Vorbau, *Consuming video on mobile devices,* in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07 (ACM, New York, NY, USA, 2007) pp. 857–866.

[53] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[54] H. Ferdinando, T. Seppänen, and E. Alasaarela, *Enhancing emotion recognition from ecg signals using supervised dimensionality reduction.* in *ICPRAM* (2017) pp. 112–118.

[55] D. Gui, S.-h. Zhong, and Z. Ming, *Implicit affective video tagging using pupillary response,* in *International Conference on Multimedia Modeling* (Springer, 2018) pp. 165–176.

[56] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '20 (Association for Computing Machinery, New York, NY, USA, 2020).

[57] D. H. Olson, C. S. Russell, and D. H. Sprenkle, *Circumplex model: Systemic assessment and treatment of families* (Psychology Press, 1989).

[58] J. Itten, *Mein Vorkurs am Bauhaus* (Otto Maier Verlag, 1963).

[59] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, *Labelling affective states" in the wild" practical guidelines and lessons learned,* in *Proceedings of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (2018) pp. 654–659.

[60] B. Zhao, Z. Wang, Z. Yu, and B. Guo, *Emotionsense: Emotion recognition based on wearable wristband,* in *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)* (IEEE, 2018) pp. 346–355.

[61] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, *Emotion recognition using physiological signals: laboratory vs. wearable sensors,* in *International Conference on Applied Human Factors and Ergonomics* (Springer, 2017) pp. 15–22.

[62] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams, *A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures,* Frontiers in psychology **8**, 2116 (2017).

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

[64] D. Handayani, A. Wahab, and H. Yaacob, *Recognition of emotions in video clips: the self-assessment manikin validation,* Telkomnika **13**, 1343 (2015).

[65] M. M. Bradley and P. J. Lang, *Measuring emotion: the self-assessment manikin and the semantic differential,* Journal of behavior therapy and experimental psychiatry **25**, 49 (1994).

# 3

# Correlation-based fine-grained emotion recognition

*This chapter proposes an emotion recognition algorithm trained with fine-grained emotion labels. It helps us build baselines about how accurate the recognition algorithms can be using fine-grained emotion labels for training. In the previous work, sequence learning methods such as LSTM 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, which can be inaccurate or misaligned temporally. If the network is trained with these labels, training errors could accumulate and affect the recognition result for other instances within the same signal. 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). Results show that for subject-independent binary classification (high-low), CorrNet yields promising recognition accuracy:* 76.37% *and* 74.03% *for V-A on CASE, and* 70.29% *and* 68.15% *for V-A on MERCA.*

*This chapter is published as Zhang T, El Ali A, Wang C, Hanjalic A, Cesar P. Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors, Sensors, 21(1):52. January 2021*

## 3.1. 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 0.5s to 4s 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 for 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.

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, 32, 33], sequential learning methods such as Long Short Term Memory (LSTM) [34] 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 sequential 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 [35, 36].

To address this challenge, this chapter 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.) [37] 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 ($\geq 1000 Hz$).

## 3.2. RELATED WORKS

Physiological signals collected from wearable sensors are widely used for recognizing emotions outside a laboratory environment. For example, Costa et al. [38] 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. [39] 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[1] and Toshiba W110 wristbands [40]) 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 [41], synchrony between presenters and audience members [42], and students' emotional engagement during lectures [43]. 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 [38, 39, 41–43], we narrow our scope on using physiological signals such as ECG, BVP, EDA and HR from wearable sensors.

## 3.3. 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 3.1. *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 latent space where intra-modal features are learned using a convolutional auto-encoder. After that, the feature vectors from the latent 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.

### 3.3.1. 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 inputs of the algorithm become $X = \{x^i\}_{i=1}^{n}$, where $x^i \in R^{L\times C}$. 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 [44] and saccadic eye movement [45]) and from ANS (skin conductance [46] and skin temperature [47]) are grouped together, respectively.

[1] https://www.empatica.com/en-eu/research/e4/

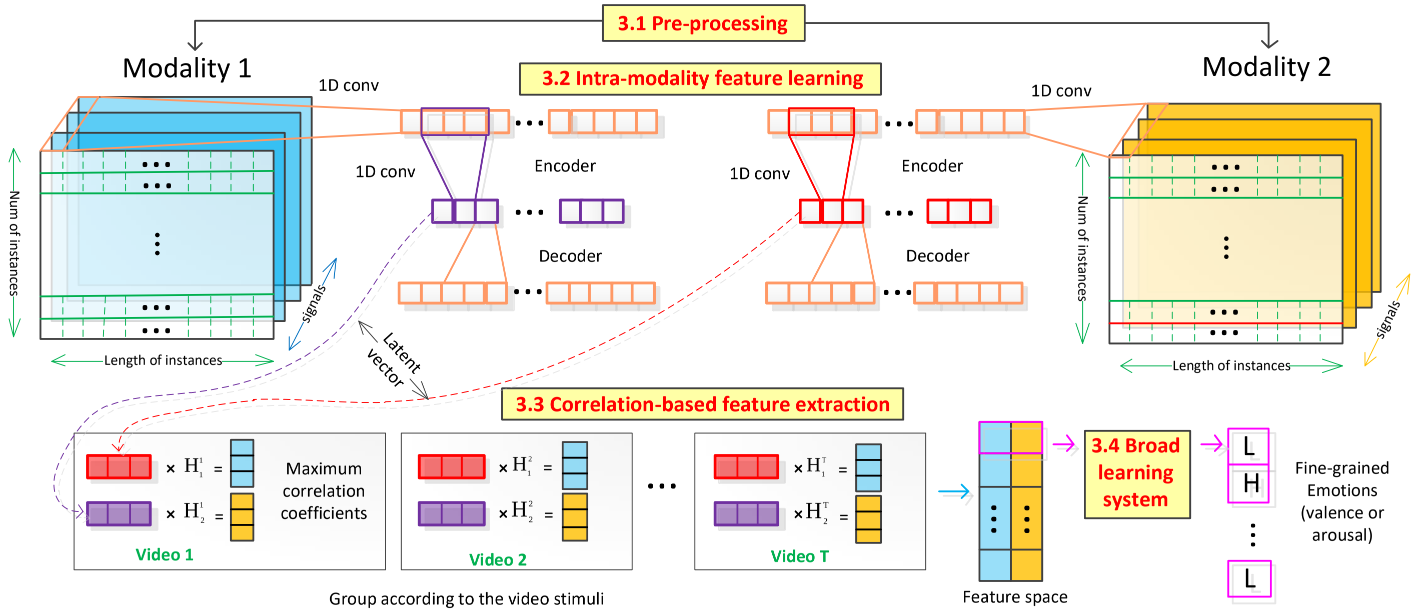


Figure 3.1: The procedure of proposed CorrNet

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

### 3.3.2. 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 [48] is implemented. We use just shallow structure (two layers) instead of deep to avoid overfitting since each instance does not contain much information.

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

$$\phi_1 = X_1 \otimes \psi(\omega, c) = X_1 \dot{\otimes} \psi_{11}(1,1) \bar{\otimes} \psi_{12}(\omega, c) \tag{3.1}$$

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:

$$\eta_1 = \phi_1 \bar{\otimes} \psi'(C_1, c) \tag{3.2}$$

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 [49, 50]:

$$H = -\frac{1}{n \cdot L} \sum_{i=1}^{n} \sum_{j=1}^{L} x_1^{ij} \cdot log(\eta_1^{ij})) + (1 - x_1^{ij}) \cdot (1 - log(\eta_1^{ij}) \tag{3.3}$$

where $x_1^{ij}$ and $\eta_1^{ij}$ are the $j$ sample point in the instance of $x_1^i$ and $\eta_1^i$, respectively. The latent vector $\phi_1$ learned from the auto-encoder is the intra-modality features we want to obtain. The latent vector $\phi_2$ for modality 2 can be calculated using the same method.

### 3.3.3. CORRELATION-BASED FEATURE EXTRACTION

In this stage, intra-modality features $\phi_1$ and $\phi_2$ are fused using a correlation-based feature extraction method [51]. 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 [52]) to lower the computational cost and avoid overfitting (where a strong constraint can make the two modalities have a correlation coefficient of $\approx 1$).

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

$$S_{11} = \frac{(\phi_1^t).^T \phi_1^t}{D^t - 1} + I^{\omega\times\omega}, S_{12} = \frac{(\phi_2^t)^T \phi_1^t}{D^t - 1}, S_{22} = \frac{(\phi_2^t)^T \phi_2^t}{D^t - 1} + I^{\omega\times\omega} \tag{3.4}$$

where $I$ is the unit matrix and $\omega$ is the dimension of the latent space, $D^t$ is the dimension of $\phi_1^t$. Then, we implement the Singular Value Decomposition (SVD) on the equation below:

$$[U, D, V] = \mathrm{SVD}(\mathrm{V_1 D_1 V_1^T \cdot S_{12} \cdot V_2 D_2 V_2^T}) \tag{3.5}$$

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 $D_1 = \mathrm{diag}(\frac{1}{\sqrt{\mathrm{D_{11}}}}, \frac{1}{\sqrt{\mathrm{D_{12}}}}, \ldots, \frac{1}{\sqrt{\mathrm{D_{1k}}}})$ 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:

$$H_1 = V_1 D_1 V_1^T \cdot U', \quad H_2 = V_2 D_2 V_2^T \cdot V' \tag{3.6}$$

where $U'$ and $V'$ consist of the first $K$ columns of $U, V$, respectively. At last, the correlation-based features of $\phi_1^t$ and $\phi_2^t$ 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.

### 3.3.4. Broad learning system for classification

While the previous two stages focus on unsupervised feature extraction, the last stage (Figure 3.1) focuses on a supervised classifier. Here, a Broad Learning System (BLS) [53] is used to map the extracted features to valence and arousal. Compared with deep learning systems such as Deep Belief Networks (DBNs) [54] and Convolutional Neural Networks (CNNs) [55], 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 overfitting [56].

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 [57]. Then, the first feature node $A_1$ can be calculated by:

$$A_1 = F'' \cdot W_{A_1} \tag{3.7}$$

where $F'' = [F'|1]$ is the augmented matrix of $F'$. $W_{A1} \in R^{2K \times N_1}$ is the sparse autoencoder [58] 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 [53]. 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 [58]:

$$\begin{gathered} W_{A1} = \arg\max ||W' \cdot W_{A1} - H''||_2^2 + \lambda ||W_{A1}||_1 \\ W_{A1} \cdot H'' = W' \end{gathered} \tag{3.8}$$

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:

$$B = \mathrm{tansig}\left[\frac{\mathrm{A}' \cdot \mathrm{orth}(\mathrm{W}'') \cdot \mathrm{S}}{\max(\mathrm{A}' \cdot \mathrm{orth}(\mathrm{W}''))}\right] \tag{3.9}$$

where $A' = [A|1]$ is the augmented matrix of $A$. $\mathrm{orth}(\mathrm{W}'')$ 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. $\mathrm{tansig} = \frac{2}{1+\mathrm{e}^{-2\mathrm{t}}} - 1$ 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 [53]:

$$W = (E^T \cdot E + I^{n' \times n'} \cdot C)^{-1} E^T \cdot y \tag{3.10}$$

$C = 2^{-30}$ 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 $y_t = E_t \cdot W$.

## 3.4. 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*.

### 3.4.1. 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:

$$S_n = \frac{S - \min(S)}{\max(S) - \min(S)} \tag{3.11}$$

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 [59]. Since the sampling rates of V-A and signals are 20Hz and 1000Hz respectively, we down-sampled all the signals to 50Hz by decimation down-sampling [60] (the choice of down-sampling rates is discussed in section 3.5.2). The EDA signals were first filtered using a low pass filter with a 2Hz cutoff frequency to remove noise [61]. For the BVP signal, we pre-processed it with a 4-order butterworth band-pass filter with cutoff frequencies [30, 200] Hz to eliminate the bursts [62]. An elliptic band-pass filter with cutoff frequencies [0.005, 0.1] was used to filter the ST signal [63]. We followed the standard filtering procedure widely used in previous works [6, 61–63] 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 3.5.1). The intra-modality features are trained using adadelta optimizer [64] since it can automatically adapt learning rate. We used the *Early-Stopping* [65] technique to terminate training intra-modality features if there is no improvement on the validation loss for 5 epochs. The choice of other hyper-parameters is listed in table 3.1.

We set $\omega$ to $L/4 = 0.5s$ 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 $\omega$ 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.

### 3.4.2. EVALUATION PROTOCOL AND BASELINES

| Hyper-parameter | Meaning | Value | |
|---|---|---|---|
| | | CASE | MERCA |
| **L** | Length of each instance | 2s (100) | 2s(64) |
| $\boldsymbol{\omega}$ | 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) |

Table 3.1: The value of hyper-parameters in *CorrNet*

### 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 3.2.

| 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) |

Table 3.2: The mapping of V-A values and discretized classes.

### 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 [66];
- ***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) [67].

These three metrics are widely used in evaluating machine learning algorithms [68]. We use weighted F1-score instead of macro and binary F1-score to take into account label imbalance.

### 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 9 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.

### 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 [69] 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 [32, 34] and Bidirectional LSTM (BiLSTM) [33, 70] because they are widely used for the classification of time series. We train the 1D-CNN, LSTM and BiLSTM with the adadelta optimizer [64], which is the same as we used for training the intra-modality features. For ML methods, we compare with Support Vector Machine (SVM) [71], K-Nearest Neighbor (KNN) [72], Random Forest (RF) [73] and Gaussian Naive Bayes (GaussianNB) [74]. These methods are commonly used as baseline methods in datasets [75, 76] and review [6, 7, 77] papers for affective computing. To train these ML models, we first pre-processed the signals using the same method we described in section 3.4.1. 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].

## 3.4.3. Experiment results

Performance of *CorrNet* on CASE and MERCA is shown in Table 3.3. 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 (4 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 2 classes on V-A, respectively). Thus, the 3-class testing (high/neutral/low) on V-A independently is more challenging than 4 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).

| | 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.

Table 3.3: Validation results for CASE and MERCA.

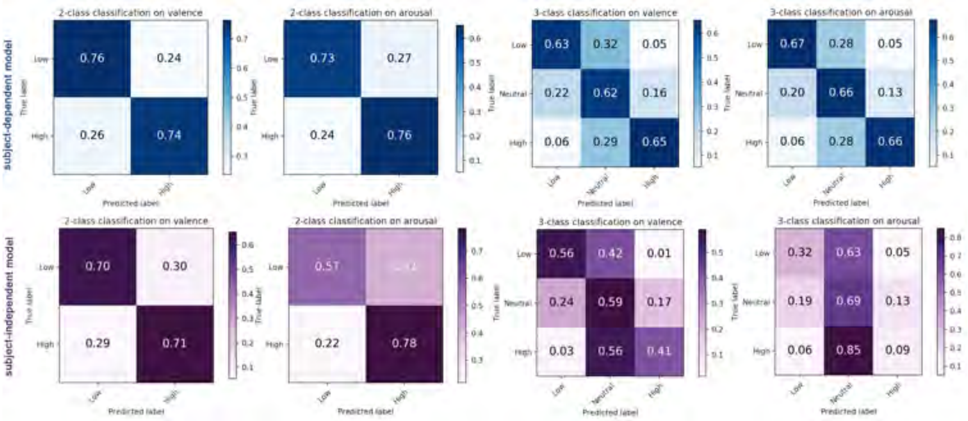


Figure 3.2: The result of 10-fold cross validation (subject-dependent model, up) and leave-one-subject-out cross validation (subject-independent model, down) on MERCA

### 3.4.4. Comparison with DL and ML methods

The comparison of DL and ML methods with *CorrNet* using a subject-independent model is shown in Table 3.4. Compared with subject-dependent models, the subject-independent model is more challenging for training, which lead to less subject-bias (overfiting 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 3.4, 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

**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. **2.** 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 3.4: Comparison between ML, DL methods and CorrNet using LOSOCV (accuracy (W-F1))

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.

### 3.4.5. 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 3.5.

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

| | 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)** |

Table 3.5: Ablation study of different components in *CorrNet* (accuracy (W-F1))

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.

### 3.4.6. 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 65.56s and 24.67s for CASE and MERCA, respectively (sampling rate = 50Hz). The average detection time for each fine-grained instance is 29.01ms, which means to recognize 2s emotions, the algorithm only spends less than 30ms after the network is trained.

## 3.5. DISCUSSION

### 3.5.1. 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

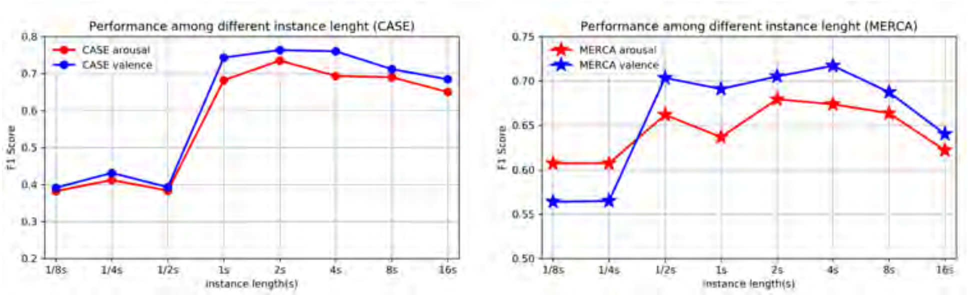


Figure 3.3: Comparison of the performance among different instance length: W-F1 of binary classification (LOSOCV)

*CorrNet* using different segmentation lengths. As shown in Figure 3.3, the W-F1 tested on CASE drops significantly after reducing the length to 0.5s while the dropping threshold for MERCA is 0.25s. This finding is in line with the finding from Paul et al. [13] that the duration of an emotion typically spans 0.5-4s. The W-F1 also decreases after increasing the length to 8s. 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 1s-4s can result in good performance (high W-F1), which can serve as an appropriate length to classify emotions using fine-grained emotion labels.

### 3.5.2. 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 ($\geq 1000Hz$). 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 1000Hz, we gradually down-sample the signals from CASE to 1Hz 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 $S = [s_1, s_2, \dots, s_N]$ and the signal after down-sampling $X$ is:

$$X = [x_{1M}, x_{2M}, \dots, x_{kM}] \tag{3.12}$$

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

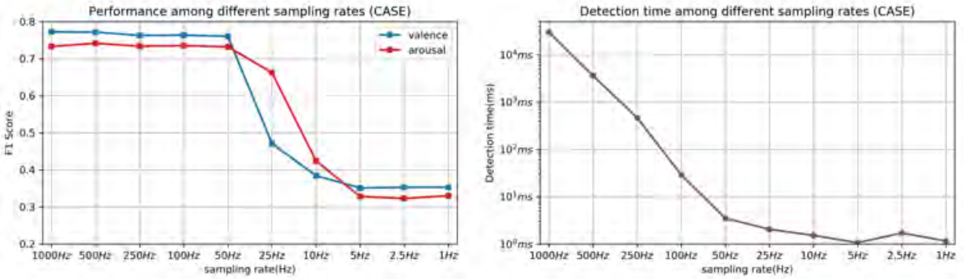


Figure 3.4: Comparison of the performance among different sampling rates: W-F1 of binary classification (LOSOCV, left) and detection time (right)

Figure 3.4 shows the weighed F1 score and detection time among different sampling rates. As shown in Figure 3.4 (left), down-sampling to 50Hz 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 50Hz. This result helps explain why for most of the wearable devices (e.g., Empatica E4 wristband, BITalino Kit [2]), the highest sampling rate of physiological sensors is less than 64Hz (e.g., 32Hz for Empatica E4 and 40Hz for BITalino Kit). The comparable recognition accuracy testing on the CASE and MERCA datasets also shows low sampling rates (32Hz) 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.

### 3.5.3. DATA IMBALANCE AND OVERFITTING IN FINE-GRAINED EMOTION RECOGNITION

As shown in Figure 3.2 (down, LOSOCV)), there is an accuracy imbalance among different classes for 3-class classification (for binary classification we did not omit neutral labels but discretize them according to Table 3.2). 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,

[2] https://bitalino.com/

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.

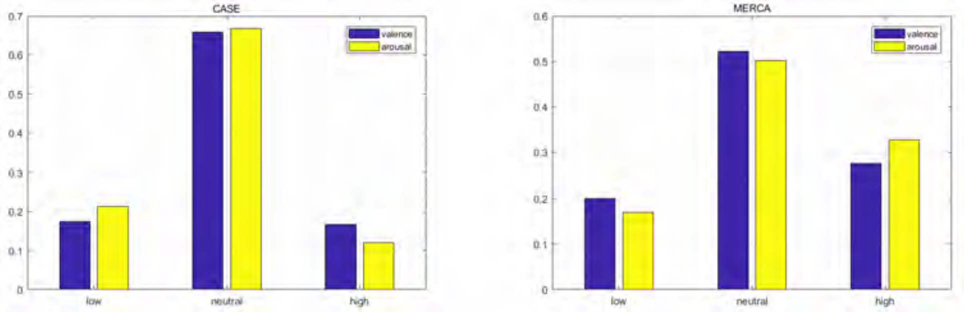


Figure 3.5: Sample percentage in each class of V-A.

As shown in Figure 3.5, 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 [78]. 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.

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 [79]. 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.

## 3.6. Conclusion and limitations

Physiological signals from different modalities contain different aspects of human emotions. In this chapter, 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 show that 1) compared with sequential learning, correlation-based instance learning offer advantages of higher recognition accuracy, less overfitting and less computational complexity (**RQ 1.1**) and 2) large amounts of neutral V-A labels, an artifact of continuous affect annotation, result in varied recognition performance in different emotion categories (**RQ 1.2**).

Given the challenges of designing for fine-grained emotion recognition, there were naturally limitations to our work. Although the instance-level fully-supervised learning methods developed in this chapter can obtain accurate results, this method require segment-by-segment, fine-grained emotion labels to train the recognition algorithm. However, experiments to collect these labels are costly and time-consuming. In addition, according to our experiments, the fully-supervised learning methods can also cause the problem of overfitting. 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. To address the limitations mentioned above, in **Chapter 4** we explore weakly-supervised learning algorithms which recognize emotions in a fine level of granularity by training with only post-stimuli emotion labels.

## REFERENCES

[1] M. Soleymani, M. Pantic, and T. Pun, *Multimodal emotion recognition in response to videos,* IEEE transactions on affective computing **3**, 211 (2011).

[2] J. Niu, X. Zhao, L. Zhu, and H. Li, *Affivir: An affect-based internet video recommendation system,* Neurocomputing **120**, 422 (2013).

[3] A. Tripathi, T. Ashwin, and R. M. R. Guddeti, *Emoware: A context-aware framework for personalized video recommendation using affective video sequences,* IEEE Access **7**, 51185 (2019).

[4] A. Yazdani, J.-S. Lee, J.-M. Vesin, and T. Ebrahimi, *Affect recognition based on physiological changes during the watching of music videos,* ACM Transactions on Interactive Intelligent Systems (TiiS) **2**, 1 (2012).

[5] M. Ali, F. Al Machot, A. Haj Mosa, M. Jdeed, E. Al Machot, and K. Kyamakya, *A globally generalized emotion recognition system involving different physiological signals,* Sensors **18**, 1905 (2018).

[6] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, *A review of emotion recognition using physiological signals,* Sensors **18**, 2074 (2018).

[7] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, *Physiological signals based human emotion recognition: a review,* in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications* (IEEE, 2011) pp. 410–415.

[8] E. Maria, L. Matthias, and H. Sten, *Emotion recognition from physiological signal analysis: a review,* Electronic Notes in Theoretical Computer Science **343**, 35 (2019).

[9] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, *Emujoy: Software for continuous measurement of perceived emotions in music,* Behavior Research Methods **39**, 283 (2007).

[10] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, *Analysis of eeg signals and facial expressions for continuous emotion detection,* IEEE Transactions on Affective Computing **7**, 17 (2015).

[11] P. J. Lang, *The emotion probe: studies of motivation and attention.* American psychologist **50**, 372 (1995).

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

[13] E. Paul, *Emotions revealed: recognizing faces and feelings to improve communication and emotional life,* NY: OWL Books (2007).

[14] R. W. Levenson, *Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity,* Social psychophysiology: Theory and clinical applications (1988).

[15] J. Domínguez-Jiménez, K. Campo-Landines, J. Martínez-Santos, E. Delahoz, and S. Contreras-Ortiz, *A machine learning model for emotion recognition from physiological signals,* Biomedical Signal Processing and Control **55**, 101646 (2020).

[16] M. M. Bradley and P. J. Lang, *Measuring emotion: the self-assessment manikin and the semantic differential,* Journal of behavior therapy and experimental psychiatry **25**, 49 (1994).

[17] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, *'feeltrace': An instrument for recording perceived emotion in real time,* in *ISCA tutorial and research workshop (ITRW) on speech and emotion* (2000).

[18] J. M. Girard and A. G. Wright, *Darma: Software for dual axis rating and media annotation,* Behavior research methods **50**, 902 (2018).

[19] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* Scientific data **6**, 1 (2019).

[20] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, *Continuous emotion detection using eeg signals and facial expressions,* in *2014 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2014) pp. 1–6.

[21] S. Haripriyadharshini and S. Gnanasaravanan, *Eeg based human facial emotion recognition system using lstmrnn,* .

[22] F. Hasanzadeh, M. Annabestani, and S. Moghimi, *Continuous emotion recognition during music listening using eeg signals: A fuzzy parallel cascades model,* arXiv preprint:1910.10489 (2019).

[23] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, *Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze,* in *2019 International Conference on Multimodal Interaction* (2019) pp. 40–48.

[24] S. Zhao, H. Yao, and X. Jiang, *Predicting continuous probability distribution of image emotions in valence-arousal space,* in *Proceedings of the 23rd ACM international conference on Multimedia* (2015) pp. 879–882.

[25] A. Craik, Y. He, and J. L. Contreras-Vidal, *Deep learning for electroencephalogram (eeg) classification tasks: a review,* Journal of neural engineering **16**, 031001 (2019).

[26] A. J. Casson, *Wearable eeg and beyond,* Biomedical engineering letters **9**, 53 (2019).

[27] M. Khamis, A. Baier, N. Henze, F. Alt, and A. Bulling, *Understanding face and eye visibility in front-facing cameras of smartphones used in the wild,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems,* CHI ?18 (Association for Computing Machinery, New York, NY, USA, 2018) p. 1?12.

[28] B. Friedman, P. H. Kahn Jr, J. Hagman, R. L. Severson, and B. Gill, *The watcher and the watched: Social judgments about privacy in a public place,* Human-Computer Interaction **21**, 235 (2006).

[29] T. L. Stanko and C. M. Beckman, *Watching you watching me: Boundary control and capturing attention in the context of ubiquitous technology use,* Academy of Management Journal **58**, 712 (2015).

[30] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, *Emotion recognition using physiological signals: laboratory vs. wearable sensors,* in *International Conference on Applied Human Factors and Ergonomics* (Springer, 2017) pp. 15–22.

[31] S. Gashi, E. Di Lascio, B. Stancu, V. D. Swain, V. Mishra, M. Gjoreski, and S. Santini, *Detection of artifacts in ambulatory electrodermal activity data,* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **4**, 1 (2020).

[32] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, *Emotion recognition using multimodal residual lstm network,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 176–183.

[33] S.-h. Zhong, A. Fares, and J. Jiang, *An attentional-lstm for improved classification of brain activities evoked by images,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 1295–1303.

[34] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, *Lstm: A search space odyssey,* IEEE transactions on neural networks and learning systems **28**, 2222 (2016).

[35] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks,* in *Advances in neural information processing systems* (2014) pp. 3104–3112.

[36] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, *Enhanced lstm for natural language inference,* arXiv preprint:1609.06038 (2016).

[37] F. Bentley and D. Lottridge, *Understanding mass-market mobile tv behaviors in the streaming era,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '19 (ACM, New York, NY, USA, 2019) pp. 261:1–261:11.

[38] A. Costa, J. A. Rincon, C. Carrascosa, V. Julian, and P. Novais, *Emotions detection on an ambient intelligent system using wearable devices,* Future Generation Computer Systems **92**, 479 (2019).

[39] A. Zenonos, A. Khan, G. Kalogridis, S. Vatsikas, T. Lewis, and M. Sooriyabandara, *Healthyoffice: Mood recognition at work using smartphones and wearable sensors,* in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)* (IEEE, 2016) pp. 1–6.

[40] S. Puke, T. Suzuki, K. Nakayama, H. Tanaka, and S. Minami, *Blood pressure estimation from pulse wave velocity measured on the chest,* in *2013 35th Annual International Conference of the IEEE engineering in medicine and biology society (EMBC)* (IEEE, 2013) pp. 6107–6110.

[41] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, *Engagemon: Multi-modal engagement sensing for mobile games,* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1 (2018).

[42] S. Gashi, E. Di Lascio, and S. Santini, *Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members,* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**, 1 (2019).

[43] E. Di Lascio, S. Gashi, and S. Santini, *Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors,* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1 (2018).

[44] E. Fernandez, C. Gangitano, A. Del Fà, C. O. Sangiacomo, G. Talamonti, F. Draicchio, and A. Sbriccoli, *Oculomotor nerve regeneration in rats: functional, histological, and neuroanatomical studies,* Journal of neurosurgery **67**, 428 (1987).

[45] M. R. Ibbotson, N. A. Crowder, S. L. Cloherty, N. S. Price, and M. J. Mustari, *Saccadic modulation of neural responses: possible roles in saccadic suppression, enhancement, and time compression,* Journal of Neuroscience **28**, 10952 (2008).

[46] R. W. Picard, *Future affective technology for autism and emotion communication,* Philosophical Transactions of the Royal Society B: Biological Sciences **364**, 3575 (2009).

[47] J. L. Greaney, W. L. Kenney, and L. M. Alexander, *Sympathetic regulation during thermal stress in human aging and disease,* Autonomic Neuroscience **196**, 81 (2016).

[48] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, *Deep features learning for medical image analysis with convolutional autoencoder neural network,* IEEE Transactions on Big Data (2017).

[49] A. Creswell, K. Arulkumaran, and A. A. Bharath, *On denoising autoencoders trained to minimise binary cross-entropy,* arXiv preprint:1708.08487 (2017).

[50] S. C. AP, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha, *An autoencoder approach to learning bilingual word representations,* in *Advances in neural information processing systems* (2014) pp. 1853–1861.

[51] T. Zhang, A. El Ali, C. Wang, X. Zhu, and P. Cesar, *Corrfeat: Correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition,* in *International Conference on Multimodal Interaction* (2019) pp. 404–408.

[52] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, *Deep canonical correlation analysis,* in *International Conference on Machine Learning* (2013) pp. 1247–1255.

[53] C. P. Chen and Z. Liu, *Broad learning system: An effective and efficient incremental learning system without the need for deep architecture,* IEEE transactions on neural networks and learning systems **29**, 10 (2018).

[54] F. Movahedi, J. L. Coyle, and E. Sejdić, *Deep belief networks for electroencephalography: A review of recent contributions and future outlooks,* IEEE journal of biomedical and health informatics **22**, 642 (2018).

[55] C. Liu, T. Tang, K. Lv, and M. Wang, *Multi-feature based emotion recognition for video clips,* in *Proceedings of the on International Conference on Multimodal Interaction* (ACM, 2018) pp. 630–634.

[56] H. Chen, B. Jiang, and S. X. Ding, *A broad learning aided data-driven framework of fast fault diagnosis for high-speed trains,* IEEE Intelligent Transportation Systems Magazine (2020).

[57] A. Jain, K. Nandakumar, and A. Ross, *Score normalization in multimodal biometric systems,* Pattern recognition **38**, 2270 (2005).

[58] B. A. Olshausen and D. J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by v1?* Vision research **37**, 3311 (1997).

[59] E. Meijering, *A chronology of interpolation: from ancient astronomy to modern signal and image processing,* Proceedings of the IEEE **90**, 319 (2002).

[60] R. W. Daniels, *Approximation methods for electronic filter design: with applications to passive, active, and digital networks* (McGraw-Hill New York, NY, USA:, 1974).

[61] J. Fleureau, P. Guillotel, and I. Orlac, *Affective benchmarking of movies based on the physiological responses of a real audience,* in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (IEEE, 2013) pp. 73–78.

[62] Y. Chu, X. Zhao, J. Han, and Y. Su, *Physiological signal-based method for measurement of pain intensity,* Frontiers in neuroscience **11**, 279 (2017).

[63] P. Karthikeyan, M. Murugappan, and S. Yaacob, *Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress,* Journal of Physical Therapy Science **24**, 1341 (2012).

[64] M. D. Zeiler, *Adadelta: an adaptive learning rate method,* arXiv preprint:1212.5701 (2012).

[65] L. Prechelt, *Early stopping-but when?* in *Neural Networks: Tricks of the trade* (Springer, 1998) pp. 55–69.

[66] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, *Multiple instance learning for emotion recognition using physiological signals,* IEEE Transactions on Affective Computing (2019).

[67] N. Chinchor, *Muc-3 evaluation metrics,* in *Proceedings of the 3rd conference on Message understanding* (Association for Computational Linguistics, 1991) pp. 17–24.

[68] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, *Comparison of evaluation metrics in classification applications with imbalanced datasets,* in *2008 Seventh International Conference on Machine Learning and Applications* (IEEE, 2008) pp. 777–782.

[69] J. Schmidhuber, *Deep learning in neural networks: An overview,* Neural networks **61**, 85 (2015).

[70] Z. Huang, W. Xu, and K. Yu, *Bidirectional lstm-crf models for sequence tagging,* arXiv preprint:1508.01991 (2015).

[71] C. He, Y.-j. Yao, and X.-s. Ye, *An emotion recognition system based on physiological signals obtained by wearable sensors,* in *Wearable sensors and robots* (Springer, 2017) pp. 15–25.

[72] L. Chen, M. Li, W. Su, M. Wu, K. Hirota, and W. Pedrycz, *Adaptive feature selection-based adaboost-knn with direct optimization for dynamic emotion recognition in human–robot interaction,* IEEE Transactions on Emerging Topics in Computational Intelligence (2019).

[73] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis, *A user independent, biosignal based, emotion recognition method,* in *International Conference on User Modeling* (Springer, 2007) pp. 314–318.

[74] D. S. Wickramasuriya and R. T. Faghih, *Online and offline anger detection via electromyography analysis,* in *2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)* (IEEE, 2017) pp. 52–55.

[75] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, *Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos),* IEEE Access **7**, 57 (2018).

[76] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *Deap: A database for emotion analysis; using physiological signals,* IEEE transactions on affective computing **3**, 18 (2011).

[77] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, *Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications,* International journal of human-computer studies **72**, 717 (2014).

[78] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI ’20 (Association for Computing Machinery, New York, NY, USA, 2020).

[79] H. D. Critchley and S. N. Garfinkel, *Interoception and emotion,* Current opinion in psychology **17**, 7 (2017).

# 4

# WEAKLY-SUPERVISED LEARNING FOR FINE-GRAINED EMOTION RECOGNITION

*This chapter investigates whether post-stimuli emotion labels (the labels user annotated after watching videos) can be used to recognize emotions at a fine granularity level. For this purpose, we propose an emotion recognition algorithm based on Deep Multiple Instance Learning (EDMIL) using physiological signals. EDMIL recognizes fine-grained valence and arousal (V-A) labels by identifying which instances represent the post-stimuli V-A annotated by users after watching the videos. Instead of fully-supervised training, the instances are weakly-supervised by the post-stimuli labels in the training stage. The V-A of instances are estimated by the instance gains, which indicate the probability of instances to predict the post-stimuli labels. Recognition results validated with the fine-grained V-A self-reports show that for subject-independent 3-class classification (high/neutral/low), EDMIL obtains promising recognition accuracies: 75.63% and 79.73% for V-A on CASE, 70.51% and 67.62% for V-A on MERCA and 65.04% and 67.05% for V-A on CEAP-360VR. Our ablation study shows that all components of EDMIL contribute to both the classification and regression tasks.*

*This chapter is published as Zhang T, El Ali A, Wang C, Hanjalic A, Cesar P. Weakly-supervised Learning for Fine-grained Emotion Recognition using Physiological Signals. IEEE Transaction on Affective Computing, March 2022*

## 4.1. Introduction

Recent years have witnessed a growing number of emotion recognition algorithms [1–4] that particularly focus on modeling the temporal dynamics of emotion states. Previous works [4–6] employ sequential machine learning algorithms such as Long Short Term Memory (LSTM) networks [7] to model the relationship between input signals and emotion states. However, sequential learning algorithms require fine-grained emotion labels for training. Here, the emotion labels and the input signals are required to have the same dimensions to train the recurrent structure of sequential learning algorithms [8]. To collect such fine-grained emotion labels (e.g., valence and arousal), there are typically three kinds of methods: (1) interrupt users at a fixed frequency for annotation [9], (2) ask users to annotate their emotions in real-time while watching videos [10, 11] or (3) let external observers annotate users' emotions segment-by-segment (e.g., using videos of users' facial expressions [12]) after watching videos [12–14]. However each of those methods has limitations. Requesting users to continuously annotate their emotions may not be feasible for longer durations (e.g., two hour film) as it may result in participant fatigue. Continuously interrupting people to self-report their emotional states can disrupt users' tasks [15]. For external observers, some emotional states are difficult or misleading for them to annotate. For example, according to the experiments of Song et al. [16] and Abdic et al. [17], negative valence is often misidentified by external annotators as positive when users smile because of sarcasm and frustration. Even if collecting fine-grained emotion labels is possible, the experiments to collect them are time-consuming and costly [11]. Researchers have to spend extra time and money collecting fine-grained emotion labels because it is an additional task other than the data collection experiment (e.g., asking users to (re-)watch videos).

Such challenges of collecting fine-grained emotion labels lead to researchers developing datasets such as DEAP [18], Mahnob-HCI [19] and ASCERTAIN [20] containing only post-stimuli labels. Instead of multiple labels for every fine-grained time segment (instance), there is only one post-stimuli label for one activity (i.e., one user watches one video). Taking advantage of these datasets can lower the cost of developing and training fine-grained emotion recognition algorithms. However, according to the *peak-end theory* [21], the post-stimuli labels represent only the most salient (peak) or recent (end) emotion during the video watching rather than the naturally dynamic and subtle emotional changes that may occur within it. According to Romeo et al. [15], this is defined as the problem of *temporal ambiguity*. When training machine learning algorithms to recognize fine-grained emotions using post-stimuli labels, the information on which fine-grained instances represent the emotion users labeled post-stimuli is missing. This can lead to overfitting [3, 22, 23] if all the instances are fully-supervised by the post-stimuli labels.

To overcome the challenge of temporal ambiguity, this chapter proposes an emotion recognition algorithm based on Deep Multiple Instance Learning (*EDMIL*) using physiological signals. *EDMIL* is trained only with post-stimuli emotion labels. However, it can provide recognition results at a finer (or higher) level of granularity (every 2s) by identifying which instances represent the emotion annotated by users after watching videos. The ground truth labels (i.e., valence and arousal (V-A)) we use are based on Russell's Circumplex model [24], which describes emotions in a continuous

2-dimensional space. Valence indicates users' positive or negative affectivity. Arousal measures how calm or excited a user is. Although we use V-A for training, the prediction of *EDMIL* can be easily mapped to discrete emotion keywords (e.g., high valence and high arousal = happy, high valence and low arousal = relax) [25]. The signals and their fine-grained segments are viewed as bags and instances, respectively. Instead of implementing fully-supervised training for all the instances using post-stimuli labels, the instances are weakly-supervised by the post-stimuli labels to avoid overfitting. The fine-grained V-A of instances are then estimated by the instance gains, which represent the probability for that instance to predict the corresponding bag label. This work makes the following contributions to Affective Computing research:

- We propose an end-to-end deep multiple instance learning framework to identify which instances represent the post-stimuli V-A in a fine level of granularity using physiological signals. Our algorithm is tested on three datasets (CASE [11], MERCA [10] and CEAP-360VR [26]) collected in three environments (desktop, mobile, and HMD-based Virtual Reality (VR)). Recognition results show good performance for subject-independent 3-class (high/neutral/low) classification on all three datasets: 75.63% and 79.73% for V-A on *CASE*, 70.51% and 67.62% for V-A on *MERCA* and 65.04% and 67.05% for V-A on *CEAP-360VR*. Our framework enables finding an optimal trade-off between recognition accuracy and the burden of fine-grained emotion annotation.

- We test both state-of-the-art weakly-supervised and fully-supervised machine learning methods and compare their performance with *EDMIL*. Results show that *EDMIL*'s recognition accuracy outperforms both weakly-supervised and fully-supervised learning methods for fine-grained emotion recognition. We also find that compared with fully-supervised learning, weakly supervised learning can reduce overfitting that results from the temporal mismatch between fine-grained annotations and input signals.

- We run validation experiments to compare the performance of *EDMIL* under different instance lengths and feature extraction methods. Results show that instance segment lengths between 1-2s result in the highest recognition accuracies (up to 60% for V-A in all three datasets). Our results also show that feature extraction using an end-to-end structure can improve recognition accuracy compared with manual feature extraction and unsupervised learning feature extraction methods.

## 4.2. RELATED WORK

In the paradigm of Multiple Instance Learning (MIL), the input is a set of *bags* which are composed of multiple *instances*. At the training stage, each bag has a corresponding label while each instance does not. Thus, not all the instances are labeled the same as the bag label at the training stage [27, 28]. MIL has been applied in previous works on emotion recognition using a variety of data modalities such as images [29], text [30], voice [31] and physiological signals [15]. For physiological signals, Romeo et al. [15] evaluated four MIL algorithms (mi-SVM [32], mil-Boost [33], MI-SVM [32] and EMDD-SVM [34]) for emotion recognition using physiological signals (without EEG) on DEAP [18] and Consumer [15] dataset. The two datasets are collected using golden-standard (for DEAP) and unobtrusive consumer devices (for Consumer). Their results show that mi-SVM and MI-SVM achieve the highest recognition accuracies (bag level) on DEAP dataset, which is 63.6% and 61.1% for valence and arousal, respectively. The hypothesis of the four methods mentioned above is that the positive bags are fairly rich in positive instances. Thus, the negative instances can be easily identified. However, positive bags can contain only a small fraction of positive instances. To solve this problem, Bunescu et al. [35] designed Balanced MIL (sb-MIL) which introduced a balancing constraint between positive and negative instances to model the sparse positive instances in different bags. Zhang et al. [36] implemented the proposed sb-MIL [35] to classify dimensional emotions using EEG signals. They achieved classification accuracies (bag level) on DEAP [18] dataset of 74.21% and 77.50% for valence and arousal, respectively.

The general idea of the MIL algorithms mentioned above is to identify the instances which contribute to maximize the probability for predicting the bag labels. However, all these methods need to manually design loss functions or constraints between instances and bags [32, 35]. The manually designed functions or constraints are usually just suitable for one condition (e.g., most of the instances are labeled the same as the bag label [32]). In addition, the previous works mentioned above only recognize emotions at the bag level, which means they recognize only one emotion instead of the fine-grained emotion response for each instance (i.e., instance-level recognition). In the work of Roman et al. [15], the authors attempt to identify the instances which make contributions to predicting the bag-level labels. However, since the two datasets they use do not have fine-grained emotion ground truth labels, they plan to validate their method for fine-grained emotion in future work.

Compared with other learning tasks, the special character of fine-grained emotion recognition using physiological signals is manifested into two aspects. First of all, most of the existing multi-instance learning methods [29, 37, 38] for emotion recognition are designed for 2D images. The purpose of these methods is to identify an object of interest embedded into the image (also known as "*emotional regions which contain objects and concepts*" according to the definition of Zhao et al. [39]). Thus, the instances (i.e., small segments of the image) are often aggregated in a dense spatial space. Strong constraint functions are often implemented to omit instances which are spatially far away from the region with the highest probability of an object of interest. For example, in the work of Rao et al. [29], a linear iterative clustering is used to merge different regions of interest representing the emotion of images. Compared with spatial differences of emotions from an image, the dynamics of emotions are sparsely distributed in the temporal space: users can have an emotion response with a short duration (0.5s to 2s) [40, 41]. Thus, we did not implement strong constraints to filter the instances which have high probability of predicting the emotions but are not densely aggregated in one temporal moment. We use a simple threshold based on the distribution of the instance gains for identifying which instances correspond to the post-stimuli emotion labels.

Secondly, compared with learning tasks using other temporal signals (e.g., video, speech, EEG signals), the physiological signals we use contain less abundant information for emotion recognition [15, 42]. This limits the performance of DNN methods: feature extraction layers with deep structures can easily overfit or fail to extract meaningful features for recognition [3]. Due to this, we use shallow convolutional layers (5-layers) with gradually increasing number of filters and lower size kernels for feature extraction. This character also motivates us to compare two different types of feature extraction methods in section 4.4.6 to find out whether the end-to-end deep-network-based feature extraction is suitable for fine-grained emotion recognition using physiological signals.

## 4.3. DEEP MIL BASED EMOTION RECOGNITION

In this section, we propose a deep multiple instance learning based emotion recognition algorithm (*EDMIL*) to identify the post-stimuli dimensional emotions (i.e., valence and arousal (V-A)) at a fine granularity level from physiological signals. *EDMIL* recognizes fine-grained V-A by identifying which instances represent the post-stimuli emotion labels annotated by users after watching the videos. In the training stage, *EDMIL* contains four parts: (1) **Pre-processing:** the obtained physiological signals are firstly filtered and grouped into bags and instances as input for *EDMIL*. (2) **Feature extraction layers**: the grouped signals are then passed into deep convolutional layers for feature extraction. (3) **Multiple instance learning layers**: the extracted features are then input into multiple instance learning layers to obtain the instance gain for each instance. The instance gain represents the probability for each instance to predict the bag label. (4) **Fully connected layer:** at last, each instance gain is fully connected with the post-stimuli emotion labels (i.e., valence or arousal). The training network is designed to learn the data representation to predict the post-stimuli emotion labels using the entire signals. The instance gains learned in part (3) are the matching scores which indicate the probability that the instance contributes to the prediction of the post-stimuli emotion label. In the prediction stage (the network has already been trained and fixed), the obtained signals are first forwarded from (1) to (3) to get the instance gains. After that, the (5) **instance regularization** is used to transfer the instance gains into the V-A for each instance. The architecture of the algorithm is shown in Figure 4.1. When describing and validating *EDMIL*, we use the physiological signals as input and specify the application scenario as video watching.

### 4.3.1. PRE-PROCESSING

We first pre-process all the physiological signals using different filters to eliminate the noise and artifacts from the measurement. The details for this process are described in section 4.4.1 (implementation details). Suppose $S_{mn} = \{s_c\}_{c=1}^{C}$ is the set of pre-processed physiological signals for one user $m$ watching one entire video $n$, where $C$ is the number (channels) of physiological signals. The signals are firstly segmented into multiple instances with a fixed instance $L$. After the segmentation, the input of the algorithm is transferred into a bag of instances: $B = \{b_g\}_{g=1}^{G}$. $G$ is the number of samples for training. $b_g = \{x_g^i\}_{i=1}^{I}$ is the bag $g$ and $x_g^i$ is the instance $i$ in bag $g$. $b_g \in R^{L\times I\times C}, x_g^i \in R^{I\times C}$, where $L$ and $I$ is the number of instances in one bag and the length for one instance, respectively. The goal of *EDMIL* is to predict the V-A for each instance. For both the training and prediction stage, only the ground truth labels for $b_g$ are available. The ground truth labels for $x_g^i$ are only used to evaluate the performance of *EDMIL*.

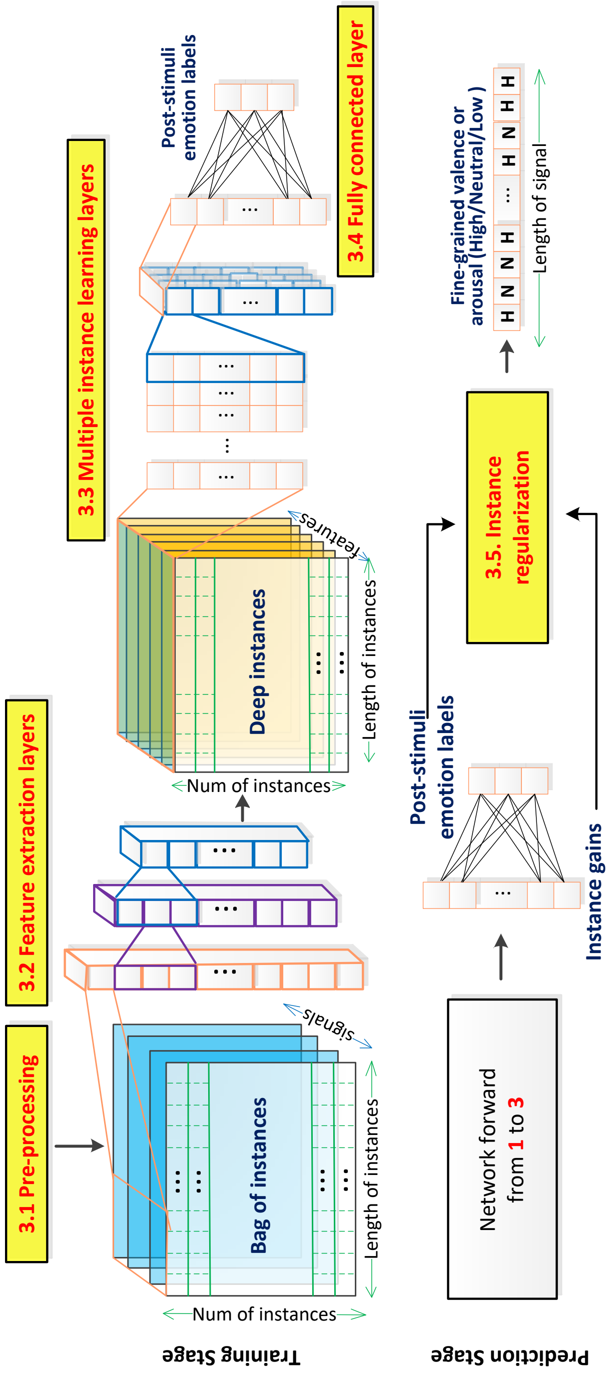

Training Stage
Prediction Stage
3.1 Pre-processing
3.2 Feature extraction layers
3.3 Multiple instance learning layers
3.4 Fully connected layer
Bag of instances
signals
Length of instances
Num of instances
Deep instances
features
Length of instances
Num of instances
Post-stimuli emotion labels
Network forward from 1 to 3
Post-stimuli emotion labels
Instance gains
3.5. Instance regularization
Fine-grained valence or arousal (High/Neutral/Low )
H N N H ... H N H H
Length of signal

Figure 4.1: The architecture of proposed EDMIL

### 4.3.2. FEATURE EXTRACTION LAYERS

The feature extraction layers are designed to learn the deep features from the physiological signals for recognizing the post-stimuli labels. The features are extracted from each instance $x_g^i$ independently, which means the feature extraction layers will not influence the independence between each set of instances (no features are extracted from multiple instances). This operation guarantees that each instance has a unique instance gain before the fully connected layer. Here, three types of feature extraction methods are implemented for comparison: (1) an end-to-end feature extraction method using one-dimensional convolutional neural network (1D-CNN) (*deepfeat*), (2) an unsupervised feature extraction method by maximizing the correlation coefficients between pairwise physiological signals (*pcorrfeat*) and (3) a manual feature extraction method using statistical features (*manualfeat*). Unless otherwise specified, we use *deepfeat* as the default feature extraction method.

Theoretically, the end-to-end model should result in the best performance as the features are directly connected with the ground truth labels[43], which means the deep representation is trained to best recognize these labels. However, according to previous studies [3, 44], if we train the network using fine-grained emotion labels and fully-supervised learning methods, the end-to-end model will overfit because of the temporal resolution mismatch between physiological signals and fine-grained self-reports due to different interoception levels across individuals [45]. Thus, we compare these three types of methods to find out whether the end-to-end, deep feature extraction (*deepfeat*, section 4.3.2) still has the problem of overfitting for weakly-supervised learning. Manual feature extraction methods (*manualfeat*, section 4.3.2) are widely used by previous works for emotion recognition [25, 46, 47]. Thus, we choose it as a baseline method for comparison. We additionally compare *deepfeat* with an unsupervised feature extraction method (*pcorrfeat*, section 4.3.2) because it can decrease overfitting compared with end-to-end models, as shown in prior work by Zhang et al. [3]. Below, we introduce the details of the three feature extraction methods.

#### DEEPFEAT

The deep features (*deepfeat*) are extracted using a 5-layer 1D-CNN [48]. The parameters for each convolutional layer are shown in TABLE 4.1. We use large (i.e., equals to half of the instance length) convolutional kernels in the beginning of the network. Large convolutional kernels commonly result in better recognition accuracy [49] because they have a large receptive field across different sampling points in one instance. However, large kernels can omit the local information and make the network more difficult to converge [50]. Thus, we follow a classical strategy that gradually increases the number of kernels and decreases the size of them when the network goes deeper [51, 52]. At last, we add a convolutional layer whose size is bigger than the previous layer to merge the local information learned by small kernels.

After the feature extraction layer, the bag of instances $B$ is transferred into deep instances $D = \{d_g\}_{g=1}^G, d_g = \{f_g^k\}_{k=1}^K$ where $K = 128$ is the dimension of features at the last 1D-CNN layer.

Table 4.1: Architecture of the 1D-CNN to extract *deepfeat*

| layer | input size | channels | kernel size | output size |
|---|---|---|---|---|
| **input** | (I,C) | 8 | I/2+1* | (I,8) |
| **conv1** | (I,8) | 16 | I/3 | (I,16) |
| **conv2** | (I,16) | 32 | I/4+1* | (I,32) |
| **conv3** | (I,32) | 64 | I/8+1* | (I,64) |
| **conv4** | (I,64) | 128 | I/12+1* | (I,128) |
| **conv5** | (I,128) | 128 | I/8+1* | (I,128) |

*We add 1 to some of the kernels to make their size an odd number

### PCORRFEAT

The pairwise correlation-based features (*pcorrfeat*) are extracted by maximizing the correlation coefficient for every two signals from users who watch the same video stimuli [44]. The idea is inspired by the hypothesis that the same stimuli will trigger relatively similar emotions across physiological responses among different users [53, 54]. To extract correlation-based features, we first calculate the covariance ($C_{11}$ and $C_{22}$) and cross-covariance ($C_{12}$) of the two signals ($S^1_n, S^2_n$) for users who watch the same video stimuli. After that, we implement the Singular Value Decomposition (SVD) on the equation below:

$$[U, D, V] = \mathrm{SVD}(\mathrm{V_1 D_1 V_1^T \cdot C_{12} \cdot V_2 D_2 V_2^T}) \tag{4.1}$$

where $D_1$ and $D_2$ are diagonal matrices whose diagonal elements are the $\omega$ biggest non-zero eigenvalues of $C_{11}$ and $C_{22}$, respectively. $D_1 = \mathrm{diag}(\frac{1}{\sqrt{\mathrm{D_{11}}}}, \frac{1}{\sqrt{\mathrm{D_{12}}}}, \dots, \frac{1}{\sqrt{\mathrm{D_{1\omega}}}})$ and $D_2$ have the same format. $V_1 = [V_{11}, V_{12}, \dots, V_{1\omega}]$ is composed of the $\omega$ corresponding eigenvectors of $[D_{11}, D_{12}, \dots, D_{1\omega}]$, respectively. $V_2$ is calculated using the same method. We then obtain two linear projections $[H^t_1, H^t_2] = [V_1 D_1 V_1^T \cdot U', V_2 D_2 V_2^T \cdot V']$, where $U'$ and $V'$ consist of the first $K$ columns of $U$ and $V$, respectively. At last, the correlation-based features of $S^t_1$ and $S^t_2$ can be obtained by: $F^t = [S^t_1 \cdot H^t_1, S^t_2 \cdot H^t_2]$. We then implement the above procedure among all the $M$ stimuli and $C$ signals (pair by pair). At last, the bag of instances $B$ is transferred into *pcorrfeat* $P = \{p_g\}^G_{g=1}, p_g = \{f^k_g\}^K_{k=1}$ where $K = \omega \prod_{i=2}^{C-1} i$ is the dimension of *pcorrfeat*.

### MANUALFEAT

For the manually selected features, we select the features both in the time and frequency domain. These are widely used features for physiological signals for the baseline comparison in dataset and review papers [25, 46, 47] for affective computing. For the features in the time domain, we choose the mean, standard variance, average root mean square, mean of the absolute values, maximum amplitude and average amplitude for the original and first-order differential of all physiological signals. For the features in the frequency domain, we choose the mean, maximum and the magnitude for the Fast Fourier Transform (FFT) [55] of all physiological signals.

### 4.3.3. MULTIPLE INSTANCE LEARNING LAYERS

The purpose of the multiple instance learning layers is to model the probability between an instance and the corresponding post-stimuli V-A labels [28]. According to the *peak-end theory* [21], the post-stimuli labels usually represent the most salient (peak) or recent (end) V-A within the entire video watching rather than the fine-grained V-A changes. Thus, only a part of the instances inside one bag represent the post-stimuli emotion labels. Traditional MIL algorithms have to make a hypothesis that instances corresponding to the bag label are densely [32] or sparsely [35] consisted of the bag. Unlike traditional MIL algorithms, we designed two multiple instance learning layers to automatically learn the instance gains without a pre-set hypothesis. The multiple instance learning layers assign each instance a matching score (instance gain) which maximize the probability to predict the post-stimuli V-A using the whole bag. That means instances which can better enable the network to predict the post-stimuli V-A will be assigned higher instance gains.

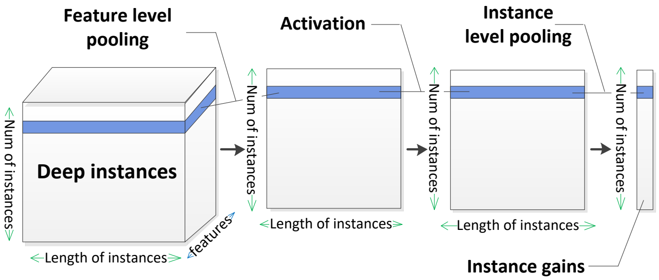


Figure 4.2: The diagram for multiple instance layers

The diagram for multiple instance learning layers is shown in Fig 4.2. For each bag $d_g \in R^{L\times I\times K}$, a maximum pooling is implemented at the feature level (at dimension of features $K$) to select the biggest features for each time point $i$. After that, we activate the features $f_g^k$ for instance $k$ as:

$$a_{k,g} = \Psi(\alpha_{k,g} f_g^k + \beta_{k,g}) \tag{4.2}$$

$\alpha_{k,g}$ and $\beta_{k,g}$ are the weight and bias for the activation operation respectively. $\Psi(\cdot)$ represents the activation function. Here, we use a *softmax* function according to previous works [28, 56]:

$$\Psi(i) = \frac{e^i}{\Sigma_j e^j} \tag{4.3}$$

The purpose of the activation operation is to (a) normalize the selected features in the range from 0 to 1 and (b) make it easier for the network to calculate the gradient during back-propagation. At last, another max pooling operation is implemented at dimension of instance length $I$. After that, we obtain the instance gains $Z = \{z_g\}_{g=1}^G, z_g \in R^{1\times L}$ with the same dimension of the number of instances $L$.

### 4.3.4. Fully-connected layer

To build the link between the instance gains and post-stimuli emotion labels, we put one fully connected layer at the end of *EDMIL*. For the multi-class (high/neutral/low V-A) classification task, we use the *softmax* in equation 4.3 as the activation function. Then we train the network using *RMSprop* [57] optimizer because *RMSprop* can automatically adjust the learning rate for faster convergence. Since the task is multi-class classification, we use the categorical cross entropy ($H_c$) as the loss function for training:

$$H_c = -\frac{1}{n}\Sigma_i[y_i \cdot \ln x_i + (1-y_i)\cdot \ln(x_i)] \tag{4.4}$$

where $x$ and $y$ are the predicted and true value for the fully-connected layer, respectively.

The target of the network is to learn the data representation to predict the post-stimuli emotion labels using the signals for the whole video watching. Thus, the training for post-stimuli labels is fully-supervised. However, the information for which instances can represent the emotion users labeled post-stimuli is not available during the training stage. The instance gain is only the probability of whether the instance makes contributions for the bag to predict the post-stimuli labels. Thus, for each instance, the training is weakly-supervised.

### 4.3.5. Instance regularization

In the prediction stage, when a new user watches a video, their physiological signals are forwarded from pre-processing (section 4.3.1) to multiple instance learning layers (section 4.3.3) to get the instance gains. After that, the instance regularization is implemented to identify which fine-grained instances are correlated with the post-stimuli V-A. Since the instance gain only represents the matching score, we need to obtain the post-stimuli label the user annotated to know which V-A value these instances match. Thus, the post-stimuli V-A is also needed in this step, which means the user needs to input his or her V-A after watching the entire video. We predict the fine-grained V-A according to both the instance gain and the annotated V-A after watching the video:

$$y_i = \begin{cases} Y, & z_i > mean(Z) \\ p, & z_i <= mean(Z) \end{cases} \tag{4.5}$$

where $Z = \{z_i\}$ is the instance gains. mean(z) is the mean value of all instance gains in one bag (signals for the user watch the entire video). $y_i$ is the predicted V-A for instance $i$. $Y$ is the post-stimuli V-A annotated by the user. $p$ is the baseline V-A (i.e., neutral).

## 4.4. Experiments and results

In this section, we first introduce the implementation details of *EDMIL* on CASE, MERCA and CEAP-360VR datasets. We then evaluate the classification and regression performance of *EDMIL* using Leave-One-Subject-Out Cross Validation (LOSOCV). After that, we compare the performance of *EDMIL* with the state-of-the-art MIL algorithms which have been applied for emotion recognition. Then, an ablation study was conducted to verify the effectiveness of each component. Lastly, we compare the performance of the three feature extraction methods mentioned in section 4.3.2.

### 4.4.1. IMPLEMENTATION DETAILS

For all three datasets, we choose four physiological signals, Electrodermal activity (EDA), Blood Volume Pulse (BVP), Skin Temperature (TEMP) and Heart Rate (HR), as the input signals of *EDMIL*. Although EEG signals can provide more abundant information according to previous works [25, 58], high-resolution EEG signals need to be captured under strict laboratory environments without any electromagnetic interference [59], which makes their use limited to an indoor laboratory environment. We choose these four signals because they can be easily measured by wearable and unobtrusive sensing devices such as smart watches or wristband (e.g., Empatica E4 and Microsoft MS Band [1]). In addition, the selected signals contain physiological responses from both autonomic nervous system (EDA and TEMP) and cardiovascular system (BVP and HR), which can provide abundant information for emotion recognition [25, 58]. Moreover, these four signals have also been widely used by previous work to recognize valence and arousal [60–62]. We first pre-process the physiological signals using the standard filtering procedure widely used in previous works [25, 63–65]. Firstly, a low pass filter with a 2Hz cutoff frequency is used to remove noise [63] from EDA signals. For the BVP signal, we implement a 4-order butterworth bandpass filter with cutoff frequencies [30, 200] Hz to eliminate the bursts [64]. At last, an elliptic band-pass filter with cutoff frequencies [0.005, 0.1] is used to filter TEMP signals [65].

To decrease measurement bias in different sessions, all signals are normalized to [0,1] using Min-Max scaling normalization:

$$S_n = \frac{S - \min(S)}{\max(S) - \min(S)} \tag{4.6}$$

The normalization is implemented on each subject under each video stimulus (session). Since signals in both MERCA and CEAP-360VR have different sampling rates, they are interpolated to 50Hz using linear interpretation [66]. We choose linear interpolation because it is the simplest interpolation method which will not change the distribution of the signals. For CASE dataset, we downsample the signals also to 50Hz by decimation down-sampling [67]. The HR for CASE is extracted from ECG using *heartpy* library [68]. Then the input signals are segmented into 2 second instances (sample size 100). The choices for different segmentation lengths are discussed in section 4.5.2. Since different sessions have different lengths, we use zero padding according to previous works [28, 69] to let all sessions (i.e., bags) have the same length. Since CASE does not collect post-stimuli V-A, we use the mean of continuous V-A as ground truth to train *EDMIL* because the mean V-A has no significant difference between post-stimuli V-A [10, 26]. Aside from the comparison of different feature extraction methods (section 4.4.6), we use *deepfeat* described in section 4.3.2 for feature extraction. The network is trained by *RMSprop* [57] optimizer since it can automatically adjust the learning rate for faster convergence. We use the *Early-Stopping* [70] technique to terminate training if there is no improvement on the training loss for 5 epochs.

[1] http://developer.microsoftband.com

The time complexity of *EDMIL* is:

$$O(\frac{15}{64}K^2 \cdot I^2 \cdot L^2 + \frac{1}{32}K \cdot C \cdot I^2 \cdot L + (\frac{69K^2}{32} + \frac{K \cdot C}{16} + 2) \cdot I \cdot L + I) \tag{4.7}$$

$L$ is the number of instances in one bag. $I$ is the number of sample points for one instance. $N = L \times I$ is the sample size of an input signal. $K$ and $C$ are constants which represent the output dimension of the feature vectors and the number of signal channels respectively. Thus, the time complexity can be simplified as:

$$O(A \cdot N^2 + B \cdot N + D) \tag{4.8}$$

where $A$, $B$ are coefficients for $N^2$ and $N$ respectively. $D$ is the constant term of the time complexity. The average training time of *EDMIL* is 218.56s, 156.39s and 192.23s for CASE, MERCA and CEAP-360VR, respectively. *EDMIL* is implemented using Keras (python). All our experiments are run on a server with NVIDIA RTX 2080Ti GPU and 32 GB RAM. The average testing time for each fine-grained instance is 19.21ms, 18.6ms and 15.34ms for CASE, MERCA and CEAP-360VR, respectively. That means to recognize 2s emotions, the algorithm only spends less than 20ms after the network is trained. The computational cost of *EDMIL* is low due to (a) the simple (5-layer) structure for the feature extraction (b) a simple threshold instead of complex constraint functions for the instance regularization module.

### 4.4.2. Evaluation protocol

To evaluate the performance of *EDMIL*, we conduct two kinds of experiments: classification and regression. The classification task tests the instance-level accuracy while the regression task validates the overall dynamics throughout one entire video watching. Below, we introduce the details of the two tasks as well as the metrics and method we use to validate them.

#### Classification task

The aim of the classification task is to test whether *EDMIL* can recognize high/neutral/low V-A for each instance, which is a standard validation method in prior works [19, 25, 71]. We use the mean V-A of instances as ground truth labels for validation. The mapping from continuous values of V-A to discretized categories is: [1,3) = Low, [3, 6) = Neutral, [6, 9] = High.

To test the performance of the classification task of *EDMIL*, we select three validation metrics:

- ***accuracy (acc):*** the percentage of correct predictions;
- ***confusion matrix:*** the square matrix that shows the type of error in a supervised paradigm [15];
- ***weighted F1-score (w-f1):*** the harmonic mean of precision and recall for each label (weighted average by the number of true instances for each label) [72].

These three metrics are widely used in evaluating machine learning algorithms [73]. We use weighted F1-score instead of macro and binary F1-score to take into account label imbalance.

### REGRESSION TASK

The performance of the classification task can only reflect the pairwise comparison between the predicted and ground truth labels for instances, not the overall difference between sequences (i.e., the predicted and ground truth V-A throughout one entire video watching). The purpose of the regression task is to test whether the instance gains (before instance regularization) learned from post-stimuli V-A have similar temporal dynamics with fine-grained V-A ground truth. In addition, as a discretization step, the instance regularization could bring bias to the classification performance. The predicted and ground truth V-A may be different but be discretized into the same category. Thus, the test of regression task can provide an additional validation of *EDMIL*.

To compare the performance of regression, we also train *EDMIL* with 3-class (low/neutral/high) post-stimuli V-A labels to get the instance gains. We then skip the instance regularization and compare the obtained instance gains and fine-grained V-A labels. Since the instance gains and V-A have different magnitudes, we also normalize them using min-max scaling normalization. To evaluate their difference, we use the mean square error (mse) as the validation metric for the regression task:

$$mse = \frac{1}{M}\Sigma_i(y_i - x_i)^2 \tag{4.9}$$

where $y_i$ and $x_i$ are the ground truth and predicted V-A for instances respectively. $M$ is the number of instances in one bag.

### EVALUATION METHOD

We train and test the proposed method using subject-independent models. The subject-independent model is tested using Leave-One-Subject-Out Cross Validation (LOSOCV). LOSOCV is a standard validation method for emotion recognition which can be used to test the generalizability among different users [18]. Data from each subject are separated as testing data and the remaining data from other subjects are used for training. We repeat the training and testing operation for $N$ times ($N$ is the number of subjects in one dataset) to make sure the data from all subjects are used for testing. The results we show are the averaged accuracy, w-f1 and mse among all subjects used as testing data.

Table 4.2: LOSOCV reuslts for CASE, MERCA and CEAP-360VR

| | | **acc** | **w-f1** | **mse** |
|---|---|---|---|---|
| **CASE** | **valence** | 75.63% | 0.72 | 0.2354 |
| | **arousal** | 79.73% | 0.77 | 0.2281 |
| **MERCA** | **valence** | 70.51% | 0.69 | 0.2673 |
| | **arousal** | 67.62% | 0.65 | 0.2051 |
| **CEAP-360VR** | **valence** | 65.04% | 0.65 | 0.2384 |
| | **arousal** | 67.05% | 0.65 | 0.2529 |

## 4.4.3. RESULTS

The classification and regression performance of *EDMIL* on three datasets are shown in Table 4.2. The accuracies for 3-class classification for all three datasets are above 65%.

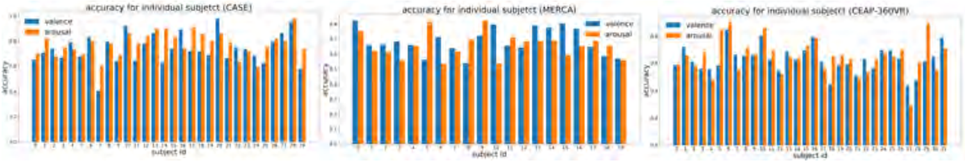
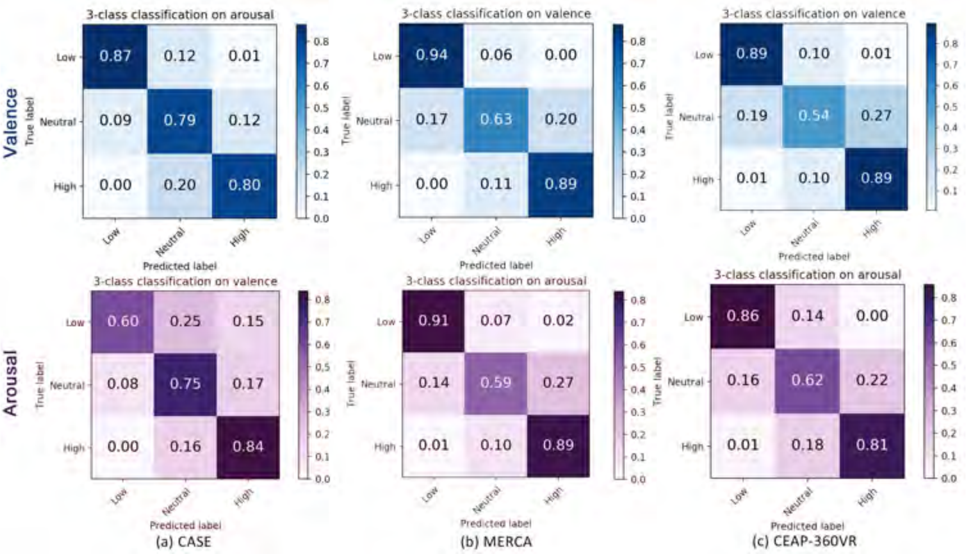


Figure 4.3: The LOSOCV accuracy for individual subject of three datasets



(a) CASE (b) MERCA (c) CEAP-360VR

Figure 4.4: The confusion matrices for leave-one-subject-out cross validation (3-class classification) on (a) CASE, (b) MERCA and (c) CEAP-360VR

The w-f1 scores are also higher or equal to 0.65, which means *EDMIL* can provide balanced recognition precision and recall for different V-A categories. Fig 4.4 shows the confusion matrices for classification. For the comparison between different datasets, *EDMIL* performs the best on CASE dataset (up to 75% accuracy for V-A). The recognition accuracies on CEAP-360VR and MERCA are similar (around 67% for V-A) but lower than the accuracies on CASE. The results indicate that the mobile and VR environments are more challenging for fine-grained emotion recognition compared with a laboratory-based desktop environment. Although the performance on different datasets are different, they all achieve promising accuracies (>65%) and w-f1 scores (>0.65). The test results on different datasets show good generalizability of *EDMIL* among different testing environments (desktop, mobile and VR).

Fig 4.3 shows the accuracy of 3-class classification for each individual user in the three datasets. From the results we can find variability of recognition accuracy among different individuals. The results are coherent with the work of Koelstra et al. [18] and Romeo et al. [15] that there is high inter-subject variability of physiological signals which affects the recognition accuracy. However, the accuracies for more than 75% users (CASE: 93%, MERCA: 85%, CEAP-360VR: 75% of the users respectively) are above 60%. Thus, *EDMIL* achieves balanced performance on different users, which shows good generalizability of *EDMIL* among different subjects.

Figure 4.5: Comparison with baseline methods

| | | **CASE** | | | **MERCA** | | | **CEAP** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **acc** | **w-f1** | **mse** | **acc** | **w-f1** | **mse** | **acc** | **w-f1** | **mse** |
| mi-SVM [74] | valence | 53.55% | 0.60 | 0.3856 | 52.41% | 0.46 | 0.3956 | 49.79% | 0.44 | 0.3845 |
| | arousal | 55.45% | 0.57 | 0.3927 | 54.93% | 0.49 | 0.4012 | 52.90% | 0.47 | 0.3822 |
| MI-SVM [74] | valence | 56.45% | 0.53 | 0.3543 | 52.41% | 0.46 | 0.3726 | 50.21% | 0.45 | 0.3733 |
| | arousal | 59.26% | 0.55 | 0.3862 | 55.07% | 0.48 | 0.3852 | 47.10% | 0.41 | 0.3871 |
| NSK [32] | valence | 56.45% | 0.54 | 0.3774 | 47.14% | 0.46 | 0.3845 | 48.09% | 0.45 | 0.4151 |
| | arousal | 59.66% | 0.55 | 0.3983 | 55.07% | 0.48 | 0.4011 | 49.10% | 0.51 | 0.4169 |
| sb-MIL [35] | valence | 57.47% | 0.53 | 0.2873 | 52.41% | 0.46 | 0.2991 | 50.21% | 0.49 | 0.2927 |
| | arousal | 58.66% | 0.55 | 0.2911 | 47.07% | 0.50 | 0.3012 | 47.00% | 0.47 | 0.2913 |
| AMIL [75] | valence | 66.45% | 0.64 | 0.2451 | 59.59% | 0.51 | 0.2745 | 55.87% | 0.45 | 0.2457 |
| | arousal | 68.57% | 0.62 | 0.2326 | 58.32% | 0.48 | 0.2677 | 57.62% | 0.48 | 0.2678 |
| WSPPG [76] | valence | 70.26% | 0.61 | 0.2382 | 65.34% | 0.53 | 0.2734 | 61.54% | 0.51 | 0.2403 |
| | arousal | 71.32% | 0.62 | 0.2387 | 62.51% | 0.54 | 0.2232 | 63.18% | 0.53 | 0.2612 |
| WCRNN [77] | valence | 61.43% | 0.51 | 0.2874 | 55.11% | 0.43 | 0.3052 | 49.19% | 0.41 | 0.3037 |
| | arousal | 63.37% | 0.53 | 0.2943 | 50.67% | 0.41 | 0.2873 | 50.13% | 0.45 | 0.3315 |
| **EDMIL** | valence | **75.63%** | **0.72** | **0.2354** | **70.51%** | **0.69** | 0.2673 | **65.04%** | **0.65** | **0.2384** |
| | arousal | **79.73%** | **0.77** | **0.2281** | **67.62%** | **0.65** | **0.2051** | **67.05%** | **0.65** | **0.2529** |

### 4.4.4. Comparison with baselines

The comparison of *EDMIL* with baseline methods [32, 35, 74–77] is shown in Table 4.5. We choose four classic multiple instance learning algorithms which have been widely used as baseline methods. In the work of Romeo et al. [15], mi-SVM and MI-SVM [74]

Figure 4.6: Ablation Study for pre-processing (PP), feature extraction (FE) and multiple instance learning (MIL) module

| | | CASE | | | MERCA | | | CEAP-360VR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **acc** | **w-f1** | **mse** | **acc** | **w-f1** | **mse** | **acc** | **w-f1** | **mse** |
| MIL | valence | 56.12% | 0.53 | 0.4052 | 51.71% | 0.49 | 0.3956 | 49.52% | 0.41 | 0.4215 |
| | arousal | 51.57% | 0.49 | 0.4132 | 50.12% | 0.47 | 0.4056 | 53.61% | 0.45 | 0.4372 |
| PP+MIL | valence | 58.21% | 0.51 | 0.4011 | 53.38% | 0.53 | 0.3327 | 51.26% | 0.50 | 0.4112 |
| | arousal | 54.38% | 0.49 | 0.3981 | 52.32% | 0.51 | 0.3152 | 55.21% | 0.51 | 0.4009 |
| FE+MIL | valence | 72.52% | 0.71 | 0.2654 | 59.67% | 0.57 | 0.2738 | 61.26% | 0.59 | 0.2953 |
| | arousal | 75.66% | 0.72 | 0.2477 | 57.21% | 0.56 | 0.2235 | 63.14% | 0.59 | 0.2817 |
| **PP+FE+MIL** | valence | **75.93%** | **0.72** | **0.2354** | **70.51%** | **0.69** | **0.2673** | **65.04%** | **0.65** | **0.2384** |
| | arousal | **79.73%** | **0.77** | **0.2281** | **67.62%** | **0.65** | **0.2051** | **67.05%** | **0.65** | **0.2529** |

achieved the best recognition accuracies (in bag-level) for emotion recognition using physiological signals. We then add another two baseline methods, NSK [32] and sb-MIL [35], to further compare the performance of *EDMIL* with state-of-the-art methods. For these four methods, we use the same hand-crafted features with [15]. In addition to the classic MIL algorithms, we also select three deep learning based weakly-supervised methods for comparison (i.e., Attentional Multiple Instance Learning (AMIL) [75], Weakly Supervised PPG (WSPPG) [76], Weakly supervised Convolutional Recurrent Neural Network (WCRNN) [77]). The baselines we choose include some widely used and more complex machine learning structures (i.e., recurrent structure [77], attention structure [75], deeper CNN [75, 75]). All these three baselines use the end-to-end learning structures which are designed for 1D signal based (voice [75, 77], PPG [76]) learning tasks. Thus, we can directly use them for testing without manually selecting features like classic MIL baselines. Similar to *EDMIL*, we use these seven methods to obtain the instance gains to compare the regression performance. For the classification performance, we use the same instance regularization to transfer the instance gains into fine-grained V-A for all the four baseline methods.

As shown in Table 4.5, the performance of *EDMIL* outperforms all four classic MIL baseline methods. The results are coherent with the finding of Romeo et al. [15] that classic MIL algorithms cannot achieve high recognition accuracy using subject-independent models. The classic MIL algorithms need to make hypotheses that the instances corresponding to the bag label are densely (mi-SVM, MI-SVM and NSK) or sparsely (sb-MIL) composed of the bag. However, for fine-grained emotion recognition, we do not know whether the post-stimuli emotions are the most salient (only small amount of the instances are correlated with the post-stimuli label) or overall (most of the instances are correlated with the post-stimuli label) emotions of users. That makes it challenging for classic MIL methods to identify the instances which are correlated with the post-stimuli labels for fine-grained emotion recognition.

For the deep learning based weakly-supervised methods, we find that all three of them provide better classification (average acc +7.75%) and regression (average mse -0.097)) results compared with the four classic MIL methods. However, we also find out that all three methods result in problems of overfitting for the classification task. The accuracies on training sets are much higher than the accuracies on testing sets: the accuracy differences for training and testing sets are 23.14%, 19.27% and 22.28% for AMIL, WSPPG and WCRNN, respectively. The results demonstrate that deeper network or more complex structures (i.e., attention structure and recurrent structure) can decrease the generalizability of the algorithms by providing more accurate results only on the training set.

*EDMIL* obtains good recognition accuracy and w-f1 score (> 65% accuracy and 0.65 w-f1) for all three datasets. By taking advantage of the end-to-end structure, *EDMIL* automatically obtains the matching scores for instances and bag labels without a pre-set hypothesis. Compared with classic MIL methods, we do not need to know whether most or only a small amount of the instances are correlated with the post-stimuli labels. Compared with the three baselines which use an end-to-end learning structure, *EDMIL* also achieves better performance (average acc +10.34%, mse -0.03). *EDMIL* does not suffer from overfitting: we only find the training accuracy is 3.46% (averaged from three datasets) higher than the testing accuracy for *EDMIL*. That is a result of the shallow feature extraction network and simple instance regularization module we use to design *EDMIL*. The result also shows that more accurate fine-grained emotion recognition can be achieved using deep neural network based (compared with traditional machine learning based) weakly-supervised learning algorithms.

### 4.4.5. ABLATION STUDY

We conduct an ablation study to verify the effectiveness of each component. Since our algorithm needs MIL layers to obtain fine-grained V-A, we begin with only using the MIL layers to train the network. The MIL layers directly use the raw signal segments without passing them through the pre-processing and feature extraction module. Then we test the performance of combining the MIL layers with the pre-processing (PP) and feature extraction (FE) layers respectively. Finally, we combine all the modules in *EDMIL* and present the results for comparison.

As shown in Table 4.6, both FE and PP contribute to the classification and regression tasks. The FE benefits the network by extracting deep features for MIL layers to learn the probability for instances to predict the corresponding post-stimuli labels. Thus, the recognition accuracies increase 12.80% and mse drops 0.148 on average after combining FE to MIL. The increased performance of adding PP is not as significant as adding FE: the recognition accuracies increase 6.07% and mse drops 0.027 on average after combining PP to the network. The reason of this is that the convolution layers of FE have already automatically filtered some of the noise and artifacts in the signals when extracting the features. In conclusion, all components contribute to both the classification and regression tasks. The observations above demonstrate the effectiveness of the components in the proposed algorithm.

### 4.4.6. COMPARISON BETWEEN FEATURE EXTRACTION METHODS

As we introduced in section 4.3.2, we compare three feature extraction methods (*deepfeat, pcorrfeat and manualfeat*) in the feature extraction layer of *EDMIL*. The purpose of this comparison is to find out whether the deep features (*deepfeat*) learned by the end-to-end neural network can provide more accurate classification and regression results compared with unsupervised feature extraction method (*pcorrfeat*) and manually selected features (*manualfeat*).

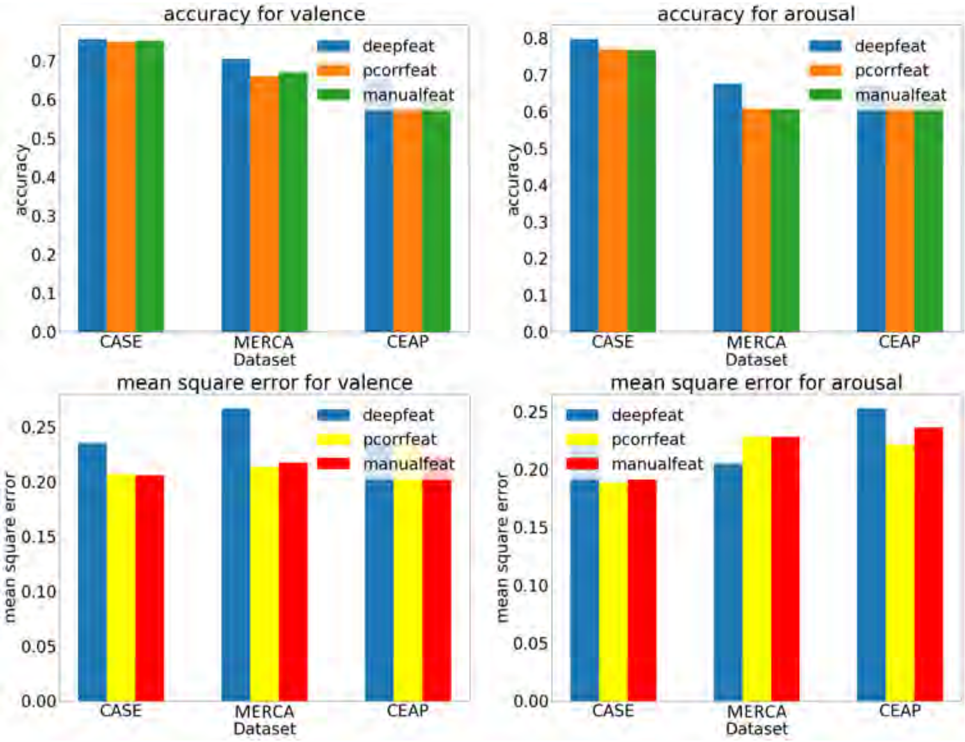


Figure 4.7: The accuracy and mse of deepfeat, pcorrfeat and manualfeat on three datasets

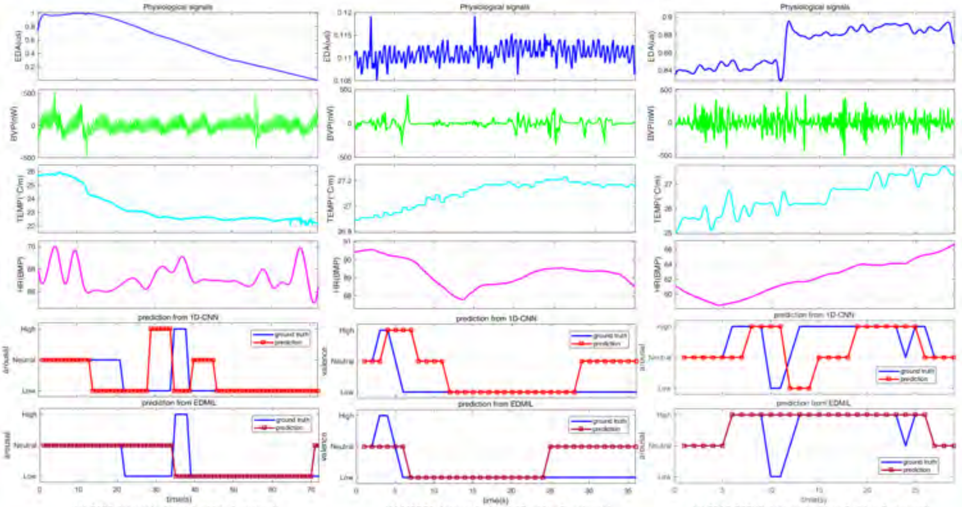
As shown in Fig 4.7, *deepfeat* results in the highest accuracy. *pcorrfeat* and *manualfeat* have similar accuracy which are lower than *deepfeat*. In addition, the accuracy on the training and testing set is similar (training accuracies are 2.6%, 3.7% and 4.1% higher than testing for CASE, MERCA and CEAP-360VR respectively). The results indicate that using an end-to-end feature extraction method does not result in overfitting (found by previous works on fully-supervised learning for emotion recognition [3, 44]) due to the weakly-supervised structure we use.

However, the mse of *pcorrfeat* and *manualfeat* are lower than *deepfeat*. Thus, using unsupervised and manually extracted features can result in better regression results. The reason of higher mse for *deepfeat* is that the deep features are learned for the classification task, not the regression task. The *pcorrfeat* and *manualfeat* however, are not constrained with the classification task, which can better represent the dynamic patterns of V-A changes. However, *EDMIL* is designed for the classification task particularly and the post-stimuli emotion labels for training are also the discretized labels. In addition, for the regression task, the maximum and minimum V-A are needed to normalize the instance gains. It means users are expected to input their highest and lowest V-A during the whole video watching, which is sometimes difficult for users if the video is long.

Thus, the classification is more applicable (users only need to input their post-stimuli V-A) in real-life scenarios.

Table 4.3: The comparison between weakly-supervised (*EDMIL*) and fully-supervised (*1D-CNN and LSTM*) methods

| | | EDMIL | | 1D-CNN[78] | | LSTM [79] | |
|---|---|---|---|---|---|---|---|
| | | **acc** | **dtw** | **acc** | **dtw** | **acc** | **dtw** |
| **CASE** | **valence** | **75.63%** | 16.0417 | 53.04% | **10.886** | 54.74% | 11.77 |
| | **arousal** | **79.73%** | 12.6875 | 52.34% | **6.997** | 53.82% | 7.267 |
| **MERCA** | **valence** | **70.51%** | 11.7688 | 51.57% | 6.183 | 53.88% | **6.031** |
| | **arousal** | **67.62%** | 10.7917 | 42.91% | 8.437 | 46.26% | **8.215** |
| **CEAP-360VR** | **valence** | **65.04%** | 9.9973 | 47.33% | **5.941** | 44.36% | 6.021 |
| | **arousal** | **67.05%** | 9.3810 | 45.67% | **5.733** | 43.27% | 5.938 |



(a) CASE dataset(subject 4, video 4, arousal)

(b) MERCA dataset (subject 19, video 5, valence)

(c) CEAP-360VR dataset (subject 2 video 0, arousal)

Figure 4.8: Examples of physiological signals and prediction results for fully-supervised (*1D-CNN*) and weakly-supervised (*EDMIL*) algorithm

## 4.5. DISCUSSION

### 4.5.1. FULLY-SUPERVISED V.S. WEAKLY SUPERVISED: ADVANTAGES AND DISADVANTAGES

The baseline methods we test in section 4.4.4 are all weakly supervised methods trained with post-stimuli emotion labels. The fully-supervised learning methods, however, learn the instance-label relationship by building the direct mapping between instances and fine-grained emotion labels. Thus, it is interesting to compare the results between training with post-stimuli labels (weakly-supervised) and fine-grained labels (fully-supervised).

The comparison can help us understand whether the additional fine-grained (i.e., instance-level) labels can improve or compromise the performance of fine-grained emotion recognition.

To implement this comparison, we choose two widely used deep learning models, 1D-Convolutional Neural Network (*1D-CNN*) and Long Short Term Memory networks (LSTM) for comparison. We choose the basic *1D-CNN* [78] and *LSTM* networks [79] from previous works for emotion recognition to avoid over-tuning. We use the mean V-A label for each instance and directly train the *1D-CNN* and *LSTM* at instance-level (i.e., each instance has a corresponding V-A label). We run the 3-class classification task as we did for testing *EDMIL*. To evaluate performance, we use two metrics: the recognition accuracy and dynamic time warping distance (DTW) [80]. DTW is one of the most prominent methods in similarity measurement for time series data [81]. The results for the comparison are shown in Table 4.3.

As shown in Table 4.3, the fully-supervised algorithms achieve lower recognition accuracy compared with our weakly-supervised algorithm (*EDMIL*). The results of fully-supervised methods are supposed to be better than *EDMIL* since the fully-supervised algorithms have additional information (i.e., the instance-level labels) for training. We then compare the training and testing accuracy for both fully-supervised algorithms and *EDMIL*. We find that both the *1D-CNN* and *LSTM* have the problem of overfitting. The accuracies on training sets are much higher than the accuracies on testing sets (average of the three datasets: 20.04% and 18.28% higher for *1D-CNN* and *LSTM* respectively). However, for weakly-supervised training, we do not find a much higher accuracy on the training set (average of the three datasets: 3.46% higher for *EDMIL*). The overfitting can be a result of the temporal resolution mismatch between physiological signals and fine-grained self-reports. When annotating their emotions in real-time, different users have different awareness (interoception) levels about their emotions [45]. The relationships between instances and labels are different among users because the interoception levels across individuals are different. Thus, the recognition algorithm can learn contradictory information if we directly build a strong constraint between the fine-grained labels and signals. That also explains the finding of Romeo et al. [15] and Kandemir et al. [42] that building a subject-independent emotion recognition model is challenging, especially for fine-grained emotion recognition.

Although recognition accuracies are lower, the DTW distances of the fully-supervised methods are lower than *EDMIL*. Lower DTW distance means higher similarity of two time sequences. That indicates that the fully-supervised algorithms can result in better recognition results for the whole video instead of individual instances. Compared with accuracy, DTW is less sensitive to the time-shift of specific values in the sequence. Figure 4.8 shows three examples of the prediction results of *1D-CNN* and *EDMIL* on three datasets. Taking the example of Figure 4.8 (c), although *EDMIL* achieves higher accuracy (86.21% v.s. 55.17%) in this specific case, the DTW of *EDMIL* is higher (6.0 v.s. 1.0) than *1D-CNN*. The results also show that there is temporal mismatch between individuals: when the evaluation metric (i.e., DTW) is less sensitive to the time-shift, fully-supervised methods have better performance. Since we run subject-independent validation, the temporal relationship between input signals and emotions learned from other subjects is different from the one used for testing. That causes shifts of predictions in the time domain, which makes the instance-level accuracies low but does not effect the sequence-level prediction.

We also add the original signals to Figure 4.8. From Figure 4.8 we can see that the arousal labels have a clear correlation between the EDA (Figure 4.8 (a), (c)), which is in line with most of the studies of previous works [60]. The heart rate and skin temperature correlate with both the valence and arousal changes [82, 83]. We also find that some of the changes which are ignored by *EDMIL* but captured by the fully-supervised algorithm are also shown in the physiological signals. For example, for Fig. 9 (a), there is a duration of high arousal predicted by the fully-supervised algorithm, which correlates to an increase in heart rate. However, *EDMIL* predicts the emotion to be neutral during this duration. The visual comparison between signals and predictions validate our conclusion that fully-supervised algorithms can result in better recognition results for whole videos instead of individual instances.

### 4.5.2. TOWARDS TEMPORALLY PRECISE EMOTION RECOGNITION: HOW FINE-GRAINED CAN THE RECOGNITION BE?

The performance of *EDMIL* is influenced by the structure of the bag: the length of the instance can affect the accuracy and the temporal resolution of recognition [3]. The shorter instance length representing a higher number of instances for one video watching will lead to a finer level of granularity in temporal resolution. However, a too-short instance length can bring challenges for feature extraction because the information inside each instance can be insufficient for accurate recognition [3, 15]. Thus, it is worthwhile to find out what are the appropriate instance lengths for fine-grained emotion recognition based on deep multiple instance learning. Thus, we conduct an experiment to test *EDMIL* with different instance lengths on CASE, MERCA, and CEAP-360VR, respectively. The recognition accuracies of different instance lengths are shown in Figure 4.9.

As shown in Figure 4.9, the recognition accuracy decreases significantly if the instance length is $\geq 5s$. This result is coherent with the finding from Romeo et al. [15] that a low number of longer instances may lose salient information related to the local physiological response. The accuracy also decreases when the instance length is $< 1s$. Since emotion states are classified based on the features from each instance, a short instance length can entail insufficient information for accurate classification [3]. The instance length of 2s achieves the best recognition accuracy for both valence and arousal. For all the datasets, instance length from $1s$ to $2s$ can result in good recognition results (up to 60%). The result is in line with the research from Paul et al. [40] that the duration of emotion typically ranges from 0.5s to 4s. The takeaway message from this experiment is that an instance length between $1s$ to $2s$ is the appropriate length for fine-grained emotion recognition using physiological signals and deep multiple instance learning.

### 4.5.3. DOES THE PERCENTAGE OF POST-STIMULI ANNOTATIONS AFFECT RECOGNITION ACCURACY?

The performance of *EDMIL* can also be influenced by the percentage of post-stimuli V-A users annotated in each session. Traditional MIL methods require pre-set hypotheses about whether the instances corresponding to the post-stimuli annotation densely or sparsely consisted of the bag [32, 35]. For example, if a user annotates that he or she ex-

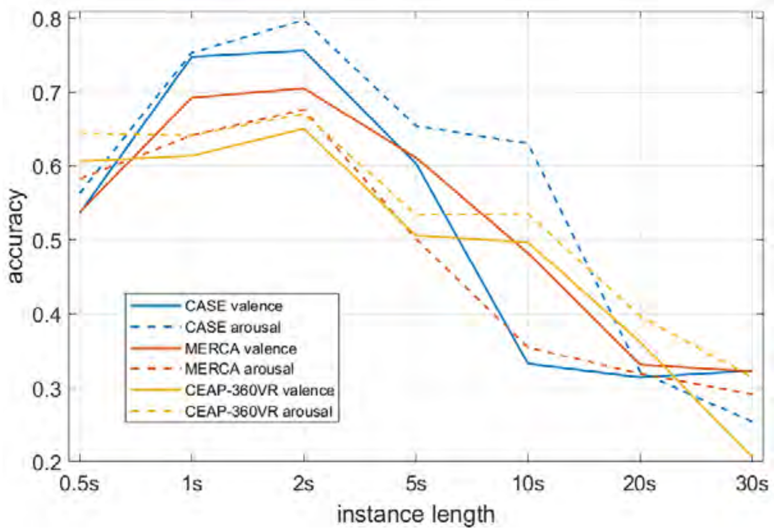


Figure 4.9: The recognition accuracy by different instance lengths on CASE, MERCA and CEAP-360VR

perienced happiness after watching one video, traditional MIL methods can only obtain accurate predictions if the user felt happy most of the time when watching the video. In addition, if all (or most) of the instances are annotated the same as the post-stimuli annotation, we can just do bag-level prediction and train all the instances using fully-supervised learning. In that case, it is not needed to develop weakly-supervised learning algorithms for identifying the instances which contribute to predicting the post-stimuli annotation.

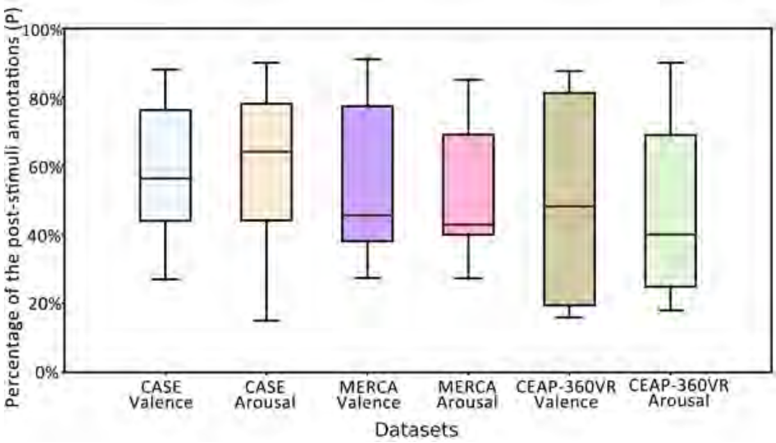


Figure 4.10: The percentage of the post-stimuli annotations in each session (one user watches one video) for CASE, MERCA and CEAP-360VR

To circumvent this, we first check the percentage of instances which are annotated the same as the post-stimuli annotations for the three datasets we use. For each dataset, suppose there are $S$ users who watch $L$ videos. For each session, the user annotates one post-stimuli V-A. Meanwhile, the user also annotates the fine-grained V-A for $K$ instances inside this session. The number of instances which the user annotated the same as the post-stimuli annotations are $N$. The percentage of the post-stimuli annotation for this session is $p = N/K$. Then for the $S \times L$ sessions we obtain $P = [p_1, p_2, \ldots, p_{S\times L}]$ of V-A for the whole dataset. As shown in Figure 4.10, the mean and standard deviation of $P$ for three datasets are: CASE-valence: 58.0%(0.18), MERCA-valence: 53.5%(0.20), CEAP-360VR-valence: 49.5%(0.27), CASE-arousal: 57.4%(0.18), MERCA-arousal: 50.2%(0.18), CEAP-360VR-arousal: 46.5%(0.25). A Shapiro-Wilk tests shows that the percentages of the post-stimuli annotations for all three datasets are not normally distributed (all $p < 0.05$ for three datasets). As we compare three unmatched groups, we perform a Kruskal-Wallis [84] test. We find significant differences of the percentages of the post-stimuli valence ($\chi^2(2) = 10.97, p < 0.05$) and arousal ($\chi^2(2) = 35.29, p < 0.05$) among the three datasets. We then run post-hoc pairwise comparison tests using Mann-Whitney test [85] with Bonferroni correction. The p-values and effect sizes for the pairwise comparison are shown in Table 4.4. In the tests, we find pairwise significant differences (all $p < 0.05$) of $P$ for both valence and arousal between datasets. The results demonstrate that the datasets contain sessions with different levels of time ambiguity (the percentage of the post-stimuli annotations for different datasets are significantly different), which makes it challenging for recognition algorithms to have a generalizable performance on all three datasets.

Table 4.4: The post-hoc pairwise comparison tests using Mann-Whitney test on the percentages of the post-stimuli valence and arousal

(a) valence

| effect size<br>p-value | CASE | MERCA | CEAP-360VR |
|---|---|---|---|
| **CASE** | | 0.566 | 0.583 |
| **MERCA** | 0.026 | | 0.553 |
| **CEAP-360VR** | 0.004 | 0.038 | |

(b) arousal

| effect size<br>p-value | CASE | MERCA | CEAP-360VR |
|---|---|---|---|
| **CASE** | | 0.611 | 0.650 |
| **MERCA** | <0.001 | | 0.576 |
| **CEAP-360VR** | <0.001 | 0.003 | |

To find out whether the percentage of post-stimuli annotation influences the recognition accuracy, we calculate the average accuracy of V-A and the percentage of the post-stimuli annotations for all the sessions in the three datasets. As shown in Figure 4.11, *EDMIL* achieves up to 60% of the recognition accuracy if the post-stimuli annotation

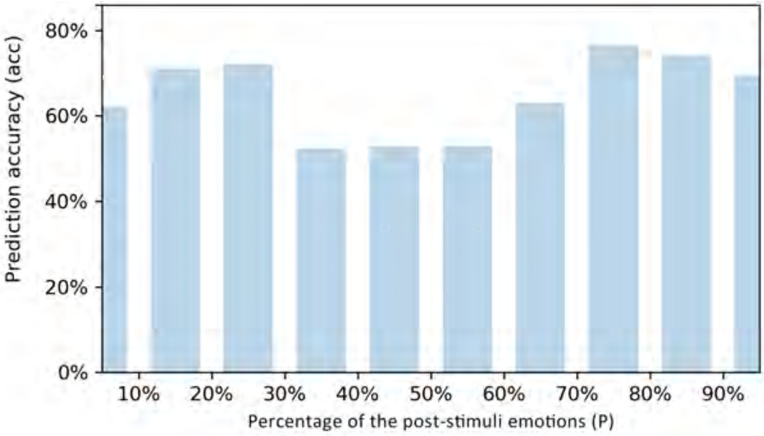


Figure 4.11: The relationship between the percentage of the post-stimuli annotations and recognition accuracy

accounts for more than 60% or less than 30% of one session. If the post-stimuli annotation accounts for 30% to 60% of one session, the accuracy is lower but still more than 55%. Although the result shows up to 50% of accuracy for all the sessions, the performance of *EDMIL* still decreases by around 10% when the post-stimuli annotation is neither densely (more than 60%) nor sparsely (less than 30%) consisted of the bag. The deep structure of *EDMIL* can automatically determine whether the post-stimuli annotation densely or sparsely consisted of the bag. Thus, for these two conditions, *EDMIL* can achieve relatively high accuracy. However, *EDMIL* recognizes the post-stimuli annotation for each instance based on the probability that the instance matches the post-stimuli annotation. Similar to traditional MIL methods, when the post-stimuli annotation neither densely nor sparsely consisted of the bag, the probabilities for instances tend to be similar [86], which makes it difficult for the network to identify the corresponding instances. The takeaway message of this experiment is that the percentage of the post-stimuli annotations do have an influence on the recognition accuracy. *EDMIL* can achieve the highest recognition accuracy if users annotate their most salient but short emotions (less than 30%), or their overall and longer (i.e., persisting) emotions (more than 60%) after watching the video.

## 4.6. CONCLUSION AND LIMITATIONS

Fine-grained emotion recognition requires training the algorithm with fine-grained emotion ground truth labels to build the mapping between segments of signals and corresponding emotions. In this chapter, we propose *EDMIL*, a deep multiple instance learning based emotion recognition algorithm to classify fine-grained valence and arousal trained with only post-stimuli emotion labels. The algorithm uses weakly-supervised learning to model the temporal ambiguity of post-stimuli emotion labels and learn the instance-label relationship according to the probability for each instance to predict the post-stimuli label. The proposed algorithm achieves reasonable performance (more than 65% on high/neutral/low classification) for subject-independent testing on three datasets collected in three different environments (i.e., desktop, mobile, and HMD-based VR). *EDMIL* also outperforms the classic multiple instance learning methods which previous work [15] used for emotion recognition. Running tests on three different datasets, we found that *EDMIL* achieves similar recognition accuracy in desktop, mobile and VR environments, which indicates its good generalizable performance. Finally, our experiments show that (1) weakly supervised learning can reduce overfitting caused by the temporal mismatch between fine-grained annotations and input signals (**RQ 2.1**), (2) instance segment lengths between 1-2s result in the highest recognition accuracies, (3) *EDMIL* can achieve the highest recognition accuracy if users annotate their most salient but short emotions, or their overall and longer-duration (i.e., persisting) emotions (**RQ 2.3**), and (4) feature extraction using an end-to-end structure can improve recognition accuracy compared with manual feature extraction as well as unsupervised feature extraction methods (**RQ 2.2**).

Given the challenges of predicting valence and arousal labels at a fine level of granularity using only post-stimuli labels, there are naturally limitations to our work. Although the weakly-supervised learning algorithm developed in this chapter can obtain accurate recognition results with only few annotations, it still needs to collect large amounts of data for training. In addition, the recognition system trained with post-stimuli emotion labels can only identify the annotated (post-stimuli) emotion from the baseline emotion (e.g., neutral) because only post-stimuli labels are used for training. The non-annotated emotions are all categorized as part of the baseline. To overcome these limitations, in **Chapter 5** we explore few-shot learning algorithms which recognize emotions in a fine level of granularity by using only few annotated samples as training data.

# REFERENCES

[1] S. Khorram, M. McInnis, and E. M. Provost, *Jointly aligning and predicting continuous emotion annotations,* IEEE Transactions on Affective Computing (2019).

[2] M. Abdul-Mageed and L. Ungar, *Emonet: Fine-grained emotion detection with gated recurrent neural networks,* in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (2017) pp. 718–728.

[3] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors,* Sensors **21**, 52 (2021).

[4] F. Hasanzadeh, M. Annabestani, and S. Moghimi, *Continuous emotion recognition during music listening using eeg signals: A fuzzy parallel cascades model,* arXiv preprint:1910.10489 (2019).

[5] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, *Analysis of eeg signals and facial expressions for continuous emotion detection,* IEEE Transactions on Affective Computing **7**, 17 (2015).

[6] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, *Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze,* in *2019 International Conference on Multimodal Interaction* (2019) pp. 40–48.

[7] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, *Emotion recognition using multimodal residual lstm network,* in *Proceedings of the 27th ACM International Conference on Multimedia* (2019) pp. 176–183.

[8] G. Van Houdt, C. Mosquera, and G. Napoles, *A review on the long short-term memory model,* Artificial Intelligence Review **53**, 5929 (2020).

[9] J. J. Rivas, F. Orihuela-Espina, L. Palafox, N. Berthouze, M. del Carmen Lara, J. Hernández-Franco, and E. Sucar, *Unobtrusive inference of affective states in virtual rehabilitation from upper limb motions: A feasibility study,* IEEE transactions on affective computing (2018).

[10] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2020) pp. 1–15.

[11] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* Scientific data **6**, 1 (2019).

[12] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, *Introducing the recola multimodal corpus of remote collaborative and affective interactions,* in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (IEEE, 2013) pp. 1–8.

[13] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, *Avec 2015: The 5th international audio/visual emotion challenge and workshop,* in *Proceedings of the 23rd ACM international conference on Multimedia* (2015) pp. 1335–1336.

[14] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, *Avec 2016: Depression, mood, and emotion recognition workshop and challenge,* in *Proceedings of the 6th international workshop on audio/visual emotion challenge* (2016) pp. 3–10.

[15] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, *Multiple instance learning for emotion recognition using physiological signals,* IEEE Transactions on Affective Computing (2019).

[16] M. Song, Z. Yang, A. Baird, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, and B. Schuller, *Audiovisual analysis for recognising frustration during game-play: Introducing the multimodal game frustration database,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 517–523.

[17] I. Abdic, L. Fridman, D. McDuff, E. Marchi, B. Reimer, and B. Schuller, *Driver frustration detection from audio and video in the wild,* Proceedings of the KI , 237 (2016).

[18] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *Deap: A database for emotion analysis; using physiological signals,* IEEE transactions on affective computing **3**, 18 (2011).

[19] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[20] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, *Ascertain: Emotion and personality recognition using commercial sensors,* IEEE Transactions on Affective Computing **9**, 147 (2016).

[21] B. L. Fredrickson and D. Kahneman, *Duration neglect in retrospective evaluations of affective episodes.* Journal of personality and social psychology **65**, 45 (1993).

[22] J. Wu, Z. Zhou, Y. Wang, Y. Li, X. Xu, and Y. Uchida, *Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction,* in *2019 International Conference on Multimodal Interaction* (2019) pp. 582–588.

[23] Z.-H. Zhou, *Ensemble methods: foundations and algorithms* (Chapman and Hall/CRC, 2012).

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

[25] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, *A review of emotion recognition using physiological signals,* Sensors **18**, 2074 (2018).

[26] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, *Rcea-360vr: Real-time, continuous emotion annotation in 360 vr videos for collecting precise viewport-dependent ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2021).

[27] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, *Multiple instance learning: A survey of problem characteristics and applications,* Pattern Recognition **77**, 329 (2018).

[28] J. Feng and Z.-H. Zhou, *Deep miml network.* in *AAAI* (2017) pp. 1884–1890.

[29] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, *Multi-scale blocks based image emotion classification using multiple instance learning,* in *2016 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2016) pp. 634–638.

[30] N. Pappas and A. Popescu-Belis, *Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis,* in *Proceedings of the Conference on Empirical Methods In Natural Language Processing (EMNLP)* (2014) pp. 455–466.

[31] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. S. Narayanan, *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, 2011) pp. 31–41.

[32] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, *Multi-instance kernels,* in *ICML*, Vol. 2 (2002) p. 7.

[33] C. Zhang, J. C. Platt, and P. A. Viola, *Multiple instance boosting for object detection,* in *Advances in neural information processing systems* (2006) pp. 1417–1424.

[34] Q. Zhang and S. A. Goldman, *Em-dd: An improved multiple-instance learning technique,* in *Advances in neural information processing systems* (2002) pp. 1073–1080.

[35] R. C. Bunescu and R. J. Mooney, *Multiple instance learning for sparse positive bags,* in *Proceedings of the 24th international conference on Machine learning* (2007) pp. 105–112.

[36] X. Zhang, Y. Wang, S. Zhao, J. Liu, J. Pan, J. Shen, and T. Ding, *Emotion recognition based on electroencephalogram using a multiple instance learning framework,* in *International Conference on Intelligent Computing* (Springer, 2018) pp. 570–578.

[37] Y. Tian, W. Hao, D. Jin, G. Chen, and A. Zou, *A review of latest multi-instance learning,* in *2020 4th International Conference on Computer Science and Artificial Intelligence* (2020) pp. 41–45.

[38] Y. Wang, J. Ma, B. Hao, P. Hu, X. Wang, J. Mei, and S. Li, *Automatic depression detection via facial expressions using multiple instance learning,* in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2020) pp. 1933–1936.

[39] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, *Exploring principles-of-art features for image emotion recognition,* in *Proceedings of the 22nd ACM international conference on Multimedia* (2014) pp. 47–56.

[40] E. Paul, *Emotions revealed: recognizing faces and feelings to improve communication and emotional life,* NY: OWL Books (2007).

[41] R. W. Levenson, *Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity,* Social psychophysiology: Theory and clinical applications (1988).

[42] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, *Multi-task and multi-view learning of user state,* Neurocomputing **139**, 97 (2014).

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition,* Proceedings of the IEEE **86**, 2278 (1998).

[44] T. Zhang, A. El Ali, C. Wang, X. Zhu, and P. Cesar, *Corrfeat: Correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition,* in *International Conference on Multimodal Interaction* (2019) pp. 404–408.

[45] H. D. Critchley and S. N. Garfinkel, *Interoception and emotion,* Current opinion in psychology **17**, 7 (2017).

[46] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, *Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications,* International journal of human-computer studies **72**, 717 (2014).

[47] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, *Physiological signals based human emotion recognition: a review,* in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications* (IEEE, 2011) pp. 410–415.

[48] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, *Wavenet: A generative model for raw audio,* arXiv preprint:1609.03499 (2016).

[49] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, *Large kernel matters–improve semantic segmentation by global convolutional network,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017) pp. 4353–4361.

[50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, *Inception-v4, inception-resnet and the impact of residual connections on learning,* in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (2017) pp. 4278–4284.

[51] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition,* arXiv preprint:1409.1556 (2014).

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks,* in *Advances in neural information processing systems* (2012) pp. 1097–1105.

[53] S. D. Kreibig, *Autonomic nervous system activity in emotion: A review,* Biological psychology **84**, 394 (2010).

[54] J. R. Loaiza, *Emotions and the problem of variability,* Review of Philosophy and Psychology , 1 (2020).

[55] H. J. Nussbaumer, *The fast fourier transform,* in *Fast Fourier Transform and Convolution Algorithms* (Springer, 1981) pp. 80–111.

[56] M. Ilse, J. M. Tomczak, and M. Welling, *Attention-based deep multiple instance learning,* in *35th International Conference on Machine Learning, ICML 2018* (International Machine Learning Society (IMLS), 2018) pp. 3376–3391.

[57] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, *A sufficient condition for convergences of adam and rmsprop,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2019) pp. 11127–11135.

[58] R. A. Calvo and S. D'Mello, *Affect detection: An interdisciplinary review of models, methods, and their applications,* IEEE Transactions on affective computing **1**, 18 (2010).

[59] A. Craik, Y. He, and J. L. Contreras-Vidal, *Deep learning for electroencephalogram (eeg) classification tasks: a review,* Journal of neural engineering **16**, 031001 (2019).

[60] J. Shukla, M. Barreda-Angeles, J. Oliver, G. Nandi, and D. Puig, *Feature extraction and selection for emotion recognition from electrodermal activity,* IEEE Transactions on Affective Computing (2019).

[61] K. Gouizi, F. Bereksi Reguig, and C. Maaoui, *Emotion recognition from physiological signals,* Journal of medical engineering & technology **35**, 300 (2011).

[62] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, *Recognizing emotions induced by affective sounds through heart rate variability,* IEEE Transactions on Affective Computing **6**, 385 (2015).

[63] J. Fleureau, P. Guillotel, and I. Orlac, *Affective benchmarking of movies based on the physiological responses of a real audience,* in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (IEEE, 2013) pp. 73–78.

[64] Y. Chu, X. Zhao, J. Han, and Y. Su, *Physiological signal-based method for measurement of pain intensity,* Frontiers in neuroscience **11**, 279 (2017).

[65] P. Karthikeyan, M. Murugappan, and S. Yaacob, *Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress,* Journal of Physical Therapy Science **24**, 1341 (2012).

[66] E. Meijering, *A chronology of interpolation: from ancient astronomy to modern signal and image processing,* Proceedings of the IEEE **90**, 319 (2002).

[67] R. W. Daniels, *Approximation methods for electronic filter design: with applications to passive, active, and digital networks* (McGraw-Hill New York, NY, USA:, 1974).

[68] P. Van Gent, H. Farah, N. Nes, and B. van Arem, *Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data,* in *Proceedings of the 6th HUMANIST Conference* (2018) pp. 173–178.

[69] A. T. Zhang and B. O. Le Meur, *How old do you look? inferring your age from your gaze,* in *2018 25th IEEE International Conference on Image Processing (ICIP)* (IEEE, 2018) pp. 2660–2664.

[70] L. Prechelt, *Early stopping-but when?* in *Neural Networks: Tricks of the trade* (Springer, 1998) pp. 55–69.

[71] H. Ferdinando and E. Alasaarela, *Enhancement of emotion recogniton using feature fusion and the neighborhood components analysis.* in *ICPRAM* (2018) pp. 463–469.

[72] N. Chinchor, *Muc-3 evaluation metrics,* in *Proceedings of the 3rd conference on Message understanding* (Association for Computational Linguistics, 1991) pp. 17–24.

[73] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, *Comparison of evaluation metrics in classification applications with imbalanced datasets,* in *2008 Seventh International Conference on Machine Learning and Applications* (IEEE, 2008) pp. 777–782.

[74] S. Andrews, I. Tsochantaridis, and T. Hofmann, *Support vector machines for multiple-instance learning,* in *Advances in neural information processing systems* (2003) pp. 577–584.

[75] S. Mao, P. Ching, C.-C. J. Kuo, and T. Lee, *Advancing multiple instance learning with attention modeling for categorical speech emotion recognition,* arXiv preprint:2008.06667 (2020).

[76] J. Du, S.-Q. Liu, B. Zhang, and P. C. Yuen, *Weakly supervised rppg estimation for respiratory rate estimation,* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021) pp. 2391–2397.

[77] Y. Chen, H. Dinkel, M. Wu, and K. Yu, *Voice activity detection in the wild via weakly supervised sound event detection.* in *INTERSPEECH* (2020) pp. 3665–3669.

[78] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, *Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos),* IEEE Access **7**, 57 (2018).

[79] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, *End-to-end learning for dimensional emotion recognition from physiological signals,* in *2017 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2017) pp. 985–990.

[80] H. Sakoe and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition,* IEEE transactions on acoustics, speech, and signal processing **26**, 43 (1978).

[81] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, *Querying and mining of time series data: experimental comparison of representations and distance measures,* Proceedings of the VLDB Endowment **1**, 1542 (2008).

[82] H. Hägglund, A. Uusitalo, J. E. Peltonen, A. S. Koponen, J. Aho, S. Tiinanen, T. Seppänen, M. Tulppo, and H. O. Tikkanen, *Cardiovascular autonomic nervous system function and aerobic capacity in type 1 diabetes,* Frontiers in physiology **3**, 356 (2012).

[83] J. L. Greaney, W. L. Kenney, and L. M. Alexander, *Sympathetic regulation during thermal stress in human aging and disease,* Autonomic Neuroscience **196**, 81 (2016).

[84] P. E. McKight and J. Najab, *Kruskal-wallis test,* The corsini encyclopedia of psychology , 1 (2010).

[85] P. E. McKnight and J. Najab, *Mann-whitney u test,* The Corsini encyclopedia of psychology , 1 (2010).

[86] W. Zhang, L. Liu, and J. Li, *Robust multi-instance learning with stable instances,* in *24th European Conference on Artificial Intelligence* (2020) pp. 718–725.

# 5

# Predicting emotions from few samples: a few-shot learning approach

*This chapter explores fine-grained emotion recognition algorithms trained with only few annotated samples. For this purpose, we propose an Emotion recognition algorithm based on Deep Siamese Networks (EmoDSN). EmoDSN recognizes fine-grained valence and arousal (V-A) labels by maximizing the distance metric between signal segments with different V-A labels. The results from our experiments show that EmoDSN achieves promising results for both one-dimension binary (high/low V-A, 1D-2C) and two-dimensional 5-class (four quadrants of V- A space + neutral, 2D-5C) classification. We achieve an average accuracy of 76.04%, 76.62% and 57.62% for 1D-2C valence, 1D-2C arousal, and 2D-5C, respectively, by using only 5 shots (5 samples in each emotion category) of training data.*

*This chapter is published as Zhang T, El Ali A, Hanjalic A, Cesar P. Few-shot Learning for Fine-grained Emotion Recognition using Physiological Signals. IEEE Transaction on Multimedia, April 2022*

## 5.1. Introduction

A growing number of emotion recognition algorithms were developed in recent years [1–3] to model the temporal dynamics of emotion states of users. Previous works[1, 3–5] on fine-grained emotion recognition rely on large amounts of training data with fine-grained emotion labels. These labels are required to be collected in a fine level of granularity (normally the same or similar frequency as the input signal) to train the recognition algorithms [6]. To collect such fine-grained emotion labels, researchers either ask users themselves to label their emotions in real-time while watching videos [7, 8] or invite external annotators to label users' emotions segment-by-segment (e.g., using videos of users' facial expressions [9]) after watching the videos [9, 10]. However, it is challenging to collect large amounts of annotated signals using any of the methods. Asking users to momentarily self-report their emotions can incur more mental workload and result in user fatigue for longer durations (e.g., a two-hour film). For external annotators, at least three annotators are usually required to get a meaningful agreement between them (e.g., high Kappa score) [9, 11]. This requires extra labeling effort and is costly when annotating large amounts of signals. Thus, the experiments to collect large amounts of continuously annotated signals are time-consuming (require additional annotation time from users) and costly (hiring professional annotators is expensive).

The challenge of collecting large amounts of annotated signals has motivated researchers to explore Few-Shot Learning (FSL) algorithms [12] for emotion recognition. FSL algorithms are designed to converge on a small amount of training data and provide relatively accurate prediction results. However, current FSL algorithms are geared towards discrete emotion recognition [13] using static data modalities such as images [14]. Thus, it is challenging to directly apply the existing FSL algorithms for fine-grained emotion recognition using physiological signals. First of all, there can be temporal mismatch between physiological signals and the fine-grained self-reports (i.e., the delay of annotation). Previous works [15–17] found that there are time delays between an emotional event and its annotation. The time of the delay ranges from 2s to 4s according to the experiments of Huang et al. [15]. Secondly, some fine-grained samples in the training set can be labeled incorrectly [1]. The mislabeled samples can be the result of a distraction of users when self-reporting their emotions momentarily or from a temporary failure of the system when collecting the labels. Both the reaction delay and mislabeled training samples can result in a mismatch between training samples and the corresponding ground truth labels. Since we only use few annotated samples for training, the mismatch can cause mis-convergence and overfitting for the recognition model. Previous works[15, 18] show that both the delay of annotation and mislabeled training samples can lower the accuracy if we directly build the recognition model between input signals and fine-grained emotion labels.

To overcome these challenges, this chapter proposes a few-shot learning algorithm (*EmoDSN*) for fine-grained emotion recognition on small data using physiological signals. *EmoDSN* is designed based on Deep Siamese Network (DSN), which can rapidly converge on a small amount of training data (typically < 10 samples per class (i.e., 10 shot) [19]). It can provide recognition results at fine level of granularity (every 2s) by maximizing the distance metric between signal segments with different emotion labels. To overcome the temporal mismatch between signals and emotion labels, we design an embedding network to automatically compensate for the delay of fine-grained emotion labels. To avoid overfitting caused by mislabeled samples, we also develop the distance fusion module which can merge the distance metrics learned from different training samples. This work makes the following contributions for multimedia community:

- We propose an end-to-end few-shot learning algorithm which can predict V-A in fine-level of granularity (2s) using physiological signals trained by a small amount (< 10 shot) of data. The algorithm can help researchers to understand the personalized experience of users watching videos by collecting only a small amount of data for training.
- We test our algorithm on three datasets (CASE [8], MERCA [7] and CEAP-360VR [17]) collected in three environments (desktop, mobile, and HMD-based Virtual Reality (VR)). Recognition results show good performance for both personalized binary (1D-2C) and 5-class (2D-5C) classification on all three datasets. we get an averaged accuracy of 76.04%, 76.62% and 57.62% for 1D-2C valence, 1D-2C arousal and 2D-5C respectively by using 5 shot of training data. Our algorithm enables finding an optimal trade-off between recognition accuracy and collecting small amounts of continuously annotated physiological signals.
- We test state-of-the-art FSL algorithms [20–22] and compare their performance with *EmoDSN*. Results show that the recognition accuracy of *EmoDSN* outperforms other FSL algorithms. Our ablation study also shows that the embedding network (+11.86%) and distance fusion module (+22.32%) we design can significantly improve the accuracy.
- We run experiments to identify training samples from which temporal moments of video watching (e.g., begin, end and changing points [23]) can better represent the distribution of emotion labels and result in better recognition results. We find that the changing points of emotion annotation and the ending moments of video watching are better temporal moments for training samples (result in higher recognition accuracy) when only few annotated samples are available.

## 5.2. RELATED WORK

In this section, we first review the previous works on emotion recognition on small data and then narrow our scope to few-shot learning based emotion recognition.

### 5.2.1. EMOTION RECOGNITION ON SMALL DATA

Fine-grained emotion recognition requires algorithms to predict multiple emotion states by relying on signals within one certain time interval. To train such recognition models, previous works [1, 3–5] need large amounts of data which are annotated in fine-level of granularity. Specifically, they usually require more than 90% of the annotated data in the datasets (e.g., CASE [8], RECOLA [9], K-EmoCon[11], MERCA [1]) to train an accurate recognition model. That means users themselves or external annotators have to continuously annotate 3 to 9 hours (e.g., CASE: 9.5h, RECOLA: 3.4 hours, K-EmoCon: 5.3 hours, MERCA: 7.5 hours) to obtain an adequate amount of data for training. That requires large amounts of labeling effort for either external annotators or users themselves. Thus, it is challenging to collect large amounts of continuously annotated data for fine-grained emotion recognition.

To overcome this challenge, previous works have applied two kinds of methods to build recognition models with a small amount of training data. The first kind of method [24–28] builds a generative model such as Generative Adversarial Network (GAN) to generate artificial signals which obey the distribution of specific emotion categories. Then the recognition models are trained with the hybrid of synthetic and real signals. For example, Chen [24] et al. design a GAN model to generate ECG samples with the corresponding emotion labels. Their experiments show that the augmented dataset help to increase the accuracy by 5% compared with using only original data. Previous works on other physiological signals (i.e., Electroencephalography (EEG) [25], Electrooculography (EOG) [26], Blood Volume Pulse (BVP) [27], saccadic eye movement [28]) have also demonstrated that the augmented signals can promote the recognition accuracy by providing more data to train the recognition model. However, to generate generalizable distributions for different emotion categories, the generative model itself also needs large amounts of signals with continuous annotation [29].

The second kind of method designs machine learning methods which can be trained by a small amount of ground truth labels. For example, Romeo et al. [30] implement four weakly-supervised learning algorithms to estimate fine-grained emotion states from post-stimuli emotion labels (i.e., the labels user annotate after each video watching). The methods they develop can identify which fine-grained signal segments (i.e. instances) can represent the post-stimuli valence and arousal. A Similar approach is used by Pei et al. [31] to model the temporal dynamics of emotional states. In their work, a weakly-supervised Bidirectional LSTM [32] is designed to predict fine-grained emotion labels according to the probability for that instance to predict the corresponding coarse labels. Although the weakly-supervised methods can predict fine-grained emotion labels with less amount of annotation, they can only identify the annotated (e.g., post-stimuli) emotion from the baseline emotion (e.g., neutral) and categorize all the remaining moments as part of the baseline. Thus, they can only predict two emotion states (i.e., the annotated emotion and neutral) in fine-level of granularity.

### 5.2.2. Few-shot learning based emotion recognition

Few-shot learning (FSL) is a kind of machine learning method which can learn a task from few (typically $<10$ samples per class [21, 22, 33]) annotated samples. Compared with weakly-supervised learning methods, FSL algorithms build direct mappings between fine-grained emotion labels and input signals, which can provide prediction with multiple emotion categories. FSL has been applied in previous works on emotion recognition using a variety of data modalities such as images [14]. To learn the representation of emotions using few annotated samples, researchers need to design different embedding networks for different data modalities. For example, Zhan et al. [14] design an affective structural embedding framework to predict the emotions of images. Their embedding network can learn an intermediate space which bridges the affective gap between low-level and high-level visual semantics.

For physiological signals, Jiang et al. [13] develop an FSL algorithm to recognize the level of stress using ECG, EDA and respiratory (RESP) signals. Their method, which is based on the Matching Network [20], achieves 80% accuracy trained by only 30% of the signals (i.e., 31.5 mins) in WESAD dataset [34]. Patane et al. [19] propose a siamese network based arousal recognition algorithm using ECG signals. Their algorithm obtains +21.5% accuracy increase compared to state-of-the-art machine learning algorithms trained with a subject-dependent model. Siamese network [35] is a kind of FSL algorithm which learns the difference between samples with different labels. Compared with other FSL algorithms (e.g., the Matching Network [20] used by Jiang et al.), Siamese network uses the pair-by-pair learning structure (learn the difference between two samples in two categories) instead of using the one-to-many learning structure (learn the difference between one sample and samples in other categories). It has been widely used for emotion recognition because of its simple and interpretable structure [36]. For example, Hayale et al. [36] use the Deep Siamese Neural (DSN) network [37] to recognize 6 basic emotions by facial expressions. For uni-dimensional signals, DSN is also used by Feng et al. [38] to predict low/medium/high arousal using speech signals. They obtain 43.4% accuracy trained with a subject-dependent model.

Although the previous works above provide useful insights on FSL or DSN based emotion recognition, they only recognize the overall emotion of an event (e.g., one video watching) instead of the fine-grained emotion responses. Our work aims to extend few-shot learning algorithms for emotion recognition with fine-level of granularity.

## 5.3. DSN BASED EMOTION RECOGNITION

In this section, we propose an Emotion recognition algorithm based on Deep Siamese Network (*EmoDSN*) to discriminate fine-grained physiological signal segments (i.e., samples) with different emotion labels. *EmoDSN* learns the difference between samples instead of building the precise mapping between samples and emotion labels. Thus it can converge with only few annotated samples as training data. In the training stage, $n$ samples are used for training. The influences of different temporal moments of training samples are discussed in section 5.5.2. *EmoDSN* contains four parts: (1) **Pre-processing:** the obtained physiological signals are firstly pre-processed using different filters to remove the noise and artifacts in signals. (2) **Embedding Network**: the pre-processed signals are then fed into an embedding network to learn embeddings representing the difference of samples between emotion labels. (3) **Siamese Learning**: the embeddings are learned based on the siamese structure. The output of siamese learning is a distance metric which can represent the probability that the two input samples belong to the same emotion label. After the network is learned, the embedding for each training sample will also be generated. In the prediction stage, the pairwise distance metrics between testing and training samples are fused by the (4) **Distance Fusion** module to obtain the probability of the testing samples corresponding to different emotion labels. The testing samples are predicted as the emotion label with the highest probability. Below we provide a detailed description of *EmoDSN*.

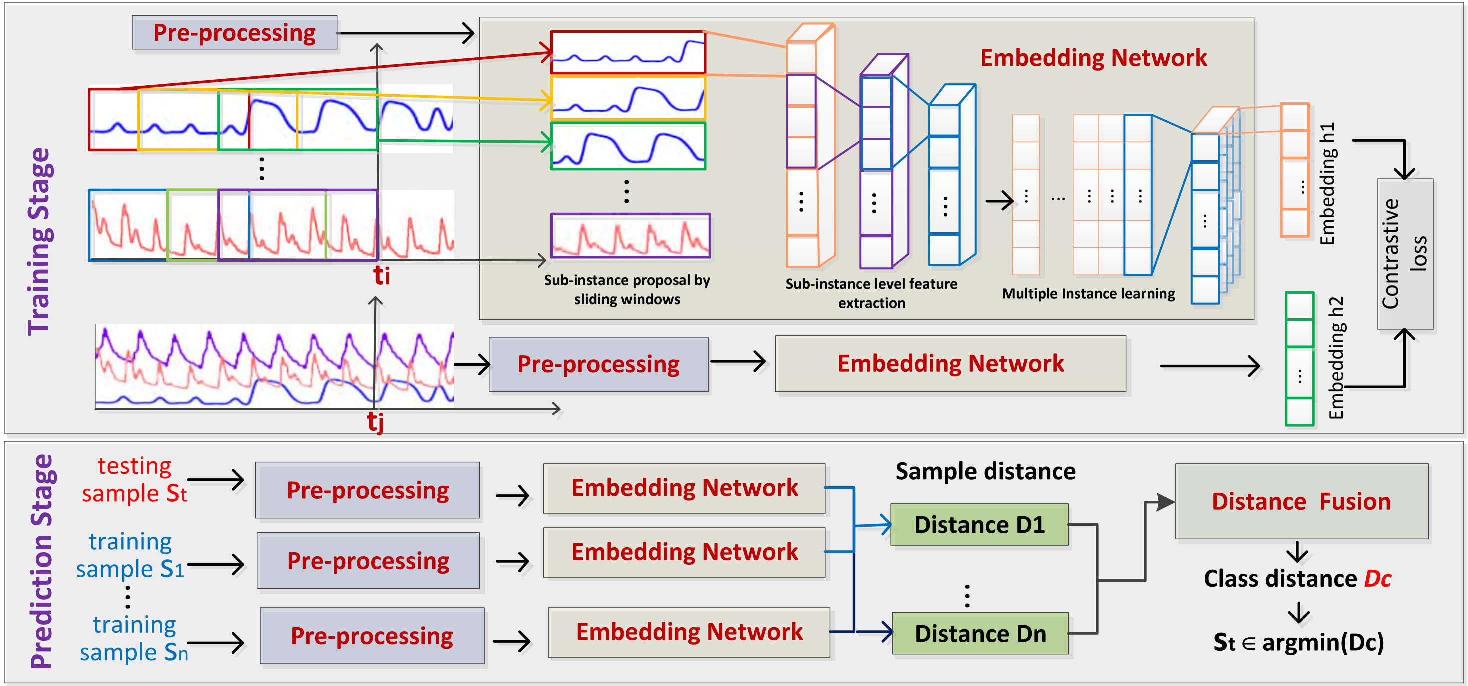


Figure 5.1: The architecture of proposed EmoDSN

### 5.3.1. PRE-PROCESSING

The physiological signals are first pre-processed by different filters to remove the noises and artifacts. We follow the pre-processing procedures which are widely used in previous works [39]. For EDA signals, a low pass filter with a 2Hz cutoff frequency is used to remove noise [40]. For the BVP signals, a 4th-order butterworth bandpass filter with cutoff frequencies [30, 200] Hz is implemented to eliminate the bursts [41]. For TEMP signals, we use an elliptic band-pass filter with cutoff frequencies [0.005, 0.1] [42]. To decrease measurement bias in different sessions (i.e., each subject under each video stimulus), all signals for each session are normalized to [0,1] using Min-Max scaling normalization.

### 5.3.2. EMBEDDING NETWORK

The purpose of the embedding network is to automatically extract features and learn latent vectors of samples to represent the difference of samples between emotion labels. Previous works [1, 43] on embedding networks for physiological signals usually use the fine-grained segments of signals (i.e., samples) as the input. However, samples can be misaligned with the fine-grained emotion labels due to the reaction delay of continuous self-reporting. When continuously annotating emotions towards videos, users first process the stimuli using their senses (at $t_s$) and then react to these changes (at $t_s + t_d$) which leads to a reaction delay ($t_d$). Thus, the sample at $t_s$ actually corresponds with emotion label at $t_s + t_d$. To address the problem of reaction delay, we use sliding windows to propose multiple signal segments (sub-instance) with different delays. Then, the embeddings of these sub-instances are learned using a weakly-supervised multiple instance learning network. Below, we describe the implementation details of each module in the embedding network.

#### SUB-INSTANCE PROPOSAL

Suppose $S = \{s_n\}_{n=1}^{N}, s_n \in R^{L\times C}$ is a set of physiological signals with the number of channels $C$ and the segmentation length $L$. For each sample $s_n$, there is a corresponding emotion label $l_m$. To generate embeddings which consider the delay of reaction, we reconstruct $s_n$ with multiple sub-instances $s'_n = [s_{n1}, s_{n2} \dots s_{nK}]^T$, where $s_{nk}$ is a sub-instance (i.e., each row of sample $s'_n$) with the delay of $t_k$. We use sliding windows with the window length $L$ and stride $k$ to generate the sub-instances. After that, the input of the algorithm become $S' = \{s'_n\}_{n=1}^{N}, s'_n \in R^{K\times L\times C}$, where $K$ is the number of sub-instances for each $s'_n$.

#### SUB-INSTANCE LEVEL FEATURE EXTRACTION

The features are extracted from each sub-instance $s_{nk}$ independently, which means the feature extraction layers will not influence the independence between each sub-instance (no features are extracted from multiple sub-instances). The independent feature extraction guarantees that each sub-instance has a unique instance gain after the embedding network. The instance gains can help us understand the duration of delay with which the network can best discriminate signal segments with different emotion labels.

The features for each sub-instance are extracted using a 3-layer (kernel: $L/2+1-L/4+1-L/8+1$, channels: 4-8-16) 1D-CNN [44]. We use a shallow structure (three layers) instead of deep to avoid overfitting since each sub-instance does not contain much information. We use large (i.e., equal to half of the sub-instance length) convolutional kernels in the shallow layer of the network. Large convolutional kernels have a large receptive field across different sampling points in one sub-instance thus can result in better recognition accuracy [45]. However, the local information can also be omitted by large kernels and result in the difficulty for the network to converge [46]. Thus, we follow a classical strategy that gradually increases the number of kernels and decreases the size of them when the network goes deeper [47]. After sub-instance feature extraction, the $S' = \{s'_n\}_{n=1}^{N}$ is mapped to the feature vectors $F = \{f_n\}_{n=1}^{N}, f_n \in R^{K \times L \times E}$, where $E = 16$ is the dimension of feature vectors.

### MULTIPLE INSTANCE LEARNING

The purpose of multiple instance learning module is to 1) merge the features learned in sub-instances to generate embeddings and 2) assign each sub-instance a instance gain representing the weights of sub-instances for discriminating samples with different emotion labels. The instance gains for all the sub-instances construct the embeddings for the sample. Here we use a weakly-supervised multiple instance learning architecture which is shown in Fig 5.2. Multiple instance learning can map the feature vectors of sub-instances to the probability for that sub-instance to a specific task (in our case, discriminating between emotion labels). Thus, it can promote the interpretability of our algorithm by helping us understand with how much delay (sub-instances with high probability) the signal segment can better predict emotions.

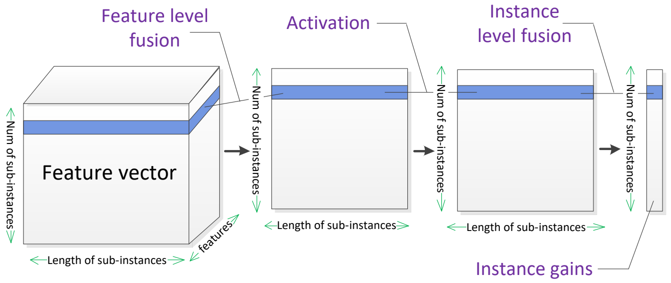


Figure 5.2: The diagram for multiple instance learning module

The feature vectors obtained from the previous module are first input into a feature level fusion module using uni-dimension convolution. The convolution is conducted on the dimension of $E$ to merge the features from different signal channels. After that, the merged features are activated by a Rectified Linear Unit (ReLU) function. Another uni-dimension convolution is implemented on the dimension of $L$ to fuse features for different sampling points inside each sub-instance. At last, we activate the results from previous modules with a softmax function. The purpose of the softmax activation is to (a) normalize the instance gains in the range from 0 to 1 and (b) make the network easier to calculate the gradient for back-propagation. After the multiple instance learning module, the feature vector $f_i = \{f_{nk}\}_{k=1}^{K}, f_{nk} \in R^{L \times E}$ is mapped into instance gain $g_n = \{g_{nk}\}_{k=1}^{K}, g_n \in R^1$. At last, the embedding of one sample $h = [g_1, g_2, \ldots, g_K]$, where $K$ is the number of sub-instances for one sample.

### 5.3.3. Siamese Learning

The purpose of using the siamese learning network is to learn a distance metric which can discriminate samples with different emotion labels. Specifically, for sample $s_i$ and $s_j$, which are two signal segments for $t_i$ and $t_j$, the siamese learning network learns a distance metric $D$ with the target of $D = 0$ if they are with the same emotion label and $D = 1$ if they are with the different emotion labels. To train the network, we first construct two embedding networks with shared weights. The two embeddings $h_i$ and $h_j$ generated from the network are trained by contrastive Loss:

$$L_{contrast} = (1 - \mathrm{Y})\frac{1}{2}(\mathrm{D_w})^2 + \mathrm{Y}\frac{1}{2}\max(0, 1 - \mathrm{D_w})^2 \tag{5.1}$$

where Y equals 0 or 1 for $s_i$ and $s_j$ have the same or different emotion labels respectively. $D_w$ is the euclidean distance for $h_i$ and $h_j$. We also tested the cosine distance metric which is also widely used for other siamese networks. However, our network cannot converge using cosine metric. The contractive loss encourages the network to learn embeddings to place samples with the same labels close to each other while distancing the samples with different emotion labels in the embedding space. The siamese learning network is trained with the *RMSprop* [48] optimizer because it can automatically adjust the learning rate for faster convergence.

### 5.3.4. Distance Fusion

In the prediction stage, when a new sample $s_t$ at time $t$ comes, we can obtain the pairwise distance metric $D = \{D_n\}_{n=1}^{N}$ by calculating euclidean distance between $s_t$ and all training samples $\{s_n\}_{n=1}^{N}$ using their embeddings. The distance metric $D$ can also be used to represent the probability of $s_t \in l_m$ if the emotion label of $s_{nm}$ is available:

$$P(s_t \in l_m | s_{nm} \in l_m) = 1 - D \tag{5.2}$$

where $P(s_t \in l_m | s_{nm} \in l_m)$ represents the probability that $s_t$ corresponds to the emotion label $l_m$ under the condition of $s_{nm} \in l_m$. Previous works [14, 20, 22] on few-shot learning simply average $D$ with the same emotion labels and predict $s_t$ as the emotion label with the closest distance (or greatest possibility). However, the hypothesis of averaging the distances is that the labels for all training samples are correct:

$$P(s_t \in l_m) = \sum_{m=1}^{M} P(s_t \in l_m | s_{nm} \in l_m) \cdot P(s_{nm} \in l_m) \tag{5.3}$$

From equation 5.3 we can conclude that if all $P(s_{nm} \in l_m) = 1$, $1 - P(s_t \in l_m)$ equals to the average of $D$. However, the fine-grained self-reports, which are used as the labels for training, can be mismatched with the physiological signals. Thus, some samples in the training set can be labeled incorrectly. This problem is not that severe when we use large amounts of samples for training. However, when we only use few annotated samples, one or multiple mislabeled samples can significantly lower the model accuracy.

To solve this problem, we propose the **Distance Fusion** module based on Bayesian Fusion to estimate $P(s_{mn} \in l_m)$. Suppose there are $N$ training samples which are annotated as $M$ emotion labels, $N = \{N_m\}_{m=1}^{M}$ are the numbers of training samples with $M$

emotion labels, respectively. $N_m$ is the number of training samples labeled as $l_m$. $s_{mn}$ represents training sample $n$ annotated as emotion label $l_m$. The probability of $s_{mn} \in l_m$ can be estimated by:

$$P(s_{mn} \in l_m) = 1 - \tfrac{1}{2}[\tfrac{1}{N_m} \textstyle\sum_{k=1}^{N_m} D_{mk} - \tfrac{1}{M-1} \textstyle\sum_{i=1}^{M,i \neq m} (\tfrac{1}{N_i} \textstyle\sum_{j=1}^{N_i} D_{ij})] \tag{5.4}$$

where $D_{ij}$ represent the distance between training sample $s_i$ and $s_j$. The first and second $\Sigma$ terms of equation 5.4 represent the probability of $s_{mn}$ similar to the training samples with the same and different emotion labels of $s_{mn}$ respectively. If $s_{mn}$ is similar to the samples with the same label and dissimilar with the samples with different labels, the probability of $s_{mn} \in l_m$ is high.

After we obtain all $P(s_{mn} \in l_m)$ for $N_m$ samples labeled as $l_m$, we can calculate $P(s_t \in l_m)$ by equation 5.3. At last, we predict $s_t$ corresponds to the emotion label with the highest probability:

$$l_t = \arg\max_m (P(s_t \in l_m)) \tag{5.5}$$

where $l_t$ is the predicted emotion label for the $s_t$.

## 5.4. EXPERIMENTS AND RESULTS

### 5.4.1. IMPLEMENTATION DETAILS

To implement a fair evaluation among the three datasets, we process the physiological signals to be as similar as possible before inputting them to *EmoDSN*. Since the three datasets have different sampling rates, we interpolate the signals in MERCA and CEAP-360VR to 50Hz using linear interpretation [49]. We choose linear interpolation because it is the simplest interpolation method which will not change the distribution of the signals. For the CASE dataset, the signals are down-sampled to 50Hz by decimation downsampling [50]. The HRs signals of CASE are extracted from ECG signals using *heartpy* library [51]. We use the mean V-A value of 2-second [1] as the labels for training and testing the algorithm. The window length $L$ and stride $k$ for the sub-instance proposal are 2s and 0.5s respectively according to previous research [1, 30]. For each timestamp $t$, we move the sliding window 12 times to cover the annotation delay for maximum 10s. The amount of time of annotation delay is discussed in section 5.5.1.

We evaluate *EmoDSN* by two tasks: the one-dimensional two-class (1D-2C) classification [52] and the two-dimensional 5-class (2D-5C) classification [53], which are widely used as the tasks for evaluating emotion recognition algorithms using physiological signals. We follow the standard labeling schemes from previous works [52, 53] to map continuous values of V-A to discretized emotion categories. The graphical illustration of this operation is listed in Fig 5.3.

For the training procedure, we train user-specific models for all the users in three datasets. We follow the standard procedure of testing few-shot learning algorithms with continuous signals [33]. We randomly sample $N$ (i.e., shot) sampling points in each emotion category as training samples from one user and use the rest of the samples for testing. The results reported in this section are the average results for all users. We also tried to train user-independent models which use only few annotated samples from one user

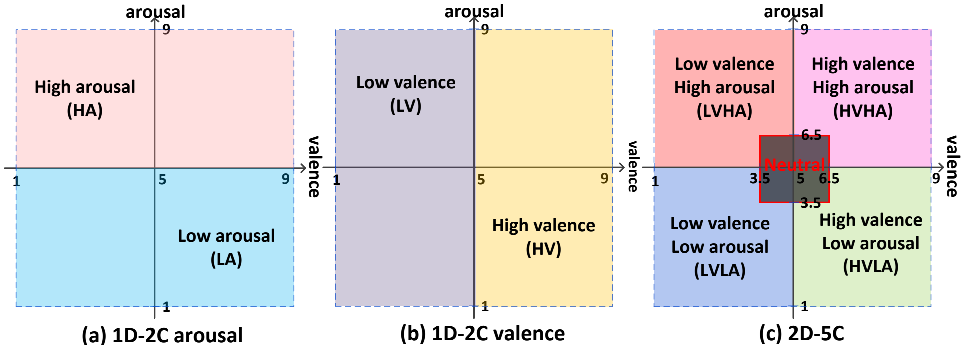


Figure 5.3: Graphical illustration of discretized emotion categories

and test the model on other users. However, due to the high inter-subject variability that affects the physiological signals, building user-independent emotion recognition model is still challenging even using large amounts of annotated data [30, 54]. In this study, we use only few annotated data for training. User-independent models did not achieve satisfactory performance (accuracy not above chance level) for all three datasets and thus the result was not reported in this study.

### 5.4.2. CLASSIFICATION RESULTS

We use accuracy (acc) and macro-F1 score (m-f1) to evaluate the performance of our algorithm. The accuracy represents the percentage of correct predictions. The macro-F1 score is the mean of precision and recall for each label. We use macro-F1 score instead of weighted and binary F1-score to take into account label imbalance. Compared with accuracy, the macro-F1 score can provide more objective evaluation results by taking into account how the data are distributed.

Table 5.1: The performance of *EmoDSN* trained with 5-shot

| **5-shot** | **CASE** | | **MERCA** | | **CEAP-360VR** | |
|---|---|---|---|---|---|---|
| | **acc** | **m-f1** | **acc** | **m-f1** | **acc** | **m-f1** |
| **1D-2C-valence** | 77.30% | 0.722 | **78.95%** | **0.765** | 71.86% | 0.695 |
| **1D-2C-arousal** | **77.72%** | 0.709 | 77.18% | **0.764** | 74.95% | 0.729 |
| **2D-5C** | **58.77%** | 0.482 | 56.08% | 0.508 | 56.93% | **0.512** |

The performance of *EmoDSN* trained with 5-shot is shown in Table 5.1. *EmoDSN* can obtain up to 70% for 1D-2C and 56% for 2D-5C, which are much higher than chance level (shown in Fig. 5.6). The results are obtained by training with only 5-shot (10 seconds of sampling points for each emotion category). That demonstrates that *EmoDSN* can converge and obtain accurate fine-grained emotion recognition with only few annotated samples.

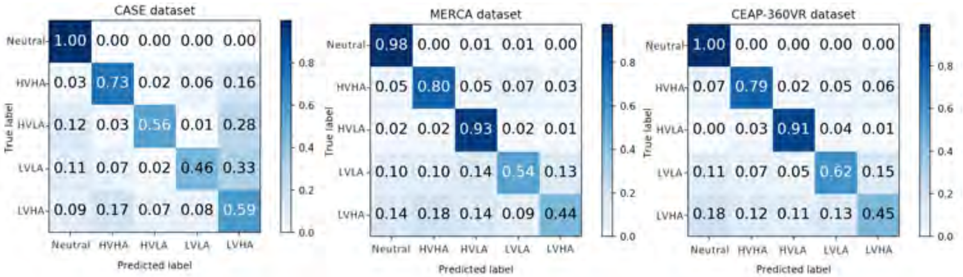


Figure 5.4: The confusion matrices for 2D-5C trained by 5-shot

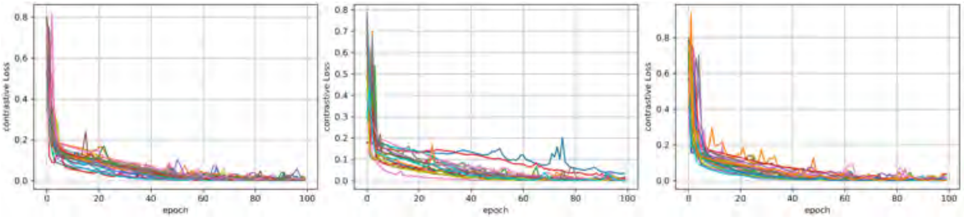


Figure 5.5: The training losses (5-shot, 2D-5C) of EmoDSN on CASE (left), MERCA (middle) and CEAP-360VR (right), each curve represents the training loss for the user-specific model trained on one user

## 5.4.3. RESULTS FOR DIFFERENT EMOTION CATEGORIES

The confusion matrices for 2D-5C are shown in Fig 5.4. We only show the confusion matrices for 2D-5C because it contains classification results for more emotion categories. From the confusion matrices we can see that *EmoDSN* performs well on discriminating the neutral and non-neutral samples. Almost all neutral samples are predicted as neutral for the three datasets. *EmoDSN* also performs well on discriminating samples with high valence. An averaged acc of 78.6% is obtained by *EmoDSN* when discriminating high/low arousal under the condition of high valence. However, the performance on discriminating samples with low valence is not as good as high valence. More than 20% of the LVHA samples are categorized as LVLA on average of the three datasets.

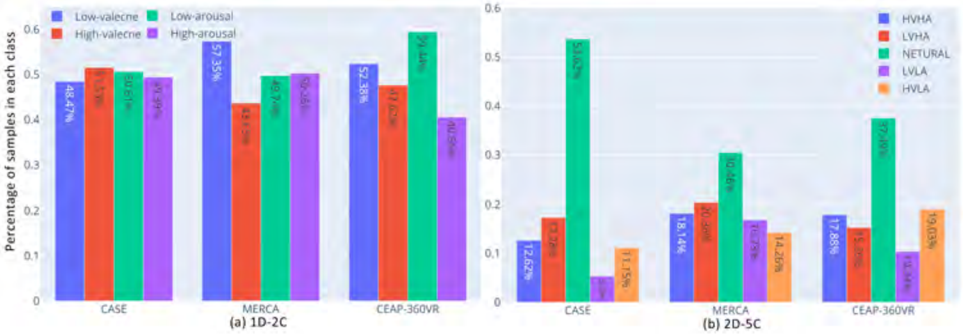


Figure 5.6: Percentage of samples in different emotion classes

The reason is the class imbalance for different emotion categories. As shown in Fig 5.6 (a), for 2D-5C, the neutral class has more than 30% of samples for all three datasets. The most imbalanced dataset is CASE, which contains more than 50% of samples with neutral labels. We also find the LVLA and LVHA classes have comparatively fewer samples (16.48%, 31.04% and 29.37% for CASE, MERCA and CEAP-360VR respectively). That is why discriminating different levels of arousal is more challenging under the condition of low valence. However, for 1D-2C, the high/low V-A classes are balanced. We do not find any classes with less than 40% of all samples. Thus, the performance for 1D-2C is comparatively balanced: the m-F1 score is 3.6% lower than the acc on average of the three datasets. For 2D-5C, the m-F1 is 7.4% lower than the acc on average of the three datasets.

However, even taking the class imbalance into consideration, the acc obtained by *EmoDSN* is still higher than the chance level. For CASE, MERCA and CEAP-360VR respectively, the accuracies are 19.07%, 24.67% and 17.75% higher than the percentage of samples in the class with the most samples (i.e., the chance level). The results show that although *EmoDSN* provides relatively imbalanced precision and recall for different V-A categories, it does not overfit into one specific V-A category and can still provide accurate predictions.

### 5.4.4. RESULTS FOR DIFFERENT DATASETS AND SUBJECTS

For the comparison between different datasets, our method performs best on CASE dataset (up to 76% and 58% acc for 1D-2C and 2D-5C respectively). The acc of 1D-2C on MERCA is similar to CASE but the 2D-5C acc on MERCA is 2.69% lower. Both the accuracies for 1D-2C and 2D-5C on CEAP-360VR are lower than the accuracies on CASE for 3.35% on average. We speculate that the different accuracies of *EmoDSN* on three datasets is a result of the different experimental environments. The data collection experiment of CASE was conducted in an indoor laboratory environment, which contains less interference and noise (e.g., environment noise, user movement, sensor detachment). Thus, the signals from CASE contain less noise and artifacts caused by both the users themselves and the outside environment. The results indicate that the mobile (MERCA) and VR environments (CEAP-360VR) are more challenging for fine-grained emotion recognition compared with a laboratory-based desktop (CASE) environment. However, the maximum difference in acc between the three datasets is less than 7%, which shows that our algorithm does not overfit on one specific dataset. The test results on different datasets show good generalizability of *EmoDSN* among different environments (desktop, mobile and VR).

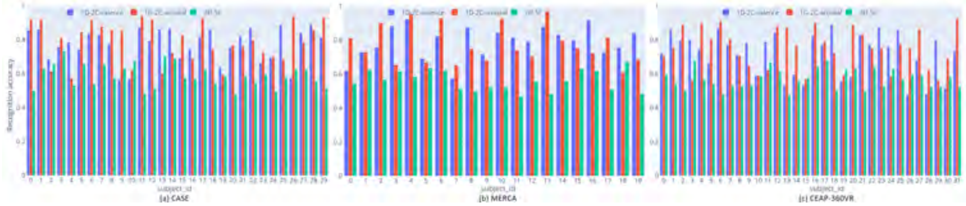
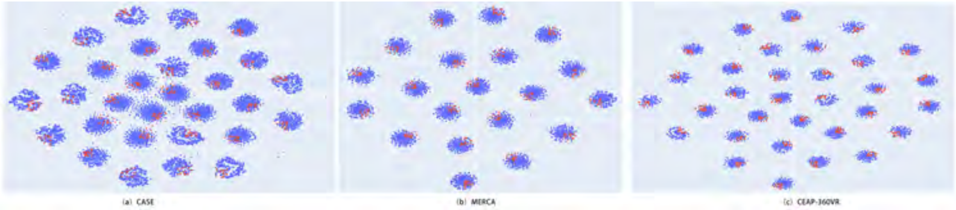


Figure 5.7: The recognition acc for individual subject of CASE, MERCA and CEAP-360VR

For the comparison between different subjects, Fig 5.7 shows the acc for each individual subject of three datasets. From Fig 5.7 we can find variability of acc between different individuals: the average SD for 1D-2C valence, arousal and 2D-5C are 10.43%, 11.31% and 6.11% respectively. Our model achieves up to the chance level (the percentage of samples in the class with the most samples) accuracies for 86.05%, 85.57% and 82.37% of the subjects for 1D-2C valence, 1D-2C arousal and 2D-5C respectively. For the subjects which our algorithm does not achieve above the chance level accuracies, we find the annotations of their data are highly imbalanced (i.e., subject annotates a high percentage of neutral emotion). For example, subject 6 in the CEAP-360VR dataset annotated 72.35% of his or her emotion as neutral when watching videos. The average percentage of neutral annotations for these subjects is 28.41% higher than the subjects whose accuracies are above the chance level. Although recognition accuracies from some of the subjects are low because of class imbalance, our model still achieves above the chance level acc for more than 80% of the subjects. The balanced performance on different subjects shows good generalizability of *EmoDSN* among different subjects.

### 5.4.5. Visualization of the embeddings



(a) CASE (b) MERCA (c) CEAP-360VR

Figure 5.8: The visualization of embedded features using t-SNE. Red and blue points denote the train and test samples, respectively. Best viewed in color.

To visualize the joint sample distribution of the training/testing set, we use T-distributed stochastic neighbor embedding (t-SNE) to reduce the dimension of the embeddings to 2D. It is widely used by previous works [55, 56] of few-shot learning approaches for visualizing the training/testing set. From Fig 7 we can see that the embedding trained by *EmoDSN* constructs the compact clusters of the training features close to the testing features (average purity score = 0.685, 0.657 and 0.614 for CASE, MERCA and CEAP-360VR respectively). The close temporal position between training/testing samples indicates that the learned embeddings can represent the joint distributions between few-shot training samples and the remaining test samples for emotion classification. Previous works [55, 57] show that the closer the training/testing sets are, the easier the classification network can complete the learning task. Our visualization results demonstrate the effectiveness of the embedding network we designed for the Deep Siamese Network.

### 5.4.6. Comparison with baseline methods

#### Implementation details

To compare the performance of *EmoDSN* with state-of-the-art emotion recognition methods, we choose two kinds of baselines: classic FSL networks (i.e., Matching network

Table 5.2: Comparison between few-shot learning methods

| Dataset | Methods | 1-shot accuracy | | | 5-shot accuracy | | | 10-shot accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1D-2C valence | 1D-2C arousal | 2D-5C | 1D-2C valence | 1D-2C arousal | 2D-5C | 1D-2C valence | 1D-2C arousal | 2D-5C |
| **CASE** | MN [20] | .364(.123) | .381(.132) | .202(.131) | .446(.145) | .501(.125) | .303(.096) | .481(.199) | .554(.174) | .332(.166) |
| | PN [21] | .311(.129) | .317(.131) | .194(.126) | .467(.124) | .488(.140) | .317(.095) | .453(.184) | .568(.191) | .336(.131) |
| | RN [22] | .335(.131) | .355(.152) | .183(.132) | .381(.146) | .392(.127) | .224(.105) | .371(.182) | .366(.184) | .281(.150) |
| | DSN [37] | .409(.156) | .338(.132) | .368(.154) | .510(.101) | .478(.093) | .414(.097) | .563(.163) | .587(.176) | .482(.182) |
| | MAML [33] | .489(.142) | .476(.150) | .263(.151) | .495(.128) | .507(.166) | .361(.139) | .519(.140) | .526(.135) | .403(.146) |
| | HetNet [58] | .353(.066) | .374(.076) | .233(.088) | .465(.073) | .533(.058) | .364(.070) | .502(.054) | .532(.050) | .425(.077) |
| | SFENet [59] | .391(.086) | .396(.079) | .268(**.053**) | .411(.083) | .422(.078) | .275(.057) | .434(**.030**) | .426(**.025**) | .350(.074) |
| | **EmoDSN** | **.668(.062)** | **.654(.054)** | **.453**(.081) | **.778(.021)** | **.769(.044)** | **.583(.045)** | **.782**(.080) | **.778**(.096) | **.586(.110)** |
| **MERCA** | MN [20] | .371(.140) | .362(.133) | .192(.156) | .457(.136) | .524(.157) | .366(.123) | .455(.197) | .511(.178) | .356(.172) |
| | PN [21] | .283(.125) | .303(.142) | .211(.123) | .416(.145) | .511(.138) | .365(.109) | .435(.188) | .546(.178) | .385(.132) |
| | RN [22] | .365(.144) | .381(.138) | .185(.150) | .402(.130) | .451(.132) | .264(.120) | .369(.160) | .384(.192) | .237(.175) |
| | DSN [37] | .370(.127) | .405(.134) | .383(.124) | .429(.106) | .562(.090) | .446(.104) | .482(.184) | .602(.158) | .466(.173) |
| | MAML [33] | .517(.180) | .497(.167) | .393(.168) | .571(.146) | .582(.146) | .404(.153) | .568(.132) | .587(.098) | .421(.139) |
| | HetNet [58] | .400(.073) | .421(**.054**) | .279(.085) | .542(.082) | .536(.057) | .426(.086) | .557(.033) | .580(**.047**) | .435(**.088**) |
| | SFENet [59] | .404(.084) | .410(.080) | .247(**.074**) | .393(.083) | .397(.060) | .312(.053) | .448(**.026**) | .443(.054) | .337(.103) |
| | **EmoDSN** | **.683(.053)** | **.633**(.064) | **.432**(.090) | **.799(.033)** | **.763(.033)** | **.558**(.033) | **.802**(.065) | **.766**(.075) | **.553(.091)** |
| **CEAP 360VR** | MN [20] | .394(.138) | .423(.154) | .176(.168) | .411(.123) | .437(.123) | .357(.101) | .396(.196) | .502(.199) | .326(.178) |
| | PN [21] | .292(.125) | .336(.133) | .186(.156) | .407(.147) | .450(.155) | .345(.112) | .402(.185) | .521(.159) | .312(.136) |
| | RN [22] | .385(.141) | .386(.149) | .185(.150) | .396(.135) | .403(.136) | .271(.109) | .413(.160) | .425(.163) | .276(.126) |
| | DSN [37] | .401(.154) | .381(.135) | .358(.138) | .434(.094) | .507(.097) | .433(.082) | .473(.157) | .554(.194) | .446(.188) |
| | MAML [33] | .481(.144) | .486(.139) | .326(.138) | .496(.174) | .499(.124) | .424(.184) | .514(.130) | .521(.128) | .421(.107) |
| | HetNet [58] | .386(.058) | .360(**.080**) | .314(**.056**) | .538(.077) | .526(.053) | .436(.055) | .546(.049) | .558(.057) | .467(.081) |
| | SFENet [59] | .406(**.057**) | .401(.087) | .261(.064) | .411(.084) | .425(.056) | .315(.072) | .443(**.027**) | .435(**.030**) | .333(.106) |
| | **EmoDSN** | **.598**(.062) | **.625**(.084) | **.487**(.097) | **.720(.044)** | **.745(.035)** | **.561(.029)** | **.725**(.066) | **.742**(.077) | **.554(.073)** |

(MN) [20], Prototype network (PN) [21], Deep Siamese Network (DSN) [37], Relation Network(RN) [22] and Model Agnostic Meta Learning (MAML) [33]) and networks designed for physiological-signal-based emotion recognition (HetEmotionNet (HetNet) [58] and SFENet [59]). We choose the five FLS baselines because they are widely used by previous works for emotion recognition using similar data modalities (i.e., uni-dimensional data modalities such as speech [38] and physiological signals [13]). To implement a fair comparison, we fine-tune the structure of these methods to make them have the same embedding network we designed in section 5.3.2. Thus, the difference between each method is only the learning structure instead of the embedding network for feature extraction. We also use the same optimizer and learning rate (lr = 0.001) as *EmoDSN* to train all four few-shot learning algorithms. For HetNet, we construct the spatial-temporal and spatial-spectral graph (by DE features) and train them using the same graph recurrent neural network. Since the folding approach used by SFENet is based on the spatial distribution of EEG electrodes, we cannot use it for other physiological signals. Thus, we only use the 3D-CNN and ensemble learning designed in SFENet for comparison. We train the above algorithms with one, five and ten shot to compare their performance trained by different amounts of annotated samples. To test the stability of each algorithm, we run all the experiments 5 times [60, 61] and report the mean and SD of the accuracies.

### Accuracy comparison

Table 5.2 shows the results of the comparison. We observe that the gradient cannot descent (losses remain constant) when training the MN, PN and RN with 1-shot and 5-shot for 2D-5C. As shown in Fig 5.5, this problem does not occur when we use DSN: the losses descend rapidly after a few epochs (<10) for all personalized models in three datasets. The performance of MAML is better (acc is 8.89% and 6.94% higher for 1D-2C and 2D-5C respectively) than other FSL methods. However, the acc increase for MAML is not as significant as other FSL methods: when the number of training samples increases from 1-shot to 5-shot, the acc increase 3.04% and 6.94% on average for 1D-2C and 2D-5C respectively. The other FSL methods however, increases 9.22% and 10.76% for 1D-2C and 2D-5C respectively.

For the two fully supervised learning methods, we find a similar problem with MN, PN, and RN that for 2D-5C, the gradient cannot descent (losses remain constant) for 1-shot and 5-shot. Their average acc for 10-shot is also 5.89% lower than DSN. We also find the problem of overfitting for them when trained with 10-shot: the training acc increases rapidly over 90% after 5 epochs but the testing acc does not increase. The results demonstrate that the fully-supervised learning methods cannot achieve good performance when only a limited amount of data are used for training.

In general, the performance of *EmoDSN* is better than both the state-of-the-art FSL algorithms and fully-supervised algorithms. To compare the performance difference between *EmoDSN* and baseline methods, we follow the previous work of Kumar et al. [62] which use *Z-test* and *Chi-square* test to compare the classification accuracies. For both the *Z-test* and *Chi-square* test, we found significant differences (all p<0.01) between EmoDSN and MN ($Z = 14.21, \chi^2 = 5.48$), PN ($Z = 14.14, \chi^2 = 6.59$), RN ($Z = 21.04, \chi^2 = 8.98$), DSN ($Z = 14.49, \chi^2 = 2.99$), MAML ($Z = 12.36, \chi^2 = 2.17$), HetNet ($Z = 11.69, \chi^2 =$

3.06) and SFENet ($Z = 23.74, \chi^2 = 5.86$). The statistical analysis shows a significant difference between the performance of *EmoDSN* and other baseline methods.

#### Stability comparison

The stability of 5 FSL methods (i.e., MN, RN, PN, DSN, MAML) is lower than the two supervised learning algorithms (i.e., HetNet and SFENet): the SD for the 5 experiments is 7.81% higher. When the number of training samples increases to 10-shot, the SD difference between FSL and fully-supervised learning methods also increases accordingly (on average 10.75% for 10-shot). FSL algorithms learn the difference (MN, RN, PN, DSN) or train a meta learner (MAML) between training samples instead of learning the exact mapping between samples and labels. Thus, their performance depends on the quality of training samples, which leads to instability if we consider all training samples to be correctly labeled [63]. The fully-supervised learning methods however, optimize the classifier among all training samples. Thus, they converge on a worse (i.e., low acc) but comparatively stable model if only few samples are used for training. The results are in line with our conclusion in ablation study that we cannot get stable and accurate recognition results if we assume all $P(s_{nm} \in l_m)$ are to be 1.

### 5.4.7. Ablation study

#### Implementation details

We conduct an ablation study to verify the effectiveness of each component in *EmoDSN*. We begin with only using the Vanilla Siamese (VS) structure to train the network. The VS structure directly uses the raw signal segments without passing them through the embedding network. Then we test the performance of combining the VS with the Embedding Network (EN) described in section 5.3.2. For the two above experiments, instead of using Distance Fusion (DF), we follow the traditional strategy of few-shot learning algorithms: average the distances with the same emotion labels and predict the samples as the emotion label with the closest distance. Finally, we replace the simple averaging with the DF described in section 5.3.4 for the complete *EmoDSN*. To test the stability of *EmoDSN*, we also repeat the experiments 5 times [60, 61] and report the mean and SD of acc.

#### Accuracy comparison

From the results (shown in Table 5.3) we can see both EN and DF contribute to the classification tasks. The EN benefits the classification tasks by extracting deep features and taking reaction delay into consideration. Thus, the accuracies increase 11.85% on average after combining EN to VS. We also observe a significant increase of accuracies (more than 20% for 1D-2C and 10% for 2D-5C) after adding the distance fusion module. This finding demonstrates that simply averaging the distances from different shot is not suitable for fine-grained emotion recognition using physiological signals. It necessitates considering the probability that some mislabelled training samples can significantly lower the model accuracy. In conclusion, the observations above demonstrate the effectiveness of the components in the proposed algorithm.

Table 5.3: Ablation study (acc (sd)) for Vanilla Siamese (VS), Embedding Network (EN) and Distance Fusion (DF)

| | **Dataset** | **VS** | **VS+EN** | **EmoDSN**<br>**VS+EN+DF** |
|---|---|---|---|---|
| **1D-2C valence** | **CASE** | .412(.119) | .480(.101) | .773(.021) |
| | **MERCA** | .337(.134) | .558(.106) | .789(.033) |
| | **CEAP-360VR** | .413(.145) | .510(.093) | .718(.043) |
| **1D-2C arousal** | **CASE** | .354(.135) | .501(.092) | .777(.044) |
| | **MERCA** | .365(.131) | .437(.089) | .772(.033) |
| | **CEAP-360VR** | .320(.147) | .440(.097) | .749(.035) |
| **2D-5C** | **CASE** | .314(.162) | .406(.097) | .587(.045) |
| | **MERCA** | .293(.159) | .443(.104) | .561(.034) |
| | **CEAP-360VR** | .322(.162) | .423(.080) | .569(.029) |

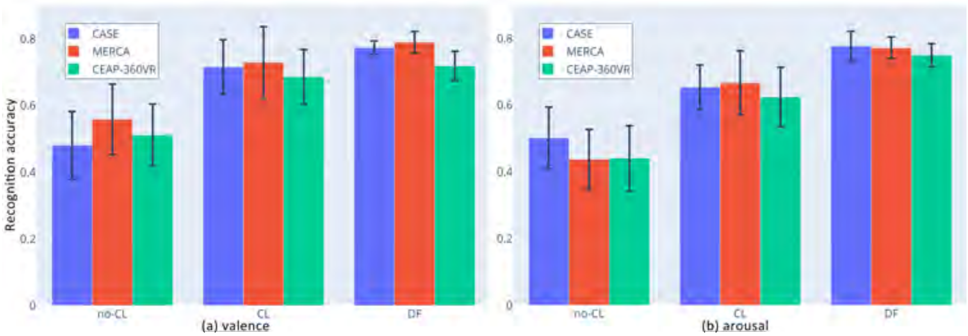
#### Stability comparison

For the comparison of SD, we find both VS and VS-EN have relatively unstable performance: the average SD is 14.39% and 9.58% for VS and VS-ED respectively. Adding DF however, can improve the stability of the network by decreasing the SD to 1.6%. When randomly selecting only few training samples, some samples with low-confidence annotation will affect the performance of the network. If all $P(s_{nm} \in l_m)$ are assumed to be 1, the network is unstable because the performance is related to the quality of labels selected for training. However, DF modules can decrease the instability by assigning less confident samples lower weights for classification. The results demonstrate the necessity and effectiveness of adding DF into *EmoDSN*.

### 5.4.8. Effectiveness of DF module

To further clarify the effectiveness of the distance fusion (DF) module for the wrong labels, we use it to identify potentially wrong labels and correct them when classifying emotions. Specifically, we first calculate the $P(s_{nm} \in l_m)$ for all training samples using equation 5.4. $P(s_{nm} \in l_m)$ represents the probability of training sample $s_{nm}$ corresponds to the emotion label $l_m$. If the probability is lower than 0.5, we assume the sample is mislabeled and correct it. Here we only run the experiments for 1D-2C because we cannot estimate the correct label of $s_{nm}$ for multi-class classification if $P(s_{nm} \in l_m)$ is low. For multi-class classification, if we know $P(s_{nm} \in l_m)$ <0.5: we do not know which $i$ can satisfy $P(s_{nm} \in l_{i,i\neq m})$. However, for binary classification, since $P(s_{nm} \in l_m)+P(s_{nm} \in l_{i,i\neq m})$ = 1, if $P(s_{nm} \in l_m)$<0.5 we can easily know $P(s_{nm} \in l_{i,i\neq m})$>0.5. Thus, the mislabeled samples are corrected as the label opposite to its original annotation. Then we average the distance (D) with the same emotion labels after the label correction and predict the testing sample as the emotion label with the greatest possibility. Then we compare the recognition accuracies among the network a) without the Correction of Labels (no-CL), b) with the Correction of Labels (CL) and c) with the DF module. To ensure the stability of the experiment, we run the experiment 5 times [60, 61] and report the average acc and the SD of 5 experiments.

As shown in Fig 5.9, after the correction of labels, the accuracies increase 19.12% on



(a) valence

(b) arousal

Figure 5.9: Recognition accuracies among the network a) without the Correction of Labels (no-CL), b) with the Correction of Labels (CL) and c) with the DF module.

average of three datasets. Since both no-CL and CL use the simple averaging distance learned by the DSN, the detection of mislabeled samples can promote the classification performance of *EmoDSN*. However, we also find that using DF can result in an average acc increase of 8.43% compared with using CL. In addition, the performance of DF is more stable than CL: the average SD of DF is 5.26% lower than CL. The difference between the network with DF and CL is that DF uses a soft weighted average of D to estimate the emotion label. CL uses an arithmetic average of D after correcting the labels of the samples whose $P(s_{nm} \in l_m) < 0.5$. Thus, for few-shot learning based fine-grained emotion recognition, assigning low weights for an inexactly labeled sample can result in better and more stable performance compared with simply correcting it according to the intra data distribution of training samples (i.e., whether the distribution of this sample is coherent with others).

### 5.4.9. RUNNING TIME AND EFFICIENCY

The average training time for different methods are shown in TABLE 5.4. Our model is implemented using Keras and Tensorflow. All our experiments are performed on a desktop with NVIDIA RTX 2080Ti GPU with 16 GB RAM. The one-to-many learning structure is used by MN, PN, RN and MAML. Thus, the number of training samples for them is $n(n-1)\cdot k^2$, where $n$ and $k$ are the numbers of shots and classes of the learning task respectively. Our method uses the pair-by-pair learning structure. The number of training samples is $\Sigma_{i=1}^{n\cdot k-1}(n\cdot k-i)$. For the fully-supervised learning methods, training samples are directly input into the network without combining them into different pairs. Thus, the number of training samples for them is $n\cdot k$.

Although the fully-supervised methods have a more complex structure, the number of samples for training is less than FSL methods. Thus, their training time is shorter than FSL methods. However, they do not achieve up-to-chance level acc because their learning structures are not designed for converging on a small amount of training samples. For the FSL methods, the pair-by-pair learning structure used by our method results in fewer training samples compared with other FSL methods using one-to-many structures. Thus, our method requires less training time: *EmoDSN* requires only 54.83% of the

Table 5.4: Average training time for different methods

| | **5-shot-2C** | **5-shot-5C** | **10-shot-2C** | **10-shot-5C** |
|---|---|---|---|---|
| MN [20] | 125.23 (s) | 652.65 (s) | 432.25 (s) | 3025.65 (s) |
| PN [21] | 118.16 (s) | 752.45 (s) | 354.72 (s) | 2546.36 (s) |
| RN [22] | 156.24 (s) | 819.65 (s) | 495.25 (s) | 3432.24 (s) |
| MAML [33] | 245.24 (s) | 792.68 (s) | 419.88 (s) | 2653.24 (s) |
| HetNet [58] | 126.54 (s) | **252.32 (s)** | 198.27 (s) | 1025.56 (s) |
| SFENet [59] | 87.26 (s) | 256.78 (s) | **175.36 (s)** | **986.71 (s)** |
| **EmoDSN** | **76.53 (s)** | 419.25 (s) | 269.54 (s) | 1543.25 (s) |

average training time of other FSL methods. The result demonstrates the good efficiency of *EmoDSN* compared with baselines using both fully-supervised and FSL methods.

Although the result of 10-shot is better, it requires much more training time compared with 5-shot. As shown in TABLE 5.4, training the model using 10-shot takes almost 4 times as long as training the model using 5-shot. The testing acc and m-F1 score however, only increases 0.15% and 0.20% on average for the three datasets. Increasing the training samples from 1-shot to 5-shot however, result in the increase of acc and m-F1 for 11.58% and 9.32% respectively. Thus, using 5-shot makes a trade-off between training time and model accuracies.

## 5.5. Discussion

### 5.5.1. Reaction delay of continuous annotation

According to the research of Metallinou et al. [64], there are time delays (e.g., due to gender, age, distraction levels) between the occurrence of an emotional event and its annotation considering that continuous annotations are performed in real-time. If we use misaligned annotation as labels to train the network, it will overfit or not converge. Most of the previous works [15–17] use visual features from video stimuli to align the annotation. In these approaches (also known as *explicit compensation* [65]), the delay compensation and the emotion prediction are performed separately. However, these approaches assume that the reaction delay is fixed for different users watching the same video stimuli. This assumption is untenable as the reaction time is both stimulus dependent and individual dependent [65].

The last layer of *EmoDSN* can identify which sub-instances (signal segments with different time of delay) can better predict the fine-grained emotion labels. Once the network is trained, we can observe the instance gains in the last layer to find out with how much delay the network can perform the best. Our approach belongs to the *implicit compensation* [65], which compensates for delays while modeling the relationship between input signals and emotion labels. The uniqueness of our approach is that we do not have to manually adjust the parameters (e.g., the width of analysis window for LSTM [18] or the receptive field for CNN [66]) in the network for compensating different delays for different individuals.

To obtain the range of reaction delay, we first run the 1D-2C task and get the delays of the sub-instances with maximum instance gain (i.e., have the highest probability to predict emotion labels). We follow the procedure of previous works [15, 17] that estimate the delay of each dimension (valence and arousal) separately. Fig 5.10 shows the box plot of reaction delays estimate by *EmoDSN* for three datasets.

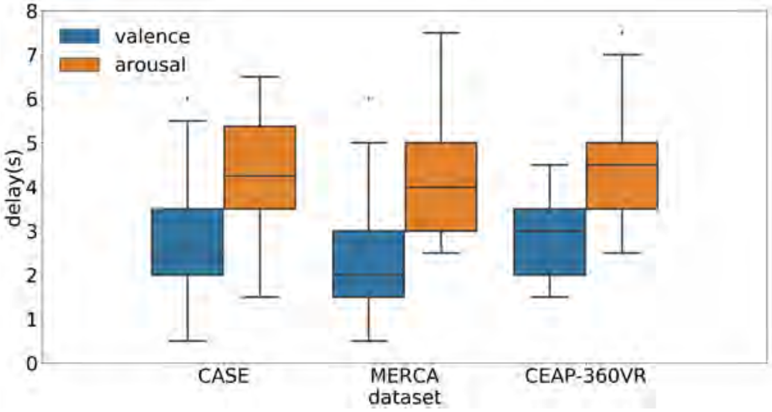


Figure 5.10: The reaction delays for valence and arousal respectively

The mean and standard deviation of delays are: CASE = 2.59(1.47), MERCA = 2.50(1.43), CEAP-360VR = 2.89(1.03) and CASE = 4.05(1.45), MERCA = 4.21(1.42), CEAP-360VR = 4.38(1.27) for valence and arousal respectively. The mean delay for arousal is higher than the delay for valence for all the three datasets. A Shapiro-Wilk test shows that the delays for both valence and arousal in three datasets are all normally distributed (all p > 0.05 for three datasets). For the comparison between different scenarios (desktop, mobile and VR for CASE, MERCA and CEAP-360VR respectively), we perform a ANOVA. Here we do not find a significant effect of scenarios on both valence ($F(2,80) = 0.795, p = 0.455, \eta_p^2 = 0.019$ and arousal ($F(2,80) = 0.416, p = 0.661, \eta_p^2 = 0.010$. However, through Welch's t tests, we do find there is significant difference between the delay of valence and arousal for CASE (t(58)=2.869, p < 0.01, Cohen's $d = 0.944$), MERCA (t(40)=3.804, p < 0.01, Cohen's $d = 1.372$) and CEAP-360VR (t(62)=5.045, p < 0.01, Cohen's $d = 1.340$) respectively.

These results show that users need more time to react for annotating arousal than valence. This finding is coherent with most of the previous works using *explicit* [15–17] compensation methods. The averaged delays (2.66s and 4.21s for V-A) obtained by our method are also similar to the results obtained by *explicit* methods (e.g., 2s and 4s from Huang et al. [15], 3.08s and 3.95s from Mariooryad et al. [16] for V-A respectively). Thus, our method for compensating reaction delay can provide similar results without using visual and audio features from stimuli. The average annotation delays in different datasets collected in different scenarios are comparable. The reason for this finding is that the annotations of all three datasets were collected using the joystick-based annotation interface.

We also conduct an experiment to find out whether sliding windows with long delays can introduce redundant information from other temporal moments for emotion recognition. Fig 5.11 shows the relationship between the steps of sliding windows and

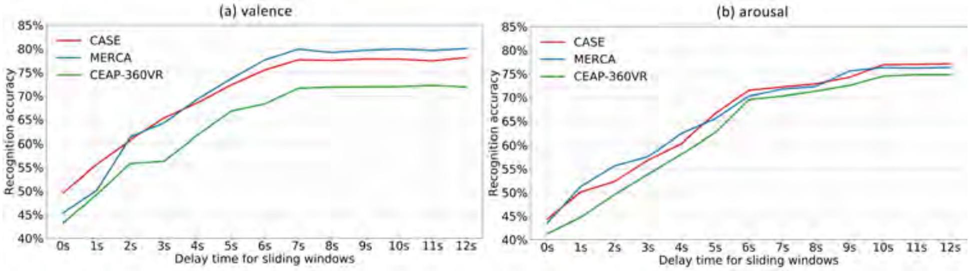


Figure 5.11: The relationship between the steps of sliding windows and recognition acc

recognition acc for 1D-2C arousal and valence respectively. The recognition acc keeps increasing for both valence and arousal recognition when the steps of delay increase from 0s to 7s. The low acc caused by the short delay time of sliding windows show that if the sliding windows cannot cover enough delay, the embedding network will fail to identify the sub-instances which represent the emotion label of that moment. The recognition acc for V-A becomes stable after increasing the delay of sliding windows for 7s and 10s respectively. Thus, the noise of adding steps of sliding windows can be filtered by the MIL module: the recognition acc do not decrease when we add the steps of the sliding window. Instead of fully-supervised learning, all the sub-instances are weakly supervised by the emotion labels. The weights learned by MIL layers represent the probability of one sub-instance for discriminating samples between different emotion categories. Thus, the redundant information from other temporal moments can be automatically filtered (i.e., assign low weights). The results are also in line with the finding that the annotation delay of arousal is higher than the delay of valence: we need to add more steps (i.e., delay time) of sliding windows to cover the corresponding sub-instances for arousal recognition.

### 5.5.2. DO THE TEMPORAL MOMENTS OF TRAINING SAMPLES AFFECT THE PERFORMANCE?

In section 5.4, we randomly sample $N$ training samples from each emotion category to train *EmoDSN*. Although it is the standard evaluation procedure to test few-shot learning algorithms, it is difficult to get randomly balanced number of samples with different emotion labels. When applying the algorithm for evaluating the user experience of watching videos, the possible methods are 1) randomly stop the video and ask users to annotate their emotions or 2) ask users to annotate at some fixed temporal moments to obtain the emotion labels for training. Since we use only few annotated samples to train the network, we want to find out samples from which temporal moments can better represent the distribution for the whole video watching and result in better recognition results. We also want to explore the amount of samples *EmoDSN* needs to obtain accurate recognition when selecting training samples in different temporal moments of video watching. Answering these two questions can help researchers maximize the performance of *EmoDSN* and minimize the amount of training samples by asking users to annotate at the most suitable temporal moments in video watching.

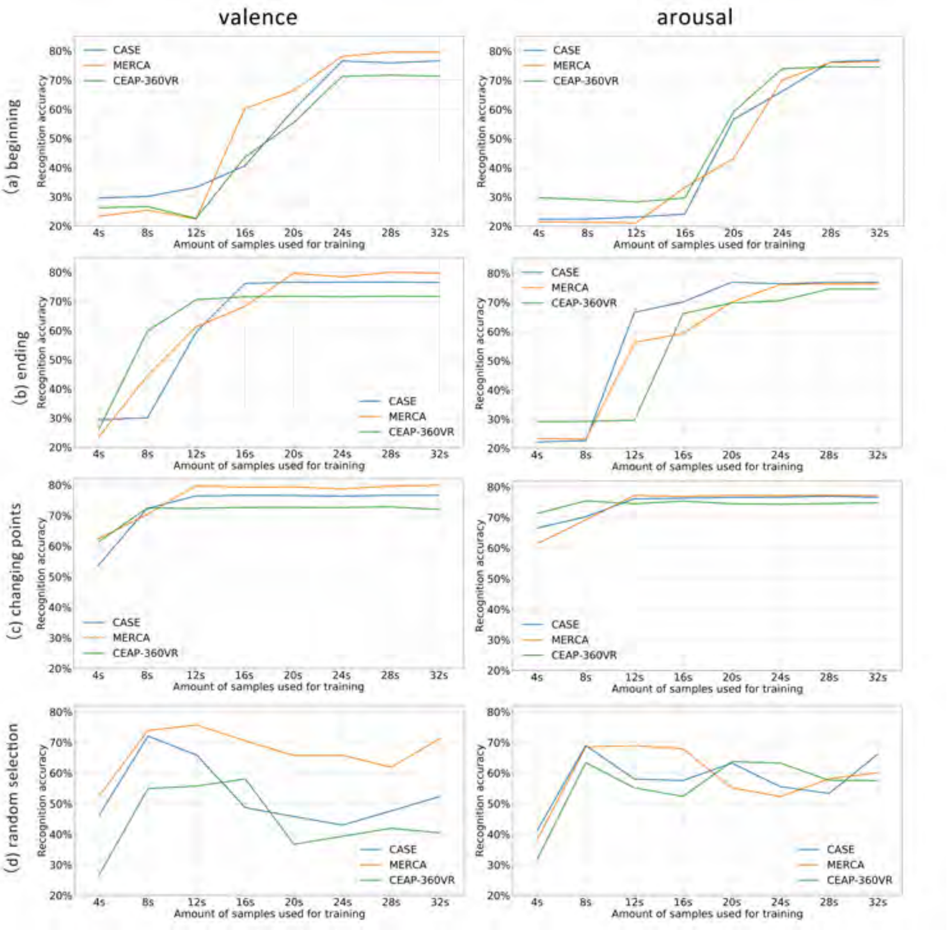


Figure 5.12: The 1D-2C recognition acc when training *EmoDSN* with samples from the a) beginning, b) ending, c) changing points and d) random position of video watching. The granularity of samples is 2 seconds. The amount of samples are shown in the unite of seconds (e.g., 4 seconds = 2 samples)

To achieve this, we select training samples from both fixed and random temporal moments of video watching and compare the recognition acc (1D-2C) when training with different amounts of samples. Specifically, we choose the beginning, ending and the changing points as fixed temporal moments and compare the result with the random moments:

- ***Beginning:*** We choose the first $K$ samples from a video watching as training samples and test on the rest.
- ***Ending:*** We choose the last $K$ samples from a video watching as training samples and test on the rest.
- ***Changing points:*** According to the research of Sharma et al. [23], the changing points

in continuous annotation can signify emotionally salient moments. Thus, we want to find out whether these samples can better represent the distribution for the whole video watching. We select samples from the changing points of annotation (obtained using the Changing Points Analysis (CPA) [23]) as the training samples and test on the remaining samples.

- ***Random:*** We randomly choose $K$ samples from one video watching as the training samples and test on the remaining samples. Unlike balanced random selection in section 5.4, it does not ensure each emotion category has a balanced number of training samples.

The results of how the acc of *EmoDSN* changes with different amounts of samples from different temporal moments of video watching is shown in Fig 5.12. From Fig 5.12 (d) we observe that random selection results in great fluctuation of the recognition acc when more samples are used for training. Selecting from fixed temporal moments however (Fig 5.12 (a)-(c)), results in relatively stable performance when inputting more training samples. Thus, selecting training samples from fixed temporal moments can result in more stable performance when we only use few annotated samples for training.

We also observe that if we choose training samples from the beginning of video watching, the algorithm needs more training samples to converge. It needs more than 10 training samples (20 seconds) to increase the recognition acc above 50%. Using the ending moments however, requires less than 8 training samples (16 seconds) to achieve 50% acc. The best temporal moments to select the training samples are the changing points: the acc exceeds 70% by only using 4 samples (8 seconds) for training. Thus, the samples at the changing points and the ending moments can better represent the distribution of the whole video watching and result in better acc with fewer samples.

The results we obtain are coherent with the *peak-end theory* [67] that the most salient (peak) or recent (end) moments can better represent the emotions of users while watching videos. We also observe that the distributions of samples with specific emotion labels are different across the temporal moments. Fig 5.13 shows the percentage of samples with high/low V-A labels in different temporal moments of video watching. Compared with the ending moments of video watching, most of the samples (more than 70% for all three datasets) from the beginning moments are labeled as high V-A. If we choose these samples as training data, the imbalanced training set can result in mis-convergence of the learning network (e.g., in Fig 5.12 (a) when the amount of samples < 12s, acc < 30%). It also explains why fewer training samples are required from the end of video watching for good results: the samples are more balanced at the end of video watching.

In conclusion, the temporal moments of training samples do have influence on the performance of *EmoDSN*. The take-way message from this experiment is that samples from the changing points of emotion and the ending moments of video watching are better training samples when only few samples are available for building up an emotion recognition system.

## 5.6. Conclusion and limitations

Fine-grained emotion recognition requires training the algorithm with large amounts of continuous emotion labels. In this chapter, we propose *EmoDSN*, a Deep Siamese

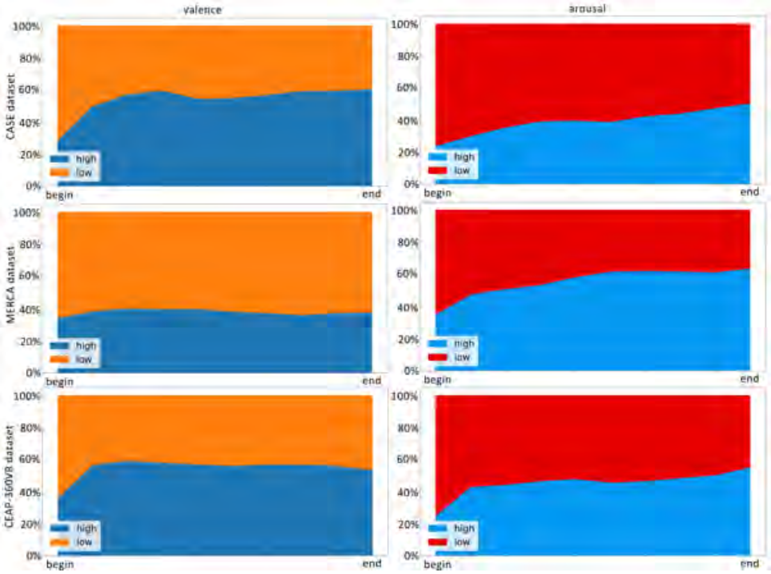


Figure 5.13: Percentage of samples with high/low V-A labels in different temporal moments of video watching for CASE, MERCA and CEAP-360VR

Network based few-shot learning algorithm to classify fine-grained valence and arousal with only a small amount of annotated signals. The embedding network of *EmoDSN* enables our algorithm to compensate the reaction delay of annotation while predicting the fine-grained valence and arousal. The distance fusion module of *EmoDSN* minimizes the overfitting problem caused by mislabeled training samples. The proposed algorithm achieves reasonable performance (averaged accuracy of 76.04%, 76.62% and 57.62% for 1D-2C valence, 1D-2C arousal and 2D-5C respectively) by using only 5 shots as training data for subject-dependent testing on three datasets collected in three different environments (i.e., desktop, mobile, and HMD-based VR). Our algorithm also outperforms other few-shot learning algorithms which are widely used for emotion recognition (**RQ 3.1**). The ablation study shows that the embedding network and distance fusion module, which are specifically designed for physiological signals based fine-grained emotion recognition, can significantly improve the recognition accuracy. Our experiment on reaction delay of annotation shows that 1) the reaction delay for arousal is longer than the delay for valence and 2) the reaction delays between different scenarios have no significant difference. We also find that the changing points of emotion annotation and the ending moments of video watching are better temporal moments for selecting training samples: if we select training samples from these two temporal moments, *EmoDSN* can provide better recognition results with fewer annotated samples (**RQ 3.2**).

Given the challenges of predicting valence and arousal labels at a fine level of granularity using only few annotated samples, there are natural limitations to our work. First, *EmoDSN* only works well for the personalized or subject-dependent emotion recognition model. Since the patterns of physiological signals are highly variable between subjects [30, 54], using few annotated samples to model it is challenging and relies on the careful selection of training samples. In the future, we will extend *EmoDSN* for subject-independent emotion recognition model by finding out which training samples can represent the inter-subject variability of physiological signals. It is also essential for us to compare the performance of *EmoDSN* on more datasets to further test its generalizability. However, the number of datasets with continuously annotated physiological signals is to date limited. It lacks benchmark results using basic few-shot learning methods. Thus, it is difficult to make comparisons with more few-shot learning methods.

In Chapter 3, 4, and 5, we propose full-supervised, weakly-supervised and few-shot learning algorithms respectively to address the challenge of recognizing emotions in a fine level of granularity with a limited amount of annotations for training. In chapter 6, the final chapter of the thesis, we wrap up important findings and answer the research questions raised in Chapter 1.

## References

[1] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors,* Sensors **21**, 52 (2021).

[2] T. Zhang, A. El Ali, C. Wang, X. Zhu, and P. Cesar, *Corrfeat: Correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition,* in *International Conference on Multimodal Interaction* (2019) pp. 404–408.

[3] M. A. Jarwar and I. Chong, *Web objects based contextual data quality assessment model for semantic data application,* Applied Sciences **10**, 2181 (2020).

[4] A. Srinivasan, S. Abirami, N. Divya, R. Akshya, and B. Sreeja, *Intelligent child safety system using machine learning in iot devices,* in *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (IEEE, 2020) pp. 1–6.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, *Affectnet: A database for facial expression, valence, and arousal computing in the wild,* IEEE Transactions on Affective Computing **10**, 18 (2017).

[6] G. Van Houdt, C. Mosquera, and G. Napoles, *A review on the long short-term memory model,* Artificial Intelligence Review **53**, 5929 (2020).

[7] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2020) pp. 1–15.

[8] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* Scientific data **6**, 1 (2019).

[9] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, *Introducing the recola multimodal corpus of remote collaborative and affective interactions,* in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (IEEE, 2013) pp. 1–8.

[10] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, *Avec 2016: Depression, mood, and emotion recognition workshop and challenge,* in *Proceedings of the 6th international workshop on audio/visual emotion challenge* (2016) pp. 3–10.

[11] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, *K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations,* arXiv preprint:2005.04120 (2020).

[12] Y. Wang and Q. Yao, *Few-shot learning: A survey,* CoRR **abs/1904.05046** (2019).

[13] S. Jiang, F. Firouzi, K. Chakrabarty, and E. Elbogen, *A resilient and hierarchical iot-based solution for stress monitoring in everyday settings,* (2021).

[14] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, *Zero-shot emotion recognition via affective structural embedding,* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019) pp. 1151–1160.

[15] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, *An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction,* in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* (2015) pp. 41–48.

[16] S. Mariooryad and C. Busso, *Correcting time-continuous emotional labels by modeling the reaction lag of evaluators,* IEEE Transactions on Affective Computing **6**, 97 (2014).

[17] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, *Rcea-360vr: Real-time, continuous emotion annotation in 360 vr videos for collecting precise viewport-dependent ground truth labels,* in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2021).

[18] D. Le, Z. Aldeneh, and E. M. Provost, *Discretized continuous speech emotion recognition with multi-task deep recurrent neural network.* in *Interspeech* (2017) pp. 1108–1112.

[19] A. Patane and M. Kwiatkowska, *Calibrating the classifier: siamese neural network architecture for end-to-end arousal recognition from ecg,* in *International Conference on Machine Learning, Optimization, and Data Science* (Springer, 2018) pp. 1–13.

[20] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, *Matching networks for one shot learning,* Advances in Neural Information Processing Systems **29**, 3630 (2016).

[21] J. Snell, K. Swersky, and R. S. Zemel, *Prototypical networks for few-shot learning,* arXiv preprint:1703.05175 (2017).

[22] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, *Learning to compare: Relation network for few-shot learning,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018) pp. 1199–1208.

[23] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek, *Continuous, real-time emotion annotation: A novel joystick-based analysis framework,* IEEE Transactions on Affective Computing (2017).

[24] G. Chen, Y. Zhu, Z. Hong, and Z. Yang, *Emotionalgan: generating ecg to enhance emotion state classification,* in *Proceedings of the International Conference on Artificial Intelligence and Computer Science* (2019) pp. 309–313.

[25] Q. Zhong, Y. Zhu, D. Cai, L. Xiao, and H. Zhang, *Electroencephalogram access for emotion recognition based on a deep hybrid network,* Frontiers in Human Neuroscience **14** (2020).

[26] Y. Jiao, Y. Deng, Y. Luo, and B.-L. Lu, *Driver sleepiness detection from eeg and eog signals using gan and lstm networks,* Neurocomputing **408**, 100 (2020).

[27] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, *Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography,* IEEE Journal of Biomedical and Health Informatics **25**, 1373 (2021).

[28] A. T. Zhang and B. O. Le Meur, *How old do you look? inferring your age from your gaze,* in *2018 25th IEEE International Conference on Image Processing (ICIP)* (IEEE, 2018) pp. 2660–2664.

[29] P. Salehi, A. Chalechale, and M. Taghizadeh, *Generative adversarial networks (gans): An overview of theoretical model, evaluation metrics, and recent developments,* arXiv preprint:2005.13178 (2020).

[30] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, *Multiple instance learning for emotion recognition using physiological signals,* IEEE Transactions on Affective Computing (2019).

[31] E. Pei, D. Jiang, M. Alioscha-Perez, and H. Sahli, *Continuous affect recognition with weakly supervised learning,* Multimedia Tools and Applications **78**, 19387 (2019).

[32] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, *Long short term memory recurrent neural network based multimodal dimensional emotion recognition,* in *Proceedings of the 5th international workshop on audio/visual emotion challenge* (2015) pp. 65–72.

[33] C. Finn, P. Abbeel, and S. Levine, *Model-agnostic meta-learning for fast adaptation of deep networks,* in *International Conference on Machine Learning* (PMLR, 2017) pp. 1126–1135.

[34] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, *Introducing wesad, a multimodal dataset for wearable stress and affect detection,* in *Proceedings of the 20th ACM international conference on multimodal interaction* (2018).

[35] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, *Signature verification using a "siamese" time delay neural network,* International Journal of Pattern Recognition and Artificial Intelligence **7**, 669 (1993).

[36] W. Hayale, P. S. Negi, and M. Mahoor, *Deep siamese neural networks for facial expression recognition in the wild,* IEEE Transactions on Affective Computing (2021).

[37] A. Mehmood, M. Maqsood, M. Bashir, and Y. Shuyuan, *A deep siamese convolution neural network for multi-class classification of alzheimer disease,* Brain sciences **10**, 84 (2020).

[38] K. Feng and T. Chaspari, *A siamese neural network with modified distance loss for transfer learning in speech emotion recognition,* arXiv preprint:2006.03001 (2020).

[39] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, *A review of emotion recognition using physiological signals,* Sensors **18**, 2074 (2018).

[40] J. Fleureau, P. Guillotel, and I. Orlac, *Affective benchmarking of movies based on the physiological responses of a real audience,* in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (IEEE, 2013) pp. 73–78.

[41] Y. Chu, X. Zhao, J. Han, and Y. Su, *Physiological signal-based method for measurement of pain intensity,* Frontiers in neuroscience **11**, 279 (2017).

[42] P. Karthikeyan, M. Murugappan, and S. Yaacob, *Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress,* Journal of Physical Therapy Science **24**, 1341 (2012).

[43] M. Awais, M. Raza, N. Singh, K. Bashir, U. Manzoor, S. ul Islam, and J. J. Rodrigues, *Lstm based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19,* IEEE Internet of Things Journal (2020).

[44] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, *Wavenet: A generative model for raw audio,* arXiv preprint:1609.03499 (2016).

[45] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, *Large kernel matters–improve semantic segmentation by global convolutional network,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017) pp. 4353–4361.

[46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, *Inception-v4, inception-resnet and the impact of residual connections on learning,* in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (2017) pp. 4278–4284.

[47] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition,* arXiv preprint:1409.1556 (2014).

[48] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, *A sufficient condition for convergences of adam and rmsprop,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2019) pp. 11127–11135.

[49] E. Meijering, *A chronology of interpolation: from ancient astronomy to modern signal and image processing,* Proceedings of the IEEE **90**, 319 (2002).

[50] R. W. Daniels, *Approximation methods for electronic filter design: with applications to passive, active, and digital networks* (McGraw-Hill New York, NY, USA:, 1974).

[51] P. Van Gent, H. Farah, N. Nes, and B. van Arem, *Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data,* in *Proceedings of the 6th HUMANIST Conference* (2018) pp. 173–178.

[52] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, *Using deep and convolutional neural networks for accurate emotion classification on deap dataset.* in *Twenty-ninth IAAI conference* (2017).

[53] Y. Li, J. Huang, H. Zhou, and N. Zhong, *Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks,* Applied Sciences **7**, 1060 (2017).

[54] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, *Multi-task and multi-view learning of user state,* Neurocomputing **139**, 97 (2014).

[55] J. Na, H. Jung, H. J. Chang, and W. Hwang, *Fixbi: Bridging domain spaces for unsupervised domain adaptation,* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021) pp. 1094–1103.

[56] Y. Wang, S. Qiu, X. Ma, and H. He, *A prototype-based spd matrix network for domain adaptation eeg emotion recognition,* Pattern Recognition **110**, 107626 (2021).

[57] M. Chen, Y. Fang, X. Wang, H. Luo, Y. Geng, X. Zhang, C. Huang, W. Liu, and B. Wang, *Diversity transfer network for few-shot learning,* in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (2020) pp. 10559–10566.

[58] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, *Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition,* in *Proceedings of the 29th ACM International Conference on Multimedia* (2021) pp. 1047–1056.

[59] X. Deng, J. Zhu, and S. Yang, *Sfe-net: Eeg-based emotion recognition with symmetrical spatial feature extraction,* arXiv preprint:2104.06308 (2021).

[60] H. Zhang, N. M. Nasrabadi, T. S. Huang, and Y. Zhang, *Transient acoustic signal classification using joint sparse representation,* in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE, 2011) pp. 2220–2223.

[61] S. Golkar, M. Kagan, and K. Cho, *Continual learning via neural pruning,* arXiv preprint:1903.04476 (2019).

[62] P. Kumar, R. Prasad, A. Choudhary, V. N. Mishra, D. K. Gupta, and P. K. Srivastava, *A statistical significance of differences in classification accuracy of crop types using different classification algorithms,* Geocarto International **32**, 206 (2017).

[63] M. Goldblum, L. Fowl, and T. Goldstein, *Robust few-shot learning with adversarially queried meta-learners,* (2019).

[64] A. Metallinou and S. Narayanan, *Annotation and processing of continuous emotional attributes: Challenges and opportunities,* in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (IEEE, 2013) pp. 1–8.

[65] S. Khorram, M. McInnis, and E. M. Provost, *Jointly aligning and predicting continuous emotion annotations,* IEEE Transactions on Affective Computing (2019).

[66] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, *Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition,* arXiv preprint:1708.07050 (2017).

[67] B. L. Fredrickson and D. Kahneman, *Duration neglect in retrospective evaluations of affective episodes.* Journal of personality and social psychology **65**, 45 (1993).

# 6
# Conclusion

*Throughout the main body of the thesis, we implemented a series of studies exploring ways to recognize emotions at a fine granularity level with a limited amount of annotations. We start with the study of limitations of fine-grained emotion recognition by training with fine-grained emotion labels and fully-supervised learning methods. These methods require segment-by-segment emotion labels, which are costly and time-consuming to collect. Thus we propose a weakly-supervised learning algorithm in Chapter 4 which recognizes emotions at a fine granularity level by training with only post-stimuli emotion labels. Although the algorithm developed in Chapter 4 requires less annotation, it can only identify the segments with post-stimuli emotion and categorize the emotions in all other segments as neutral. To address this limitation, in Chapter 5 we propose a few-shot learning network which can also be trained with a limited amount of annotations and provide more abundant predictions (instead of only post-stimuli and neutral emotions) compared to the algorithm in Chapter 4. In this final chapter, we first answer the research questions formulated in the introductory chapter (Section 6.1). After that, we distill some important takeaway messages and thoughts collected throughout the thesis (Section 6.2). At last, we reflect on the limitations and potential alternative methods related to the topic of this thesis and propose the way forward (Section 6.3).*

## 6.1. REVISITING THE RESEARCH QUESTIONS

The main challenge of the thesis is to explore whether accurate fine-grained emotion recognition can be achieved with a limited amount of emotion labels as ground truth for training. To answer this question, we need to first understand how accurate the recognition can be if we use the fine-grained emotion labels. In Chapter 3, we build up the baselines for recognizing emotions at a fine granularity level by using fine-grained emotion labels as the ground truth for training. Our objective is to explore which kind of machine learning structure can learn the joint features between signal segments without having the problem of overfitting (RQ1). Hence, we propose a correlation-based instance learning approach (*CorrNet*) and compare it with state-of-the-art fine-grained emotion recognition methods to answer the research question 1.1:

**RQ 1.1:** *Can correlation-based instance learning offer advantages for fine-grained emotion recognition compared with sequence learning?*

The evaluation of *CorrNet* shows that the correlation-based instance learning (CorrNet) can provide more accurate recognition results compared to sequence learning (i.e., LSTM and BiLSTM) and other deep learning methods. We found the performance of sequence learning to be similar to other deep learning based methods (e.g, 1D CNN), which means the recurrent structure does not help to increase the recognition accuracy. Our finding is consistent with the one of a previous study on continuous speech emotion recognition [1] (speech signals and physiological signals are similar because they are both time series data) stating that the recurrent structures are not necessary to accomplish the task of time-continuous emotion recognition. Compared to sequence learning methods, correlation-based instance learning offers three advantages for fine-grained emotion recognition:

- **Higher recognition accuracy:** The recognition of *CorrNet* outperforms the sequence learning methods since it takes advantage of information both across modalities and their correlation. Validation results show that *CorrNet* achieves the accuracy of >70% averaged on all the testing datasets for binary valence and arousal classification. The accuracy of sequence learning methods however, is 50.71% averaged on the testing datasets.

- **Less overfitting:** For 3-class classification (low/neutral/high valence and arousal), the sequence learning approaches achieve higher accuracy (>1.52%) but lower F1 score (<10%) compared to *CorrNet*. High accuracy and low F1 score means that the network has the problem of overfitting because it performs well only on a specific class (i.e. neutral emotion). Thus, compared to sequence learning approaches, *CorrNet* has less overfitting problems and generalizes across different classes.

- **Less computational complexity:** The computational cost of *CorrNet* is low due to the simple (2-layer) structure for intra-modality feature learning and the linear mapping (instead of other complex transformations) in correlation-based feature extraction. The sequence learning methods however, need a large amount of time to train the recurrent structure. Thus, the average training time for sequence learning methods

(189.72s) is more than 4 times of the training time for *CorrNet* (45.11s) on our testing datasets.

When we compare the performance of baseline methods, we found that all the state-of-the-art deep learning methods for fine-grained emotion recognition suffer from the problem of overfitting: the recognition accuracy on the neutral class is much higher than on the other emotion categories. Thus, we want to find out whether *CorrNet* also has this problem by answering research question 1.2:

**RQ 1.2:** *Does correlation-based instance learning provide imbalanced prediction accuracies for different emotion categories?*

According to the study in Chapter 3, there is an accuracy imbalance among different classes for 3-class classification for *CorrNet*. We found that the accuracy of the high and low class (for both arousal and valence) is much lower than the accuracy on the neutral class. Compared to the subject-independent model, subject-dependent model provides more balanced recognition results, while there is still overfitting (about 30% of samples from high and low) on the neutral class. We found this is a problem due to the data imbalance when recognizing emotions using fine-grained emotion labels: more than 50% of samples from our testing datasets belong to the neutral class. The problem of sample imbalance is challenging to overcome because users tend to annotate their default emotion as neutral for most of the time when watching videos and annotate non-neutral only for specific scenes (e.g., kissing scenes for happiness). The answer to this research question underscores the importance of carefully treating data imbalance and the problem of overfitting when designing fine-grained emotion recognition algorithms using fine-grained emotion labels for training.

Although the correlation-based instance learning methods developed in Chapter 3 can obtain accurate results, it still requires large amounts of fine-grained emotion labels to train the recognition algorithm. Collecting these labels is costly and time-consuming. To address this limitation, we explore the possibility of predicting fine-grained emotions by post-stimuli emotion labels (**RQ2**) in Chapter 4. The goal is to identify which signal segments represent the post-stimuli emotion users annotated after watching the entire video. We try to achieve this goal by answering research question 2.1:

**RQ 2.1:** *Does weakly-supervised multiple instance learning solve the problem of time ambiguity if only post-stimuli emotion labels are available for training?*

To address this research question, we evaluate the regression and 3-class (high/neutral/low) classification performance of *EDMIL* (Emotion recognition based on Deep Multiple Instance Learning) proposed in Chapter 4. *EDMIL* solves the problem of time ambiguity by learning the probability for signals segments to predict the corresponding post-stimuli emotion labels. We found that *EDMIL* can provide temporally more accurate predictions compared with both the state-of-the-art weakly-supervised and fully-supervised learning algorithms. Specifically, *EDMIL* obtains > 65% accuracy and >0.65 F1 scores on all our testing datasets for three-class classification. Our studies also show that even there is only a small percentage (e.g., <30%) of the emotions elicited along a video that corresponds to the post-stimuli emotion (i.e., the level of time ambiguity is high), *EDMIL* can still obtain good recognition accuracy (>60%) for three-class classification.

Since the end-to-end feature learning structure we use for *EDMIL* may result in the problem of overfitting according to our experiments on fully-supervised learning methods, we want to find out whether this is also a problem for weakly-supervised learning methods. We do so by answering the research question 2.2:

**RQ 2.2:** *Which feature extraction methods are suitable for emotion recognition trained with only post-stimuli emotion labels?*

The performance comparison of three feature extraction methods (end-to-end, unsupervised learning and manual feature extraction) shows that using an end-to-end feature extraction method does not result in overfitting in the classification task (the accuracy on the training and testing set is similar +3.47%). The end-to-end model also provides the highest recognition accuracy compared to the other two methods. However, using unsupervised and manually extracted features can result in better regression results because the end-to-end feature learning is designed specifically for the classification task, not the regression task. The other two feature extraction methods however, are not constrained by the classification task. Thus, they can better represent the dynamic patterns of valence and arousal changes.

At the end of Chapter 4, we compare the performance of weakly-supervised and fully-supervised methods in detail to answer the research question 2.3:

**RQ 2.3:** *What are the advantages and disadvantages of recognizing emotions using weakly-supervised training by post-stimuli labels compared to fully-supervised training with fine-grained labels?*

We compare the performance of these two kinds of methods by using both the instance-level (recognition accuracy) and bag-level (dynamic time wrapping (DTW)) evaluation metrics. Compared to fully-supervised training with fine-grained labels, the advantages and disadvantages of weakly-supervised training by post-stimuli labels are:

**Advantages:**

1. Weakly-supervised training can provide more accurate prediction at the instance level, which means it can better learn the instance-label relationship from the post-stimuli emotion label.
2. Weakly-supervised training can reduce the problem of overfitting (the accuracy on training sets is not much higher than the accuracy on testing sets). It can therefore avoid learning the contradictory information (instance-label relationship) caused by the different interoception levels across individuals.

**Disadvantages:**

1. Fully-supervised algorithms can result in better recognition results for the whole video instead of individual instances (low DTW). For the subject-independent model, the temporal relationship between input signals and emotions learned from other subjects is different than the one learned from the testing user. This causes shifts of predictions in the time domain, which makes the instance-level accuracies low but does not affect the sequence-level prediction.
2. Fully-supervised algorithms can recognize more than one emotion, except for neutral, during one video watching event. The weakly-supervised methods however, can only identify which fine-grained instances are correlated to the post-stimuli emotion the user annotated and categorize all non-annotated emotions as neutral.

Although the deep multiple instance learning algorithm developed in Chapter 4 can obtain accurate recognition results with only few annotations, it can only distinguish the post-stimuli emotion from the neutral emotion and categorize all non-annotated emotions as neutral. To overcome this limitation, in Chapter 5 we investigate the possibility of recognizing emotion at a fine granularity level by using only a small amount of annotated samples to train the recognition system (**RQ3**). We want to provide more abundant predictions (instead of only post-stimuli and neutral emotions) compared to the weakly-supervised method developed in Chapter 4. We do so by answering research questions 3.1:

**RQ 3.1:** *Do deep siamese networks enable learning useful representation from few annotated signal segments that are competitive with other few-shot learning counterparts for fine-grained emotion recognition?*

We investigate this research question by comparing the performance of the Emotion recognition algorithm based on Deep Siamese Networks (EmoDSN) with other state-of-the-art few-shot learning (FSL) methods for fine-grained emotion recognition. We found that the pair-by-pair learning structure used by *EmoDSN* can result in more accurate prediction results compared to the one-to-many learning structure used by other FSL methods. The comparison also shows that the pair-by-pair learning structure results in fewer training samples compared to the one-to-many structure. Therefore, *EmoDSN* requires less training time compared to other FSL counterparts. In general, *EmoDSN* achieves 76.04% (1D-2C valence), 76.62% (1D-2C arousal) and 57.62% (2D-5C valence) accuracy (subject-dependent model) averaged on our testing datasets by using only 5 samples from each emotion category.

Then we research the influence of the temporal moments of training samples on the performance of EmoDSN by answering the research question 3.2:

**RQ 3.2:** *Do the temporal moments for selecting training samples affect the performance of deep siamese network based fine-grained emotion recognition?*

We provide the answer to this research question by selecting training samples from the beginning, ending and changing points of emotion and comparing the recognition accuracy with different amounts of training samples. We found that the temporal moments of training samples do have an influence on the performance of *EmoDSN*: samples from the changing points of emotion and the ending moments of the video watching are better training samples when only few samples are available for building up an emotion recognition system. Our finding is in line with the *peak-end theory* [2] stating that the most salient (peak) or recent (end) moments can better represent the emotions of users while watching videos. We also found that the samples at the beginning of video watching are more imbalanced compared with the ending moments, which can result in misconvergence of the learning network and provide less accurate prediction results.

## 6.2. DISCUSSION

This thesis focuses on investigating the trade-off between fine-grained emotion recognition accuracy and the amount of labels to train the recognition model using physiological signals. We embed our research in the scenario of video watching to help the video providers to better understand the temporal dynamics of their users' emotions. To close the last chapter of the thesis, we summarize the remarks throughout the thesis in the following lessons we learned:

**The imbalance of training samples in different emotion categories.** We find this problem in Chapter 3, 4 and 5 when we train the recognition network using fully-supervised methods. Due to the imbalanced nature of users' emotion annotations (annotate as neutral by default and non-neutral only for specific scenes), it is challenging to get a balanced amount of samples in different emotion categories. The network can easily overfit on the category most of the samples belong to, which may result in a "good" accuracy when the samples are severely unbalanced (e.g., 80% of the samples are in the neutral category and the recognition accuracy is 80%). Thus, we should always check whether this happens after training a network regardless of the learning structure we use.

To avoid overfitting because of the data imbalance, we can implement data augmentation methods to achieve an equal number of samples in different emotion categories. One promising approach is using the collected data to train a generative model (e.g., Generative Adversarial Network (cGAN)) to extend the size of the data for specific emotion categories (e.g., high arousal) by artificially generating more samples. We can also decrease the number of neutral samples to achieve an equal number of training samples in different emotion categories. However, that can cause the misconvergence of the network because the number of training samples is not enough to train an accurate emotion recognition model. Solving this problem requires designing specific learning structures which can converge with a limited amount of training samples (e.g., the few-shot learning network we designed in Chapter 5).

In conclusion, the takeaway messages for future researchers are 1) we should always check whether we get imbalanced accuracies on different emotion categories after training a network and 2) it is essential to develop data augmentation or learning structures that can converge on a small amount of data to avoid overfitting resulting from data imbalance.

**Towards more accurate recognition: which machine learning models are suitable for**

**fine-grained emotion recognition?** By comparing the performance of fully-supervised, weakly-supervised, and few-shot learning methods, we find that they have different advantages and disadvantages. Fully-supervised algorithms can result in better recognition results for the whole video watching. However, this requires a large number of signals with fine-grained annotations to train the recognition system. In Chapter 4 we find that weakly-supervised training methods can provide more accurate prediction at the instance level compared to fully-supervised methods. However, their overall prediction results are worse than the fully-supervised methods and can only identify post-stimuli emotion labels. The few-shot learning method in Chapter 5 can recognize more emotion categories (better than the weakly-supervised learning) by using less training samples (better than the fully-supervised learning). However, it can only obtain accurate results by the subject-dependent model.

In conclusion, if the annotation quantity is enough, fully-supervised learning methods can provide more accurate recognition results. If only a limited amount of annotations is available for training, weakly-supervised learning methods are more suitable for building up subject-independent models which identify one emotion (e.g., the post-stimuli emotion) from the baselines. Few-shot learning methods are more suitable for building a personalized recognition model for each user, which can predict multiple emotions (except for neutral and the post-stimuli emotion in Chapter 4) for users watching one video.

**The temporal dynamics of physiological changes for emotions.** In this thesis, we did not implement strong constraints to filter out the instances which have a high probability of predicting the emotions, but are not densely aggregated in one temporal moment (which is widely used for image-based emotion recognition [3]). In Chapter 3 and 5, we consider instances as independent samples and extract features from each of them respectively. In Chapter 4, we use a simple threshold based on the distribution of the instance gains for identifying which instances correspond to the post-stimuli emotion labels. We find that the recognition accuracy is higher compared to implementing strong constraints (e.g., the recurrent structure and linear iterative clustering) on different instances. We therefore conclude that the dynamics of physiological changes related to emotions are sparsely distributed in the temporal space: users can give a response of physiological changes of a short duration. If we implement strong constraints between instances, some of the responses of short-lasting physiological changes could be omitted, resulting in lower recognition accuracy. This finding also allows us to understand that the length of the instance cannot be too long: longer instances may lose salient information related to the local physiological response. However, short instance length can entail insufficient information for accurate classification.

Therefore, in Chapter 3 and 4, we run experiments by testing the proposed algorithms using different instance lengths. In conclusion, the results show that the instance length between 1s-2s can result in a good performance for the algorithms proposed in this thesis, which can serve as an appropriate length to classify emotions using fine-grained emotion labels.

**Information sparsity of physiological signals for emotion recognition.** Compared to machine learning tasks using other temporal signals (e.g., video, speech, EEG signals), the physiological signals we use contain less abundant information for emotion recognition (information sparsity of physiological signals) [4, 5]. We find this problem in the baseline comparison in Chapter 3, 4 and 5: deepening the networks (keep increasing the number of convolution layers) can easily cause them to overfit quickly on the training data. Here we can speculate that the information inside each instance is limited and insufficient to train a deep discriminative model. This limits the performance of deep neural networks: feature extraction layers with deep or sophisticated structures can easily overfit or fail to extract meaningful features for recognition.

The information sparsity of physiological signals makes it challenging for neural networks to predict users' emotions by simply deepening the network. In Chapter 3 and 4, we find that shallow feature extraction structures (e.g., < 5 convolutional layers with a gradually increased number of filters and decreased size of kernels) can result in a better recognition accuracy and less overfitting.

In conclusion, the take-away message for future researchers is: 1) shallow feature extraction structures are more suitable for fine-grained emotion recognition using physiological signal because of the information sparsity of physiological signals, and 2) when the networks cannot accurately predict the emotions, we should reconsider the learning paradigm (e.g., weakly-supervised instead of fully-supervised learning, Chapter 4) and learning structure (e.g., pair-by-pair instead of one-to-many, Chapter 5) instead of simply deepening the neural networks.

## 6.3. LIMITATIONS AND FUTURE WORK

The 'data hunger' of machine learning algorithms requires a large amount of fine-grained emotion annotations to train an accurate emotion recognition model. This thesis focuses on addressing this challenge by investigating the trade-off between fine-grained emotion recognition accuracy and the amount of annotations used for training. Given the challenges of predicting valence and arousal labels at a fine level of granularity using only limited amount of annotations, there are natural limitations to our work.

Due to the high inter-subject variability of physiological signals, the performance of the subject-independent models (in Chapter 3 and 4) trained with limited amount of data can still be improved. One possible solution is to provide more annotated samples to fulfill the data hunger of machine learning algorithms. Data augmentation methods can be used to generate artificial signals which obey the distribution of specific emotion categories. Then a 'hybrid' of synthetic and real signals can be used to train the machine learning based emotion recognition algorithms. In our preliminary study [6], we provide initial steps towards further investigating how generative models could be applied to diverse physiological signals for the task of physiology-driven emotion recognition. We use an Auxiliary Conditioned Wasserstein Generative Adversarial Network with Gradient Penalty (AC-WGAN-GP) to generate synthetic data and compare the recognition performance between real and synthetic signals as training data in the task of binary arousal classification. Experiments on the CASE dataset using GSR and ECG signals show that generative data augmentation significantly improves model performance (avg. 16.5%) for binary arousal classification in a subject-independent setting. The synthetic data generated by our method significantly improves the classification performance in subject-independent testing by providing more balanced classification results for different arousal categories. Our study presents the initial steps towards further investigating how generative models can be applied to diverse physiological signals for fine-grained emotion recognition. In the future, others may plan to extend our model for other physiological signals, such as EEG, EMG, eye-movement, respiration (RSP) and skin temperature. Others may also want to explore more generative models, such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Naive Bayes Model to find out which data augmentation methods are the most suitable ones across different physiological signals.

All the algorithms designed in this thesis require discretization of continuous labels for fine-grained recognition since they are designed specifically for classification instead of regression. Although we test *EDMIL* in Chapter 4 on the regression task, its regression performance is not as good as classification. Compared to classification methods, the regression methods can present continuous changes of emotions instead of discrete emotion labels for each time segment. Therefore, regression is temporally more precise. In the future, others may want to extend the algorithms designed in this thesis for regression tasks. Specifically, we will extend the *EDMIL* and *EmoDSN* into a multi-instance multi-label formulation [7] and a few-shot regression [8] algorithm, respectively, to obtain a continuous output for emotion recognition.

In this thesis, we only consider physiological signals (i.e., EDA, BVP, Heart Rate and Skin Temperature) as input modalities. Previous works also show that other modalities, such as eye-tracking data from electrooculography (EOG) sensors and Electroencephalography (EEG), also contain abundant information for emotion recognition [9]. The application scenario of our work is to analyze the personalized experience (emotions) of users while watching videos. Thus, the semantic features (e.g., audio-visual content, text captions and speech transcripts of the content) from video stimuli can also help to improve the recognition results by providing the context information for recognition. The text-derived fingerprints are more accessible to generate this context information compared to video data because of their low-cost computation [10]. In the future, others can extend the algorithm designed in this thesis for the data in other modalities and investigate whether the recognition accuracy can be further improved. For example, others can build a weak constraint (e.g., using Canonical Correlation Analysis (CCA) (Chapter 3) ) between the emotion semantic features derived from the text of videos (e.g., speech transcripts, emotion tags of videos) to promote the recognition accuracy.

At last, we only focus on the physiological signals with emotions annotations and ignore the large amount of unlabeled signals. Although reliably annotating a large amount physiological signals is costly and time-consuming, collecting massive amounts of physiological measurements from sensors is a trivial task [11]. Therefore, it is worthwhile for other researchers to explore the possibility of recognizing emotions by learning the task-unrelated features from a large amount of unlabeled signals and fine-tune the models for emotion recognition with a small amount of annotated signals. To achieve this target, one possible solution is to use self-supervised learning. Self-supervised learning is a kind of machine learning method that can automatically learn embeddings and features from unlabeled samples. Unlike supervised learning which learns task-specific features, self-supervised learning learns the task-unrelated features according to the inner structure of the data (e.g., correlation). For example, Sarkar et al. [12] designed a self-supervised learning network that first learns ECG representations by signal transformation recognition tasks and then uses the learned task-unrelated features for emotion recognition. Although previous work provides some useful insights on using self-supervised learning for emotion recognition, there are still challenges in designing good (learned features can provide accurate recognition with limited number of fine-tuning samples) and universal (generalized among different input modalities) pre-training tasks to enable the self-supervised learning networks to learn these task-unrelated features [13]. In the future, others can have the chance to explore this possibility by exploring which pre-training tasks are the most suitable ones for fine-grained emotion recognition.

To conclude, we expect future researchers in the field of affective computing to focus more on the problem of machine learning by small data. Collecting a large amount of annotated data for emotion recognition is costly and time-consuming which is a large extra burden and investment for any potential stakeholder (e.g., video providers). Future researchers can focus on developing data-augmentation methods (e.g, generative models), which can generate more annotated samples for training the recognition algorithm, or design machine learning methods (e.g., self-supervised learning) which can be trained by a small amount of annotated signals.

# REFERENCES

[1] M. Schmitt, N. Cummins, and B. Schuller, *Continuous emotion recognition in speech: do we need recurrence?* (2019).

[2] B. L. Fredrickson and D. Kahneman, *Duration neglect in retrospective evaluations of affective episodes.* Journal of personality and social psychology **65**, 45 (1993).

[3] Y. Tian, W. Hao, D. Jin, G. Chen, and A. Zou, *A review of latest multi-instance learning,* in *2020 4th International Conference on Computer Science and Artificial Intelligence* (2020) pp. 41–45.

[4] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, *Multiple instance learning for emotion recognition using physiological signals,* IEEE Transactions on Affective Computing (2019).

[5] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, *Multi-task and multi-view learning of user state,* Neurocomputing **139**, 97 (2014).

[6] A. Furdui, T. Zhang, M. Worring, P. Cesar, and A. El Ali, *Ac-wgan-gp: Augmenting ecg and gsr signals using conditional generative models for arousal classification,* in *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers* (2021) pp. 21–22.

[7] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, *Multi-instance multi-label learning,* Artificial Intelligence **176**, 2291 (2012).

[8] Y. Loo, S. K. Lim, G. Roig, and N.-M. Cheung, *Few-shot regression via learned basis functions,* (2019).

[9] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, *A review of emotion recognition using physiological signals,* Sensors **18**, 2074 (2018).

[10] Z. Erenel, O. R. Adegboye, and H. Kusetogullari, *A new feature selection scheme for emotion recognition from text,* Applied Sciences **10**, 5351 (2020).

[11] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, *A dataset of continuous affect annotations and physiological signals for emotion analysis,* Scientific data **6**, 1 (2019).

[12] P. Sarkar and A. Etemad, *Self-supervised ecg representation learning for emotion recognition,* IEEE Transactions on Affective Computing (2020).

[13] H. H. Mao, *A survey on self-supervised pre-training for sequential transfer learning in neural networks,* arXiv preprint:2007.00800 (2020).

# Acknowledgements

The 4-year voyage for pursuing the Ph.D. degree is challenging. I could not finish this journey without the company of wonderful people. I want to say appreciation here for those who made this journey possible.

First, I would like to thank my promotor and daily supervisor in CWI, Pablo Cesar and Abdallah El Ali. Thank you so much for your steering and supervision! Pablo, thank you so much for allowing me to be part of this amazing group of researchers! I really appreciate the great amount of effort you made to help me develop the ability of scientific independence. Thank you so much for pushing and motivating me. I remember coming into the meetings thinking ”I cannot find any post-doc positions after I graduate” and coming out thinking ”Of course I can find post-doc positions, maybe assistant professor directly!”. You always believed in me and pushed me to my best. Abdo, thank you so much for all the help you gave me in all these years. You taught me a great deal about HCI and affective computing, and most importantly, to care about not only the accuracies of algorithms but also how they can benefit the scientific community. We drank when our paper gets accepted, we drank double when our paper gets rejected, and we drank triple when nothing happens. I am thankful for what you have done as a supervisor and as a friend, for your patience towards my endless “quick questions”, for supporting, for disagreeing, and for many enjoyable smokings on the balcony.

Alan, you were also the one who allowed me to start my Ph.D. and guided me to finish. Thank you for all discussions and advice we have had for our work. Your insights always helped me how to efficiently communicate the research to others. I also appreciate the constructive and long editorial work you have helped with my thesis.

I also want to thank the people from the DIS group: Jack Jansen, Thomas Röggla, Irene Viola, Silvia Rossi, Alina Striner, Sueyoon Lee, Xuemei Zhou, and Nacho Reimat. These last 4 years in DIS were wonderful, and I will bring the memories with me forever. I would like to thank Shishir Subramanyam who also works as a Ph.D. candidate in the DIS group. We always share our experiences and options about our work and happy moments in our life. You (and Abdo) taught me a lot of English words which I will never learn from the English classes in China. Most importantly, I understand that I still owe you a lot of cigarettes but we should all quit smoking after we finish our Ph.D.

I want to give my greatest thank to my parents who always support me during these 4 years. My father believes getting a Ph.D. is necessary for all human beings while my mother believes I can do whatever I want if I am happy. Their support, although manifested from different perspectives, helped me get through my toughest times in the Netherlands. I will always love you, just as you always loved me.

在此时，我要特别感谢我的奶奶。虽然不了解我的工作，但是奶奶一直无条件的支持我，坚信我能做到最好。即使相隔万里，我也能感受到您对我的牵挂与关心。离家多年，我一直最记挂的就是独自在老家生活的奶奶，希望您可以一直健康平安，笑口常开。我还要感谢住在和曾经住在 Carolina MacGillavrylaan 的小伙伴们，是你们的陪伴让我度过了四年快乐的时光。我也要特别感谢一直陪伴支持我的各位博士，包括但不限于刘洋博士，王祯博士，郭若琛博士，郭宝烘博士，王嘉楠博士，聂忠辉博士，陈星宇博士，是你们的陪伴让我有了继续下去的希望与动力。我也要感谢新华网给予我博士项目的资助，感谢杨溟院长和王晨副院长对我学术上的指导和工作上的关心。

子在川上曰：逝者如斯夫，不舍昼夜。这是我写在硕士毕业论文最后的话。如今我将博士毕业，往者不可谏，来者犹可追，希望在以后的学术道路上我仍有动力、有信心、有能力去追求理想。最后，值此毕业之际，我又将开启人生的新篇章，我想再念一遍来时念过的两句诗：“苟利国家生死以，岂因祸福避趋之”。

# Curriculum Vitæ

## Tianyi Zhang

09-03-1993 Born in Nanjing, China.

## Education

2011–2015
**B.S.**, *Electrical Engineering and Automation*
Nanjing University of Aeronautics and Astronautics, China

2015–2018
**M.S.**, *Control Engineering*
Nanjing University of Aeronautics and Astronautics, China
Promotor: Prof. dr. Z. Yang
Thesis: Obstacle avoidance for mobile robot based on stereo vision

2018-2022
**Ph.D.**, *Computer Science*
Delft University of Technology, the Netherlands
Promotor: Prof. dr. P. Cesar
Promotor: Prof. dr. A. Hanjalic
Daily Supervisor: Dr. A. El Ali
Thesis: On fine-grained temporal emotion recognition in video
How to trade off recognition accuracy with annotation complexity?

## Experience

2017
Research Assistant
AE2 Department
KOSTAL Asia R&D Center
Shanghai, China

2018
Research Assistant
Future Media Convergence Institute
Xinhuanet
Beijing, China

# List of Publications

## Main publications for the thesis

1. **Zhang, Tianyi**, Abdallah El Ali, Alan Hanjalic, and Pablo Cesar. "Few-shot Learning for Fine-grained Emotion Recognition using Physiological Signals." *IEEE Transactions on Multimedia* (2022).

2. **Zhang, Tianyi**, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. "Weakly-supervised Learning for Fine-grained Emotion Recognition using Physiological Signals." *IEEE Transactions on Affective Computing* (2022).

3. **Zhang, Tianyi**, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. "Corrnet: Fine-grained emotion recognition for video watching using wearable physiological sensors." *Sensors* 21, no. 1 (2020): 52.

4. **Zhang, Tianyi**, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. "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*, pp. 1-15. 2020.

5. **Zhang, Tianyi**, Abdallah El Ali, Chen Wang, Xintong Zhu, and Pablo Cesar. "CorrFeat: correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition." In 2019 *International Conference on Multimodal Interaction*, pp. 404-408. 2019.

6. **Zhang, Tianyi**. "Multi-modal Fusion Methods for Robust Emotion Recognition using Body-worn Physiological Sensors in Mobile Environments." In 2019 *International Conference on Multimodal Interaction*, pp. 463-467. 2019.

## Other publications

1. Xue, Tong, Abdallah El Ali, **Tianyi Zhang**, Gangyi Ding, and Pablo Cesar. "CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos." *IEEE Transactions on Multimedia* (2021).

2. Furdui, Andrei, **Tianyi Zhang**, Marcel Worring, Pablo Cesar, and Abdallah El Ali. "AC-WGAN-GP: Augmenting ECG and GSR Signals using Conditional Generative Models for Arousal Classification." In Proceedings of the 2021 *ACM International Conference on Pervasive and Ubiquitous Computing*, pp. 21-22. 2021.

3. Xue, Tong, Abdallah El Ali, **Tianyi Zhang**, Gangyi Ding, and Pablo Cesar. "RCEA-360VR: Real-time, Continuous Emotion Annotation in 360 VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels." In Proceedings of the 2021 *CHI Conference on Human Factors in Computing Systems*, pp. 1-15. 2021.