Multi-modal Fusion Methods for Robust Emotion Recognition using Body-worn Physiological Sensors in Mobile Environments

Tianyi Zhang
Centrum Wiskunde & Informatica
Delft University of Technology
Amsterdam, The Netherlands
tianyi.zhang@cwi.nl

ABSTRACT

High-accuracy physiological emotion recognition typically requires participants to wear or attach obtrusive sensors (e.g., Electroencephalograph). To achieve precise emotion recognition using only wearable body-worn physiological sensors, my doctoral work focuses on researching and developing a robust sensor fusion system among different physiological sensors. Developing such fusion system has three problems: 1) how to pre-process signals with different temporal characteristics and noise models, 2) how to train the fusion system with limited labeled data and 3) how to fuse multiple signals with inaccurate and inexact ground truth. To overcome these challenges, I plan to explore semi-supervised, weakly supervised and unsupervised machine learning methods to obtain precise emotion recognition in mobile environments. By developing such techniques, we can measure the user engagement with larger amounts of participants and apply the emotion recognition techniques in a variety of scenarios such as mobile video watching and online education.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Emotion recognition; Physiological sensors; Mobile Environments; Machine learning

ACM Reference Format:
Tianyi Zhang. 2019. Multi-modal Fusion Methods for Robust Emotion Recognition using Body-worn Physiological Sensors in Mobile Environments. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340555.3356089

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10.
https://doi.org/10.1145/3340555.3356089

1 INTRODUCTION

My PhD thesis aims to build a robust sensor fusion system for emotion recognition in mobile environments. The research on emotion recognition using physiological signals focuses on analyzing the physiological behavior of the human neural systems [20], in which emotion changes according to Cannon's theory [7]. According to Cannon's theory, we feel emotions and experience physiological reactions such as sweating, trembling, and muscle tension simultaneously. Unlike the facial expressions, physiological behaviors in human neural systems are involuntarily activated and therefore cannot be easily controlled. Thus, it is more objective to measure the emotion of people through physiological sensors [21].

An important step for emotion recognition, using physiological signals, is multi-modal fusion. Mobile environments are uncontrolled environments that users' mobility is not constrained by any devices. [19]. Thus, this kind of environments present a challenge for multi-modal fusion because the signals have noise and different physiological signals have different temporal characteristics [15]. In addition, obtaining large amounts of data and accurately labeling them are also big challenges in mobile environments. While previous work has combined body-worn sensors (e.g., eye tracking, GSR) to address the low accuracy rates for recognizing emotion states (e.g., arousal, valence and dominance [18]), more work is needed for addressing some of the existing challenges of multi-modal fusion algorithms.

There are three kinds of methods to fuse signals from different sensors: data level fusion, feature level fusion and decision level fusion, each depending on when information from the different sensors is combined [10]. Traditionally, physiological signals are often fused in decision level using classifiers such as Support Vector Machine (SVM) and K Nearest Neighbor (KNN) [6, 20]. However, decision level fusion depends highly on the precision of ground truth, which

is difficult to get in mobile environments. Additionally, fusion methods in decision level cannot generate new features from multi-modal signals. That means if the selected features are not appropriate for classification, the system will be inaccurate no matter how well-designed the fusion method is. To solve these problems, we propose to explore fusion methods in data and feature level and use weakly supervised or unsupervised machine learning methods to fuse the physiological signals for robust emotion recognition in mobile environments.

The proposed model will be based on principles of what neural and psychological activities these signals reflect. Thus, we will develop machine learning algorithms which are interpetable by psychological theories. The system will both consider the signals from the Peripheral Nervous System (PNS), such as Electrodermal activity (EDA), and from the Central Nervous System (CNS), such as pupil dilation and saccade amplitude. We believe the weakly supervised or unsupervised machine learning models, such as Generative Adversarial Networks (GAN) [13] and Deep Canonical Correlation Analysis (DCCA) [1], will ensure the robustness of the fusion method towards noise and inaccurate ground truth caused by mobile environments. Unlike supervised machine learning models which fuse different signals according to the ground truth labels, unsupervised machine learning models fuse multi-modal signals according to the relationship between elements of data (i.e., correlation, energy or entropy of data) [25, 26]. Thus, we believe that it will make the model robust to inaccurate ground truth labels. The final goal of this research is to build a robust sensor fusion system based on psychological models, which could increase the accuracy of recognizing emotional status such as arousal, valence and dominance in mobile environments.

2 TECHNICAL PROBLEMS

Generally, there are three major technical problems in the thesis:

P1. Pre-processing signals across physiological sensors
The signals from different physiological sensors have different characteristics. That includes **different temporal characteristics (P1.1)** and **different error models to the noise (P1.2)**.

Different sensors have different sampling rates and different response time for emotions. For example, both pupil diameter [4] and skin conductance response (SCR) [3] contain information to predict the arousal of the Autonomic Nervous System (ANS). However, the change of pupil diameter occurs 200ms [23] after the stimulus, while SCR takes 1-2 seconds [14]. Such asynchronicity makes it challenging to fuse multi-modal sensor data to extract joint psychological features from the raw signals.

In addition, mobile environments can result in noise and sparsity in the raw data collected by physiological sensors [16]. Since different sensors have different hardware structures and error models, it is a challenge to design filters or neural networks to erase noise from the signals. The movement of the subjects will also result in some unpredictable noise. For example, the electrodes of GSR sensors could suddenly detach from the skin because of the movement of subjects' hands. Such uncertainty makes it more challenging to erase noise from physiological signals.

P2. Limited amount of data
A fusion system based on machine learning models requires a large amount of data for training [9, 24]. However, it is challenging to **collect large amounts of data (P2.1)** and **label them (P2.2)**.

It is costly to collect physiological sensor data since we need to recruit users for experiments. In addition, it is also difficult to equip a large number of users with multiple physiological sensors. Thus, the challenge is how to automatically augment data with suitable artificial samples when the amount of data is limited.

Labeling a large amount of data is very costly and time-consuming since we need users to reflect on and label their own emotional states (e.g., through self-report questionnaires). Thus, the challenge is how to train the fusion system with a small amount of labeled data and a large amount of unlabeled data.

P3. Labeling ground truth data
A fusion system based on machine learning methods requires precise labels for training. However, the labels collected in mobile environments may not be precise enough, be misaligned temporally to the actual state at which they were experienced, or be altogether inaccurate.

Inexact labelling (P3.1) is common and sometimes inevitable for emotion recognition. For example, signals with a duration of one hour can be labelled as happy according to the self report of subjects. However, it does not mean that the subjects feel happy all the time during the entire process of the one-hour experiment. If we train the network with inexact labels, the fusion system will easily over fit. [25].

In mobile environments, some of the signals could be **inaccurately labelled (P3.2)**. This is due to the high variance and even inaccurate self-reports when users label their own past emotional states. Classic supervised machine learning methods need precise labels to build discriminative models. Thus, if the ground truth labels are not precise, most widely-used supervised machine learning methods will have problems such as mis-convergence and over-fitting [12].

3 RESEARCH QUESTIONS

To solve these technical problems, we set out to answer the following research questions. Each research question is asked to address the corresponding problem.

Research question 1: *Can machine learning methods automatically pre-process the signals across multiple physiological sensors?*

The signals from different physiological sensors have different characteristics. Filters and alignment techniques are used in traditional methods for the pre-processing of physiological signals across sensors [2, 8, 17]. However, these methods need to be designed manually and are inefficient to unpredictable noise. One possible solution is to use machine learning methods to automatically pre-process signals across sensors.

RQ 1.1: *Which learning techniques are most suitable for fusing the asynchronous physiological signals to increase the accuracy of mobile emotion recognition?*

To answer this question, we propose to design deep learning networks to automatically extract features from the raw data of different psychological signals with different temporal characteristics. By comparing it with methods which need manual feature extraction and synchronization, we will answer the question whether or not the deep structure could also have excellent performance on physiological signals.

RQ 1.2: *Can unsupervised machine learning methods eliminate the effect of signal noise that is an outcome of mobile environments?*

Classic supervised machine learning methods need precise raw signals to build discriminative models. Thus, if the raw signals are not precise, most widely-used supervised machine learning methods will have problems such as misconvergence and over-fitting. That is why we propose to use unsupervised methods such as Restricted Boltzmann Machine (RBM) and DCCA to extract features considering the probability distribution of the data set.

Research question 2: *Do semi-supervised and unsupervised learning methods benefit the training process of mobile emotion recognition across sensors when the amount of data are limited?*

The amount of data for emotion recognition is comparatively small due to the difficulty of equipping users with multiple physiological signals. That could cause problems such as mis-convergence and over-fitting when the fusion system is trained. Thus, it remains a challenge how to develop machine learning algorithms for mobile emotion recognition with limited number of data.

RQ 2.1: *How to adapt data augmentation methods that are suitable across different physiological signals when the amount of sample data is limited?*

One possible solution is using the collected data to train a generative model (e.g., GAN), and then extend the size of data set by artificially generating more samples by this model. Many generative models such as Generative Adversarial Network (GAN), Hidden Markov Model (HMM), Gaussian Mixed Model (GMM) and Naive Bayes Model can be used to generate probability distributions for physiological signals. It is essential to adapt appropriate generative models to be suitable for one or more physiological signals.

RQ 2.2: *Can semi-supervised learning methods benefit the training processing of the fusion system when only a small amount of data are labelled?*

Labeling a large amount of physiological signals is very costly and time-consuming. However, it is difficult to train a fusion system without precise ground truth. One possible solution is that we only label a small amount of signals and train the fusion system with both labeled and unlabeled signals with semi-supervised learning methods.

Research question 3: *Do weakly supervised machine learning methods increase the accuracy of mobile emotion recognition across sensors when some of the ground truth labels are inexact or inaccurate?*

Collecting accurate ground truth labels is quite difficult for emotion recognition in mobile environments. Weakly supervised learning methods take potential inaccuracy of ground truth into consideration during the training process, which makes them robust to inexact and inaccurate labels.

RQ 3.1: *Can multi-instance learning methods increase the accuracy of emotion recognition when the signals from different sensors are labeled in an inexact way?*

Inexact labelling is common and sometimes inevitable for emotion recognition. If we train the network with inexact labels, the fusion system will easily over fit. Multi-instance (MI) learning is a variant of inductive machine learning, where each learning example contains a bag of instances instead of a single feature vector [11]. For sensor-based emotion recognition, the entire signals and their segments can be viewed as bags and instances respectively. The methods have been successfully implemented in the fields for image and voice recognition. Thus, it is worthwhile finding out whether they can benefit the fusion system for mobile emotion recognition as well.

RQ 3.2: *Which mathematical principle is suitable for identifying potentially inaccurate labels in signals across multiple physiological senors?*

In mobile environments, some of the signals could be inaccurate labelled due to subjective self-reports when users label their own past emotional states. In practice, a basic idea to solve this problem is to identify the potentially mislabeled examples [5], and then try to make some correction [26]. There are several mathematical principles in previous

research, such as minimax entropy principle [25], that attempt to infer ground-truth labels from the data. Thus, it is essential to find out which principle is the most appropriate one to identify the potentially mislabelled data for emotion recognition.

4 RESULTS TO DATE

To answer the research questions, we have conducted one experiment to collect user affect data in mobile settings. In addition, we have developed a feature extraction algorithm to extract joint features from skin conductance response and pupil diameters for the prediction of valence and arousal.

1) Mobile user affect data collection experiment

The purpose of this experiment is to collect ecologically valid user affect data in mobile environments, which can be used to train emotion recognition algorithms to robustly classify user affective states in mobile environments. The experiment was conducted in a mobile setting. The participants was told to walk freely or stand as their wish during the experiment. They were equipped with a wearable eye-tracker to capture their eye movement and a Empatica E4 wristband. The participants were also asked to annotate their valence and arousal in real-time using an mobile application developed by us. All the signals including the real-time annotation by users were synchronized by a NTP sever.

From this experiment, we got the eye movement and physiological signals in a mobile environment. The data collected in this experiment has two features: 1) we collected the real-time emotion annotations which are synchronized with physiological signals. Compared with the "inexact labels" described in section 2 (P3.1), the continuous annotation could help us to analyze emotions for each segmentation of the videos, which could promote the accuracy of emotion recognition in time domain. 2) previous experiments on affect data collection are mostly conducted in a static and desktop environment. Our experiment was conducted in an uncontrolled environment where the participants can move freely without any constrain, which is more similar to the application scenarios such as watching a video when waiting for a train, on a bus or on a metro. **The design of experiment and annotation method will be submitted to CHI 2020.**

2) Correlation-based feature extraction algorithm

To overcome the challenge of different temporal characteristics and noise model among signals, we designed a feature extraction algorithm that maximizes the correlation coefficient of pupil diameter and skin conductance responses for participants watching the same video clip. To boost performance given limited data, we implement an incremental learning system without a deep architecture to classify emotions along the arousal and valence dimensions. We test our method on the MAHNOB-HCI [22] database, and achieve accuracies of 82.9% and 82.1% for arousal and valence, respectively. Our method outperforms not only state-of-art approaches, but also widely-used traditional machine learning and deep learning methods. **The details of the algorithm and the testing results were submitted to ICMI 2019.**

5 FUTURE PLAN

In the future, we will further explore unsupervised, semi-supervised and weakly supervised models to answer the research questions. We will conduct more experiments in different scenarios to validate the robustness of our algorithms. Generally, our future plan can be summarized into three stages:

1) **Adapt and validate the fusion algorithms on the data we collected in mobile environments.**

In this stage, we will compare the testing result for our feature extraction algorithm between the static environment (data from MAHNOB-HCI [22] database) and the mobile environment (data we collected). Moreover, we will develop a regression model to predict the valence and arousal using the continuous emotion annotation we got from mobile user affect data collection experiment. The analysis will try to explore which parameters in the fusion system will improve or decrease the accuracy of emotion recognition in different environments. This will help us answer **RQ 1.1** and **RQ1.2**.

2) **Semi-supervised learning algorithms to fuse the sensor signals when only partial of the samples are labeled.**

In this stage, we will conduct an experiment which only partial of the samples are labeled and try to develop a semi-supervised learning algorithms for emotion recognition. This will enable us to collect large amounts of unlabelled data for training, which will low the cost and simplify the process of affect data collection. This will help us answer **RQ 2.1** and **RQ 2.2**.

3) **Multi-instance learning algorithms for long duration emotion recognition.**

In this stage, we will conduct an experiment which users will wear physiological sensors for a long duration such as one day or one week. The data collected in this experiment can only have inexact ground truth from users' self-report. Thus, we plan to design multi-instance learning algorithms to solve this problem. This will help us to answer **RQ 3.1**.

Since the users will be equipped with physiological sensors for a long duration, his or her self-report on what emotion he or she feels may be inaccurate. We will develop an algorithm to detect this inaccuracy and base on that, we will be able to answer **RQ3.2**.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. 1247–1255.

[2] GJ Boyle, E Helmes, G Matthews, and CE Izard. 2015. Multidimesnional measures of af-fects: Emotions and mood states.

[3] MM Bradley, MK Greenwald, and AO Hamm. 1993. The Structure of Emotion: PsychoPhysiological, Cognitive and Clinical Aspects. (1993).

[4] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 4 (2008), 602–607.

[5] Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research* 11 (1999), 131–167.

[6] Rafael A Calvo and Sidney D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1, 1 (Jan. 2010), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

[7] Walter B Cannon. 1931. Again the James-Lange and the thalamic theories of emotion. *Psychological Review* 38, 4 (1931), 281.

[8] Hongtian Chen and Bin Jiang. 2019. A review of fault detection and diagnosis for the traction system in high-speed trains. *IEEE Transactions on Intelligent Transportation Systems* (2019).

[9] Hongtian Chen, Bin Jiang, Tianyi Zhang, and Ningyun Lu. 2019. Data-driven and deep learning-based detection and diagnosis of incipient faults with application to electrical traction systems. *Neurocomputing* (2019).

[10] Rana El Kaliouby and Peter Robinson. 2005. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*. Springer, 181–200.

[11] James Foulds and Eibe Frank. 2010. A review of multi-instance learning assumptions. *The Knowledge Engineering Review* 25, 1 (2010), 1–25.

[12] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[14] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

[15] M Shamim Hossain and Ghulam Muhammad. 2017. An emotion recognition system for mobile applications. *IEEE Access* 5 (2017), 2281–2287.

[16] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 493–504.

[17] Yisi Liu and Olga Sourina. 2014. Real-time subject-dependent EEG-based emotion recognition algorithm. In *Transactions on Computational Science XXIII*. Springer, 199–223.

[18] Albert Mehrabian. 1997. Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression. *Journal of psychopathology and behavioral assessment* 19, 4 (1997), 331–357.

[19] S Nagamani and DR Nagaraju. 2018. A Mobile Cloud-Based Approach for Secure M-Health Prediction Application. *International Journal for Innovative Engineering & Management Research* 7, 12 (2018).

[20] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A Review of Emotion Recognition Using Physiological Signals. *Sensors* 18, 7 (2018), 2074.

[21] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A Review of Emotion Recognition Using Physiological Signals. *Sensors* 18, 7 (June 2018), 2074. https://doi.org/10.3390/s18072074

[22] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55.

[23] Henk Van Steenbergen and Guido PH Band. 2013. Pupil dilation in the Simon task as a marker of conflict processing. *Frontiers in human neuroscience* 7 (2013), 215.

[24] Tianyi Zhang and Olivier Le Meur. 2018. How Old Do You Look? Inferring Your Age From Your Gaze. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2660–2664.

[25] Zhi-Hua Zhou. 2012. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

[26] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (2017), 44–53.