U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

INESCTEC

Inside Out: Fusing ECG and Face Information to Recognise Emotions

Mariana Silva Xavier

Mestrado em Bioengenharia

Supervisor: Jaime dos Santos Cardoso, PhD

Co-Supervisor: João Tiago Ribeiro Pinto

July 28, 2022

© Mariana Silva Xavier, 2022

Inside Out: Fusing ECG and Face Information to Recognise Emotions

Mariana Silva Xavier

Mestrado em Bioengenharia

Approved in public examination by the Jury:

President: Miguel Velhote Correia, PhD

Referee: João Carlos Monteiro, PhD

Referee: Jaime dos Santos Cardoso, PhD

July 28, 2022

Resumo

O reconhecimento de emoções tem atraído um grande interesse em diversas áreas de aplicação, tendo várias metodologias sido estudadas e propostas ao longo dos anos. O reconhecimento de emoções através da face tem sido a modalidade mais explorada, partindo da capacidade de uma pessoa facilmente reconhecer o estado emocional de outra através das suas mudanças de expressão. No entanto, é bastante fácil para alguém esconder o seu verdadeiro estado emocional, o que se traduz numa redução da fiabilidade destas abordagens. Pelo contrário, sinais fisiológicos têm surgido como alternativa devido à sua relação com o sistema nervoso. Este é ativado de forma involuntária, tornando esta uma modalidade mais fiável, visto que as pessoas não conseguem controlar como as suas verdadeiras emoções se traduzem nos diversos sinais. O uso do ECG surge devido à variação do batimento cardíaco conforme os diferentes estados emocionais.

Estudos mais recentes consideram a fusão de diferentes modalidades, trazendo a vantagem de recolher informação complementar destas, o que leva a uma maior robustez dos modelos. O trabalho realizado ao longo desta dissertação teve como objetivo explorar o uso do sinal ECG juntamente com informação facial de forma a permitir o reconhecimento de emoções, tentando minimizar os desafios de cada uma das modalidades. Assim, metodologias individuais foram desenvolvidas e comparadas com a fusão de ambas.

Ao replicar uma abordagem presente na lituratura de forma a obter um ponto de partida para o trabalho, alguns desafios foram identificados e tidos em conta para futuras melhorias. Estes incluem: a extração automática de características a partir de modelos de *deep learning*, pré-treinados; a adição de LSTMs de forma a explorar a informação temporal presente nos dados e, desta forma, agregar os segmentos dos sinais ECG e as *frames* dos vídeos. Os modelos sem agregação permitiram atingir uma precisão de 47 % para os sinais ECG e de 75 % para os vídeos faciais, enquanto que, com a agregação temporal, obteve-se uma precisão de 33 % para os sinais ECG e de 22 % para os vídeos faciais. Esta última estratégia demonstrou resultados inferiores em comparação com os modelos sem agregação. No entanto, a divisão destes dados é feita de forma *signal-independent* que, apesar de ser mais realista, é também mais desafiante.

Por fim, a fusão das duas modalidades foi explorada através de três diferentes abordagens. O método de *early fusion* consistiu em usar dois LSTMs, um para cada uma das modalidades, cujos *outputs* foram concatenados e mapeados para as categorias finais através de uma camada linear. Os diferentes passos deste modelo foram treinados em conjunto e permitiram atingir uma precisão de 27 %, que não se mostrou superior aos resultados obtidos com o ECG apenas. Sendo um modelo bastante complexo, mais dados seriam necessários para permitir a aprendizagem deste de forma eficaz. As duas estratégias de *late fusion* utilizadas foram a regra das somas e a regra dos máximos. Estas regras atuam sob as probabilidades retornadas pelos modelos individuais. Ambas as abordagens provaram serem capazes de integrar os pontos fortes de ambos os modelos, atingindo uma precisão de 58 % com a regra das somas e de 60 % com a regra dos máximos.

Abstract

Emotion recognition has attracted great interest in many research areas, and therefore numerous emotion recognition approaches have been studied and proposed throughout the years. Facial emotion recognition has been the most explored modality, following the intuition that one can easily feel others' emotions by the changes in their expression. However, among other challenges, reliability cannot be guaranteed because it is easy for people to purposely hide their real emotions. On the contrary, physiological signals are emerging as a way to recognise emotions due to their relationship with the nervous systems. These are largely involuntarily activated and people cannot control how their true emotions are reflected on these signals. The intuition behind the use of ECG signals for emotion recognition is the variation in heart rate due to emotional states.

More recent research has considered fusing multiple modalities, which brings the advantage of gathering complementary information from these, leading to increased robustness of the emotion recognition models. The work developed in this dissertation aims to explore the use of ECG signals together with face information for the task of emotion recognition, attempting to minimize the challenges of each modality. Therefore, individual methodologies were developed and then compared with the fusion of the two.

While replicating one literature approach to serve as a baseline, some challenges were identified and addressed in further improvements. These included the extraction of deep learning features from pre-trained networks and the addition of LSTMs to explore the temporal information present in the data and hereby aggregating the different segments of the signals or frames of the videos. The non-aggregated models enabled to achieve an accuracy of 47 % for ECG signals and 75 % for face videos, while with the temporal aggregation, an accuracy of 33 % was obtained for the ECG signals and of 22 % for the face videos. The temporal aggregation brought lower accuracies in comparison to the non-aggregated models. However, the data was split in a signal-independent way, which is more realistic but more challenging as well.

Finally, the two modalities were fused with three different approaches. The early fusion method consisted of using two LSTMs, one for the ECG signals and another for the face videos, whose outputs were concatenated before passing through a linear layer which mapped them into the five categories. The different steps of this model were trained altogether and achieved an accuracy of 27 %, which was not able to improve the results from the ECG modality separately. Being such a complex model, more data is needed to enable the model to learn well. Two late fusion strategies were considered: the sum rule and the maximum rule. These act on the probabilities returned from the individual models either by summing the confidences from each classifier or by taking the one which is most confident in its result, respectively. Both approaches proved to be able to take the best out of the two individual models, reaching an accuracy of 58 % with the sum rule and 60 % with the maximum rule.

Acknowledgements

Chegou ao fim mais uma etapa, e vejo-me agora a desejar poder começar de novo e repetir tudo da mesma maneira. Foram 5 anos incríveis e só me resta agradecer a todos os que se cruzaram no meu caminho e permitiram que assim fosse.

Gostava de agradecer, em primeiro lugar, aos meus orientadores. Ao Professor Jaime, pelo acompanhamento e orientação durante esta dissertação. Ao João Pinto, pelo apoio, motivação e ideias que foram fundamentais para a realização desta dissertação.

À minha mãe, por ser a responsável por tudo aquilo que consegui e por sempre me incentivar a seguir os meus sonhos, só espero um dia poder ter tanta força como ela.

Ao Miguel, pela paciência por me aturar tanto ao longo desta tese como ao longo do resto dos dias, não é trabalho para qualquer um.

Aos que a faculdade me deu e que vieram para ficar. Ao Ricardo e ao Duarte, os melhores companheiros de curso e que estiveram sempre lá para dar a motivação necessária para chegar ao fim. Às meninas de metal - Mariana, Áurea, Dri, Lacri -, que me acolheram no meio delas e fizeram destes 5 anos inesquecíveis. À Feta, uma das pessoas com quem mais me identifico e que levo para a vida. A 17/18!

A Vila do Conde e à Póvoa e a todas que nunca se afastam, mesmo com o passar dos anos e os desafios que a faculdade nos trouxe: Marianinha, Filipa, Rosa, Sofia, Eduarda e Xana. Um agradecimento especial à Marianinha, que mesmo em circunstâncias bastante diferentes, viveu comigo e por mim o espírito académico de uma forma muito especial.

A ti ⋆

Mariana Xavier

"A journey of a thousand miles
begins with a single step."

Lao Tzu

Contents

1 Introduction **1**
- 1.1 Context and Motivation . . . 1
- 1.2 Dissertation Goals and Contributions . . . 3
- 1.3 Dissertation Outline . . . 3

2 Fundamental Concepts **5**
- 2.1 Introduction . . . 5
- 2.2 Emotion Modeling . . . 5
  - 2.2.1 Discrete Emotional Models . . . 5
  - 2.2.2 Dimensional Emotional Models . . . 6
- 2.3 Emotion Recognition Modalities . . . 7
  - 2.3.1 Emotion Recognition from Behavioural Signals . . . 8
  - 2.3.2 Emotion Recognition from Physiological Signals . . . 11
  - 2.3.3 Multimodal Emotion Recognition . . . 12
- 2.4 Conclusion . . . 14

3 Multimodal Databases **15**
- 3.1 Introduction . . . 15
- 3.2 Available Databases . . . 15
- 3.3 Conclusion . . . 18

4 Prior Art **21**
- 4.1 Introduction . . . 21
- 4.2 Pre-processing . . . 21
  - 4.2.1 ECG Pre-processing . . . 21
  - 4.2.2 Face Video Pre-processing . . . 22
- 4.3 Feature extraction and selection . . . 23
  - 4.3.1 Handcrafted methods . . . 24
  - 4.3.2 Deep learning methods . . . 33
- 4.4 Classification . . . 34
- 4.5 Fusion . . . 37
  - 4.5.1 Early fusion . . . 37
  - 4.5.2 Late fusion . . . 39
- 4.6 Conclusion . . . 41

5 Implementation of Prior Art Baselines **43**
- 5.1 Introduction . . . 43
- 5.2 Methodology . . . 43

5.2.1 ECG . . . 43
5.2.2 Face . . . 46
5.2.3 Data division . . . 48
5.3 Results . . . 48
5.3.1 Dataset overview . . . 48
5.3.2 ECG . . . 49
5.3.3 Face . . . 53
5.4 Conclusion . . . 55

6 Improvements on Emotion Recognition from ECG Signals 57
6.1 Introduction . . . 57
6.2 Methods . . . 57
6.3 Results . . . 60
6.4 Conclusion . . . 61

7 Improvements on Facial Emotion Recognition 63
7.1 Introduction . . . 63
7.2 Methods . . . 63
7.3 Results . . . 64
7.4 Conclusion . . . 66

8 Multimodal Emotion Recognition 67
8.1 Introduction . . . 67
8.2 Methods . . . 67
8.3 Results . . . 69
8.4 Conclusion . . . 70

9 Conclusion and Future Work 71

References 75

List of Figures

2.1 Plutchik's wheel of emotions, from [78]. 6
2.2 2D emotion space model [78]. 7
2.3 3D emotion space model, from [78]. 8
2.4 Facial action units considered in "Lie to me" series [66]. 9
2.5 Male actor portraying two different emotions, from [54]. 10
2.6 A typical ECG trace of the cardiac cycle, identifying its main events, from [56]. . 12
2.7 Accuracy of emotion recognition model obtained using different modalities, from [88]. 13
2.8 Schematic representation of early fusion techniques. 14
2.9 Schematic representation of late fusion techniques. 14

4.1 ECG baseline wander due to breathing and body movement, from [83]. 21
4.2 ECG affected by powerline interference, from [45]. 22
4.3 Face detection and illumination normalization example, from [68]. 23
4.4 Example of LBP calculation, from [21]. 30
4.5 Example for optical flow computation, from [91]. 31
4.6 Schema of the Inception-ResNet-v2 network, from [88]. 34
4.7 Schematic representation of an SVM, from [8]. 35
4.8 The architecture of the proposed system for face expression classification from [94]. 36
4.9 Structure of bimodal DAE, from [87]. 38
4.10 Architecture of the typical LSTM block, from [40]. 38

5.1 The MDBN-based model proposed by Zhongmin et al., from [88]. 44
5.2 ECG pipeline used to obtain a baseline. 45
5.3 Video pipeline used to obtain a baseline. 47
5.4 Example of pre-processing of an ECG signal. 49
5.5 Confusion matrices obtained using the different windows individually. 50
5.6 Confusion matrices obtained using the different windows individually, while balancing the weights of the classifier. 51
5.7 Train and test accuracies obtained by varying the value of C for 5 s windows. . . 53
5.8 Example of pre-processing face video. 54
5.9 Train and validation loss evolution over the number of epochs using SIFT features. 54

6.1 An overview of the proposed framework for self-supervised emotion recognition presented by Sarkar et al., from [65]. 58
6.2 Schematic representation of the ECG pipeline adopted to aggregate the features of the different segments. 60

7.1 Schematic representation of the pipeline adopted to aggregate the SIFT and deep features of the different frames. 64

8.1 Schematic representation of the pipeline adopted for the early fusion model. . . . 68
8.2 Schematic representation of the pipeline adopted for the late fusion model. 68
8.3 Comparison between accuracies obtained for the modalities, individually and combined. 69

List of Tables

3.1 Summary of current available databases for emotion recognition containing data on facial expression and ECG. 20

4.1 Summary of the handcrafted feature extraction and features selection methods used on the surveyed approaches for ECG emotion recognition. 28
4.2 Summary of the handcrafted feature extraction and features selection methods used on the surveyed approaches for facial expression/emotion recognition. . . . 32
4.3 Summary of current available algorithms for multimodal emotion recognition regarding classification and fusion and the obtained results. 40

5.1 Number of examples for each class both in the original dataset and after pre-processing. 48
5.2 Classification results obtained using the different windows individually. 50
5.3 Classification results obtained using the different windows individually, while balancing the weights of the classifier. 51
5.4 Classification results obtained using the different windows individually, while balancing the weights of the classifier, followed by majority voting of the predictions. 52
5.5 Classification results obtained using the mean and/or standard-deviation aggregation of features. 52
5.6 Classification results obtained using the SIFT features. 54
5.7 Classification results obtained using the features extracted from Inception-Resnet-V2 and Inception-Resnet-v1. 55

6.1 Classification results with four-layer MLP for ECG improvements. 60
6.2 Classification results with LSTM for segments aggregation. 61

7.1 Best classification results obtained using the SIFT features and/or the InceptionResNet-V1 features. 65
7.2 Classification results with LSTM for frames aggregation. 65

8.1 Classification results with LSTM for frames aggregation. 69

Abbreviations

A	Accuracy
AAM	Active Appearance Model
ANS	Automatic Nervous System
AU	Action Unit
BPM	Beats per Minute
BoVW	Bag of Visual Words
BW	Baseline Wander
C3D	3D Convolutional Neural Network
CAGR	Compound Annual Growth Rate
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
CNS	Central Nervous System
DAE	Deep Autoencoder
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DRMF	Discriminative Response Map Fitting
ECG	Electrocardiogram
EDA	Electrodermal Activity
EDR	ECG-derived Respiration
EEG	Electroencephalogram
EOG	Electrooculogram
EMG	Electromyogram
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
FFT	Fast Fourier Transform
FLD	Fisher's Linear Discriminant
GCN	Global Contrast Normalization
GDA	Generalized Discriminant Analysis
GSR	Galvanic Skin Response
hEOG	Horizontal Electrooculogram
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
ICA	Independent Component Analysis
KNN	K-nearest neighbor
LBP	Local-Binary Patterns
LBP-TOP	Local Binary Pattern from Three Orthogonal Planes

LDA	Linear Discriminant Analysis
LF	Low Frequency
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
LSTM	Long-Short Term Memory
MEG	Magnetoencephalogram
MKL	Multiple Kernel Learning
MLP	Multilayer Perceptron
MoT	Mixtures of Trees
MTCNN	MultiTask Cascaded Convolutional Neural Network
NIR	Near-infrared
NLD	Normalized Length Density
PCA	Principal Component Analysis
PNS	Peripheral Nervous System
PSD	Power Spectral Density
QDA	Quadratic Discriminant Analysis
RBM	Restricted Boltzmann Machines
RF	Random Forest
RNN	Recurrent Neural Network
RSP	Respiration
SBS	Sequential Backward Selection
SCL	Skin Conductance Level
SDM	Supervised Descent Method
SFS	Sequential Forward Selection
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
T	Temperature
tEMG	Trapezius Electromyogram
V&J	Viola-Jones
VLF	Very Low Frequency
ZCR	Zero-crossing Rate

Chapter 1

Introduction

1.1 Context and Motivation

Emotions have a significant impact on how people think and act. The emotions that people experience on a daily basis can motivate them to act a certain way and influence the choices they make in their lives. Recognising and reacting to each other's emotions enables us to communicate effectively in a range of social circumstances.

There is widespread concern about taking users' emotional states into account in various research fields. Affective computing is a growing field and can be defined as the study and development of systems and devices that can recognise, interpret, process, and simulate human affect [26]. It is particularly important in areas such as:

- **Robotics:** robots are becoming an important part of our daily lives as they can be designed to interact with humans for very different tasks such as caretaking, reception tasks, and providing services in the living space, among others. Designing robots that can recognise how a person is feeling and react accordingly allows for a better and more pleasing interaction with humans [58; 71; 82];

- **Marketing:** knowing how customers feel has become a key goal in business, which is why companies must keep the customer perspective in mind as they make important decisions [31]. The effectiveness of advertisement on the target audience has been studied, both for a general population [79] and discriminated by gender [19] or race [4], to see its influence on the responses. Another study intended to see whether salespersons who more accurately read nonverbal emotional expressions were more successful [13], which aims for the training of salespersons to bring higher effectiveness to their work.

- **Entertainment industries:** like marketing, the entertainment industry may benefit from emotion recognition to evaluate how the potential customers/users feel and optimize the entertainment according to that target audience. One example is in interactive play environments [73] by capturing usability and playability through metrics relevant to ludic experience; and videogames, by automatically adjusting controllable game parameters to

augment the entertainment value [96]. Another example is in software usability testing, software process improvement [47] and webpage design [27] to aid interface designers in improving the quality of user experience by building software/webpages that fulfill the emotional needs of the users.

- **Education:** emotion recognition research in education has focused on intelligent tutoring systems [10; 80]. These systems help the tutors to provide emotional or motivational feedback and also to construct and successfully adapt their explanations. Furthermore, it has been demonstrated that after interacting with an emotional pedagogical agent, low-achieving students improved their affective outcomes and experienced less frustration and anxiety [10]. Research has also been conducted to develop tools for recognizing human emotions during e-learning processes [75].

- **Healthcare systems:** emotion recognition has been shown to bring advantages in the identification of mental health issues. These are characterized by the absence of positive emotions and a range of associated cognitive, physical, emotional, and behavioural symptoms [3]. By joining the availability of wearable technology with the advances in emotion recognition models, there is an opportunity to monitor those more prone to be diagnosed with a mental health condition, namely people undergoing medical treatments and those in nursing homes. Some studies have already used facial emotion recognition to assess the stress levels of patients [85] or by analyzing biological signals [22; 70], and daily activity patterns [93]. Moreover, in healthcare, the ability to interpret emotions is crucial in shaping the communication between healthcare professionals and patients and the current widespread face mask usage severely impacts it. This functional relationship with patients must be fostered in healthcare professionals and students who are developing their communication and relationship-building abilities [9].

According to a new market research report, in the post-COVID-19 scenario, the global Emotion Detection and Recognition Market size is projected to grow at a Compound Annual Growth Rate (CAGR) of 11.3% between 2020 and 2026, reaching a total market value of $37 billion [59]. This shows that there is still a huge demand for emotion recognition models that drives both the research behind it and the creation of commercial solutions for the task.

Shifting from research to the market, emotion recognition models are well consolidated in some main companies like Affectiva, Noldus, Emotibot Technologies, NVISO, among others. These use primarily facial information to recognise emotions, which brings problems associated with occlusions, illumination, facial alignment, and reliability [78]. This reliability issue is the biggest setback since it would be very problematic to have people intentionally fake their real emotions, specifically for mental health assessment purposes. However, some companies also use multimodal approaches but these focus mostly on joining audio information to facial expression recognition which are both physical characteristics. The ECG signal could bring these companies a competitive advantage by being able to extract useful physiological information that has been shown to correlate with emotional states, while maintaining a simple acquisition system.

1.2 Dissertation Goals and Contributions

As exposed, there are many applications for emotion recognition, and the options available within and beyond research still show plenty of margin for improvement, more specifically for effective multimodal approaches. The work done during this dissertation aimed to explore the use of ECG signals together with face information for the task of emotion recognition, attempting to minimize the individual challenges of each modality. Specifically, this work can be divided into five main goals:

1. design and implement methodologies for emotion recognition based on ECG signals;
2. design and implement methodologies for emotion recognition based on face video;
3. explore the use of temporal models for emotion recognition;
4. implement fusion strategies for multimodal recognition;
5. evaluate and compare the performance of single-modality and multi-modality methods and compare them with the state-of-the-art.

1.3 Dissertation Outline

Following this introduction, the fundamental concepts of emotion recognition research are covered in Chapter 2. This includes concepts behind emotion modeling and the different modalities used.

In Chapter 3, the most relevant multimodal databases available for emotion recognition are reviewed and compared.

After that, Chapter 4 gives a state-of-the-art analysis, describing and discussing some of the most important approaches used in emotion recognition nowadays, both through handcrafted and machine/deep learning techniques.

Subsequently, in Chapter 5, state-of-the-art techniques are implemented to serve as a baseline for the following work, together with a discussion on how some parameter modifications can affect the results.

Chapters 6 and 7 show the methodology implemented for emotion recognition based on ECG signals and face video in order to improve the results from the baseline.

Finally, Chapter 8 presents the methodology and results from fusing the two modalities and Chapter 9 concludes the work while outlining potential directions for future research.

Chapter 2

Fundamental Concepts

2.1 Introduction

Before addressing the state-of-the-art, it is necessary to discuss some basic concepts related to emotion recognition. These concepts include both emotional models and the different modalities used to recognise emotions, looking at their advantages and disadvantages.

Hence, this chapter presents the fundamental concepts needed to build emotion recognition models. It includes a discussion on the different types of emotional models and an overview of the different modalities that may be used for it, with a special focus on the face and the ECG signal.

2.2 Emotion Modeling

One of the starting points for emotion recognition systems is how to define and model emotions. There is not yet a universally accepted emotional model. However, two types of models prevail in the emotion recognition literature: discrete emotional models and dimensional emotional models.

2.2.1 Discrete Emotional Models

The discrete emotional model was first described by Ekman [24] and considers six basic emotions: sadness, happiness, fear, anger, disgust, and surprise. All other emotions are considered to be combinations or reactions to these. Moreover, these basic emotions are considered to be innate and culturally independent - *i.e.* universal regardless of one's gender, age and origin [23]. In the literature, a seventh emotion category is also commonly found: neutral.

Plutchik [69] then extended this list by considering eight emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Figure 2.1 represents the Plutchik's wheel of emotions that describes emotions ranked by intensity, being the stronger ones positioned in the centre. The use of colour emphasizes the fact that fundamental emotions may be combined to form complex emotions. Additionally, its disposition depicts the existent duality in emotions, implying that each one has a polar opposite: sadness is the polar opposite of joy, trust is the polar opposite of disgust, fear is the polar opposite of anger, and surprise is the polar opposite of anticipation.

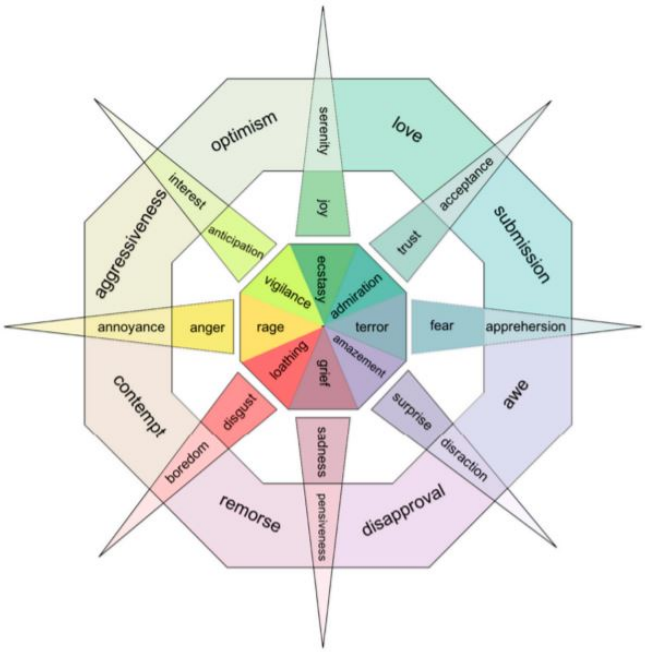


Figure 2.1: Plutchik's wheel of emotions, from [78].

Izard [42] extended these models further by defining ten basic emotions: interest, joy, surprise, sadness, fear, shyness, guilt, anger, disgust, and contempt.

Defining the observed emotions following discrete labels is simple and intuitive when developing a database. In fact, we can easily define others' emotions by how they appear to us. However, these models fail to define more complex emotional states. Besides, they do not consider that a person may be experiencing two labeled emotions at the same time. Finally, many factors may influence one's emotional response; therefore, considering a universal approach as explained can be unsatisfactory.

2.2.2 Dimensional Emotional Models

The dimensional emotional model is an alternative model which suggests that emotions are not discrete but continuous.

Lang et al. [49] suggested a two-dimensional space model based on valence and arousal. Valence is a measure of pleasure that ranges from negative/unpleasant to positive/pleasant. Arousal refers to the amount of activation and intensity concerning emotional stimulus, ranging from low/passive to high/active. Figure 2.2 shows the plot of different emotions over these two dimensions.

A 3D model emerged from this one, considering another dimension: dominance [60]. Dominance corresponds to someone's level of control over the emotion, varying from submissive to dominant. This allows for more complete discrimination between emotions. Figure 2.3 shows the plot of different emotions over these three dimensions.

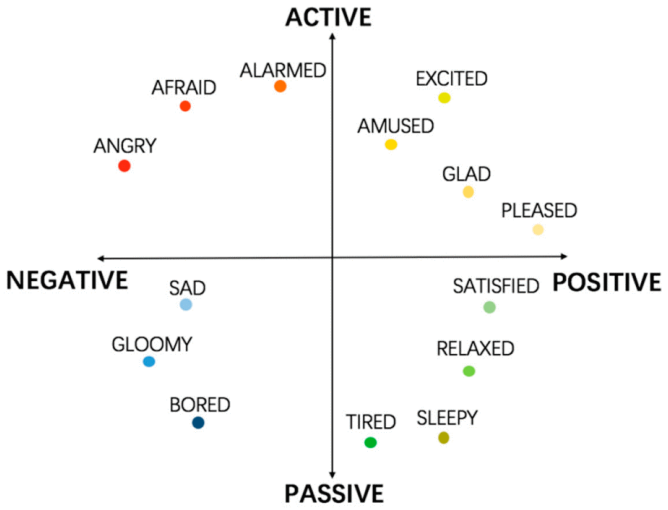


Figure 2.2: 2D emotion space model with valence in the x-axis and arousal in the y-axis, from [78].

These models have been further developed, and Fontaine et al. report consistent results using a set of four dimensions, adding unpredictability [29].

In these dimensional approaches, emotions are not independent of each other, which allows for a more complete characterization. However, when labelling a database, it is not so intuitive to say the value of the dimensions as it is to say the discrete emotion. Therefore, labelling emotions following dimensional models requires special training and a considerable level of know-how, since it is a subjective task.

2.3 Emotion Recognition Modalities

Shu et al. [78] says that "emotion refers to a mental state that arises spontaneously and is often accompanied by physical and physiological changes that are relevant to the human organs and tissues such as brain, heart, skin, blood flow, muscle, facial expressions, voice, etc.". These physical and physiological changes result in two distinct categories of emotion recognition.

The first category is emotion recognition through physical and behavioural signals, which is the most researched category. This is driven by the fact that individuals in human interactions interpret each other's emotions through facial expressions, gestures, posture, and even speech.

The second category is emotion recognition from physiological signals, which include the electroencephalogram (EEG), body temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP), among others. These physiological signals are in response to the central nervous system (CNS) and the peripheral nervous system (PNS) of the human body, specifically the automatic nervous system (ANS), which changes with emotions, following Cannon's theory [89].

More recently, multimodal approaches have been proposed where the information from different types of data is joined. Since multimodal learning is able to take advantage of

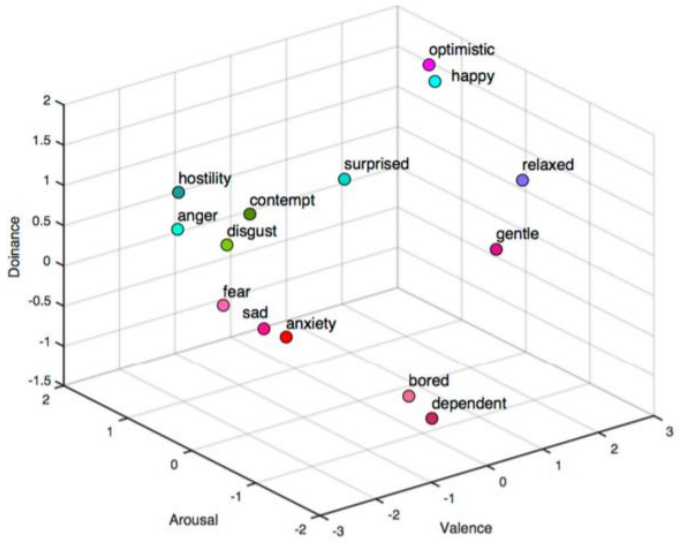


Figure 2.3: 3D emotion space model, from [78].

the complementary nature of the individual signals, the performance of multimodal emotion recognition usually surpasses that based on a single modality [48].

The current section focuses on the modalities listed above for emotion recognition, specifically facial expressions and ECG signals, discussing the benefits and drawbacks of each.

2.3.1 Emotion Recognition from Behavioural Signals

2.3.1.1 Facial Emotion Recognition

A person's face provides an intuition on different aspects of that person, namely regarding their age, gender, cultural background, and even what they are feeling. Therefore, among the different physical and behavioural signals, facial expression has been the most explored modality over the years.

In 1978, Ekman and Friesen developed the Facial Action Coding System (FACS) [25], one of the leading systems in facial expression recognition. This system can characterize facial actions to express emotions based on subtle facial muscle movements. These specific muscles are called Action Units (AUs) and allow for the detection of micro-expressions.

Figure 2.4 demonstrates the coding system used in the "Lie to me" television series, supervised and guided by Ekman himself. It is shown that this system focuses mostly on the eyes, eyebrows, and mouth to detect subtle movements and infer the specific emotional state. However, these movements are not necessarily categorizable into one of the main emotions. Some movements may indicate other things like physical discomfort, injuries, muscular twitches, etc. For example, if a person squints to better see something, that does not indicate an emotion but that a person is short-sighted [6].

Facial emotion recognition (FER) brings challenges that need to be addressed during pre-processing. Pre-processing aims to prepare the image for the next steps by addressing image

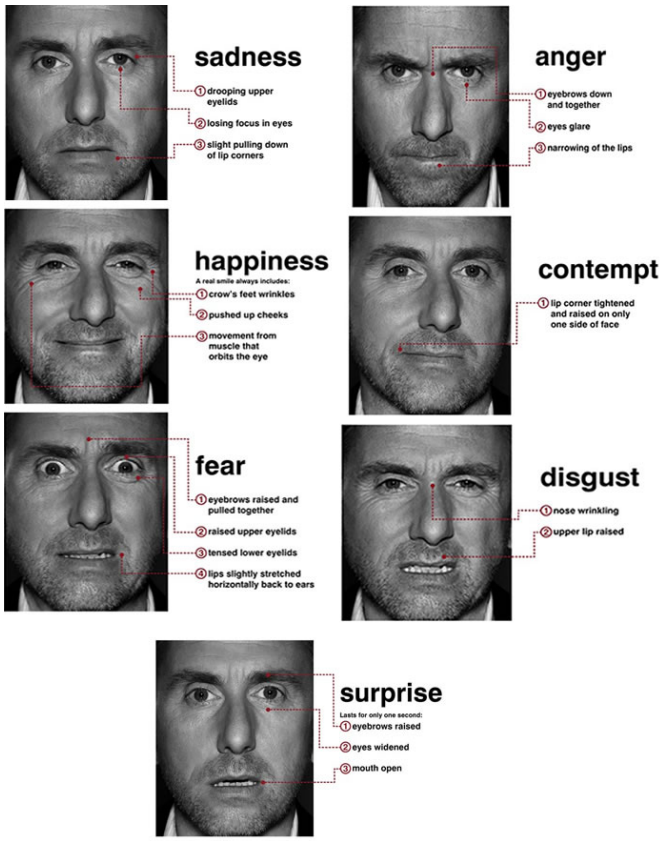


Figure 2.4: Facial action units considered in "Lie to me" series for expression evaluation, from [66].

acquisition challenges, such as occlusions, illumination, and face alignment. Different types of occlusions can occur, such as the use of face masks covering the nose and mouth, the use of sunglasses covering the eyes, pose variations - *e.g.* profile images of the face don't allow for the detection of all the action units -, among others. In the wild, environments are very prone to illumination changes that also difficult the task. Finally, aligning the faces of the different images allows for easier identification of the action units. In videos, an additional challenge appears regarding the continuous detection of the face throughout time [6].

To conclude, facial emotion recognition has the advantage of being easily interpreted by humans, and also it allows for an easy collection of data. Moreover, it has shown promising results over the last years, being the most studied modality. However, among other previously discussed challenges, reliability can't be guaranteed because it is easy for people to purposely hide their real emotions [78]. For example, it is common for people to put on a fake smile in a social situation even if they may feel sad or angry on the inside.

2.3.1.2 Emotion Recognition from Gestures and Posture

Along with facial information, gestures and posture can also give an intuition about the emotional state of a person. However, research focused on postural communication of emotion is scarce due

to the small variety of validated emotion poses.

The notion of emotional body posture is usually determined by the placement of the hands, arms, and shoulders. Figure 2.5 shows a male actor portraying two different emotions: anger on top, and disgust at the bottom. However, these portray an exaggeration of real movements, with the often subtle movements of the body in real life difficulting the recognition of these emotions. Especially with disgust, it is still not reliably recognised when communicated through posture alone, despite studies trying to achieve that [54].



Figure 2.5: Male actor portraying two different emotions: anger on top, and disgust at the bottom, from [54].

Moreover, in the image anger is very much portrayed with an aggressive pose, with the position of the hands in protest and raising of the fist, but anger is often not displayed aggressively. This is another limitation of the modality: there is not one single pose for each emotion. Similarly, the same pose may be used to express different emotions.

2.3.1.3 Speech Emotion Recognition

The final behavioural modality is speech emotion recognition. For most humans, speech characteristics such as pitch, loudness, firmness, and speech rate are indicative of the emotional state besides the actual words being said.

The quality of the data utilized for the model is a significant consideration in assessing speech emotion [7]. Data collection is influenced by factors such as speaker characteristics, length and word count in sentences, the language of speech, environment, recording method, and so on [50]. Furthermore, similar to gestures and body position, when actors portray a given emotion,

they frequently do so as an exaggeration of reality, which does not translate well into real-life circumstances.

Besides being an interesting modality, its achievements are still few due to the exposed challenges.

2.3.2 Emotion Recognition from Physiological Signals

Physiological signals are emerging as a way to recognise emotions due to their relationship with the nervous systems. The nervous systems are largely involuntarily activated and people cannot control how their true emotions are reflected in these signals, which is the biggest advantage of this category. However, signal acquisition is more difficult than the previous one since it includes electrodes properly placed on the human body or wearable devices [78].

As mentioned in section 2.3, a variety of different physiological signals are used to recognise emotions.

The EEG is perhaps one of the most researched signals, with several neuropsychological studies reporting correlations between EEG signals and emotions. The EEG measures electrical activity on the scalp, which has been proven to represent the macroscopic activity of the brain's surface layer underneath. Due to the brain's complexity and functionality of the different areas, many studies have been conducted to establish which parts of the brain are most correlated with emotional activity, being them the amygdala (located close to the hippocampus, in the frontal portion of the temporal lobe), and the pre-frontal cortex (covers part of the frontal lobe) [2].

The ECG is a recording of the electrical activity of the heart caused by the contraction of the cardiac muscles, which indirectly represents the flow of blood inside the heart. It is known to be successful in biometric identification and to vary according to the emotional state [11]. The variation in heart rate caused by emotional states is the intuition behind the usage of ECG signals. Different time and frequency domain characteristics have been employed and demonstrated their efficacy in emotion recognition. Since this work will focus on the use of ECG signals, it is important to review some of its characteristics.

The heart is a muscular organ crucial to human life because it is responsible for contracting and forcing the blood through the body's blood vessels. The heart is formed by four chambers: two upper chambers called the right and left atria, and two lower chambers called the right and left ventricles. The low-oxygen blood enters the heart through the right atrium and is guided to the right ventricle by atrial contraction before being pumped to the lungs. The left atrium collects blood from the lungs and pumps it back into the body once it reaches the left ventricle [35]. With each heartbeat, a cycle of electrophysiological events occurs, which translates into the distinct waves visible in an ECG signal, namely the P wave, the QRS complex, and the T wave, as illustrated in Figure 2.6. The P wave is caused by depolarization of the atria, whereas the QRS is caused by ventricle depolarization and simultaneous atria repolarisation. Finally, the T-wave represents ventricular repolarisation [56]. The identification and analysis of these events are frequently supported in the study of ECG signals for monitoring, diagnosis, and biometrics, thus its prominence in the literature.

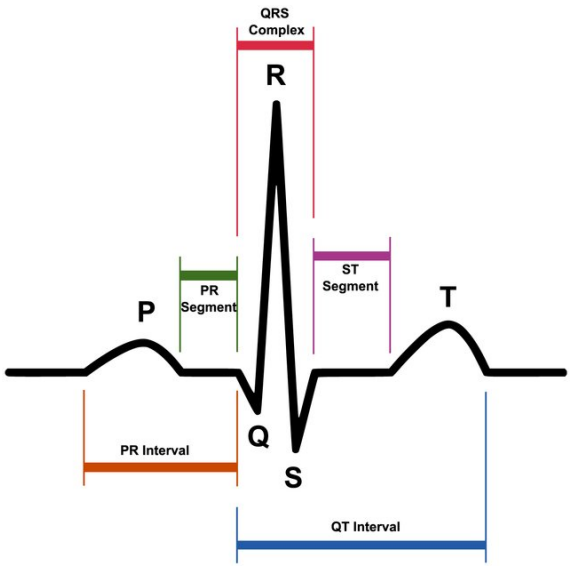


Figure 2.6: A typical ECG trace of the cardiac cycle, identifying its main events: P wave, QRS complex, and T wave, from [56].

Despite the common events present in an ECG signal, it also has a significant degree of variability. This variability may be designated as intra-subject - differences between heartbeats of a single subject, which are primarily interesting for medical purposes - or inter-subject - changes between heartbeats of different individuals, which allow for biometric recognition [67]. Both forms of variability can be caused by several factors, including heart geometry, personal characteristics such as age and weight, physical exercise, cardiac diseases, posture, emotions, and fatigue, and electrode characteristics/placement [67].

Along with heart rate variations, employing respiratory characteristics is a powerful strategy since its changes according to the emotional state are apparent. Pulmonary ventilation and respiration are considered to be largely engaged in the oxygenation of blood for oxygen supply to cells throughout the body for metabolic reasons. However, there is substantial evidence that there is some physiological specificity across the emotional spectrum, with breathing patterns commonly characterized as an indication of emotional state [43].

Overall, physiological signals have proven useful in emotion recognition tasks because of their involuntary nature, but they also introduce problems related to the complexity of data collection and the noise associated with them, which can hinder key features. Taking all factors into consideration, ECG is a promising modality compared to the other physiological signals since much information can be extracted from it, it has been shown to correlate with emotional states and its acquisition is simpler than that of the EEG, which is one of the most studied modalities.

2.3.3 Multimodal Emotion Recognition

The idea that emotions cause both behavioural and physiological changes and that these forms of emotional information are all present when people communicate with each other to manifest

their emotions has led to the thought that fusing these different signals could be beneficial for emotion recognition tasks. Therefore, multimodal methods have emerged as a way to use different modalities that complement each other to fight their individual weaknesses and achieve a more robust and accurate model. An infinite number of combinations of modalities can be used. For instance, joining different physiological signals together, different behavioural signals together, or even different physiological signals with different behavioural signals.

Most multimodal studies present not only the final accuracy reached but also the accuracy that each modality had individually for comparison purposes. In [88], the authors showed that not only merging the features from three physiological signals (ECG, SCL, tEMG) but also joining them with facial features significantly improved the performance of the emotional recognition model. Figure 2.7 shows the accuracy obtained from physiological signals alone, facial video alone, and the final multimodal approach. Similarly, [74] has shown that joining video, audio, and text allowed for increased performance. Also, [36] reached the same conclusion using facial information and the EEG signal.

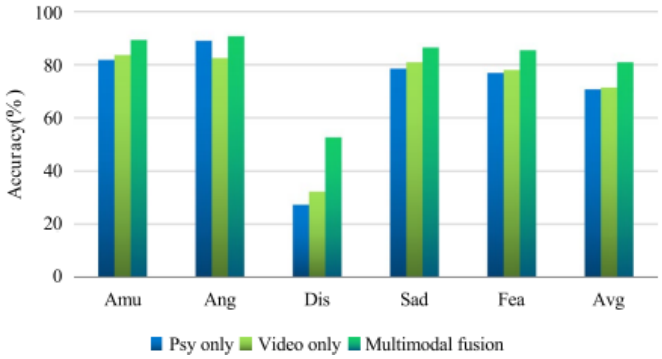


Figure 2.7: Accuracy of emotion recognition model obtained using different modalities, from [88].

Despite the observed increase in performance in multimodal approaches, it also brings higher complexity to the models. This may not be ideal when applied to the different areas of application. Therefore, the trade-off between performance and simplicity may be taken into account when considering multimodal approaches.

Fusing different modalities may be mainly approached in two different ways, depending on the stage at which it is introduced: early and late fusion.

Early fusion consists of extracting the features from the modalities separately and then joining the feature vectors all together, as shown in Figure 2.8. The simplest approach to joining these feature vectors is by concatenating them, called plain early fusion. This resulting feature vector is then used as input for the classification model. This usually results in a high number of features and even redundancy of features from different modalities; therefore, a feature selection approach may be applied [15]. Other more complex early fusion methods aim at obtaining high-level features from the previously extracted ones.

Late fusion merges the results of multiple classifiers into a final prediction [15], as shown in Figure 2.9. Different late fusion strategies are used for emotion recognition with different levels of

complexity. Some of the most common ones are the maximum rule (which selects the maximum of all posterior probabilities [12]), the sum rule (which sums the confidences for a given class from each classifier and chooses the one with the highest confidence as output [12; 33; 52]), the product rule (which works like the sum rule but multiplying the confidences [12; 33]), using a weight criterion (which assigns a value of confidence to each classifier output, outputting a weighted linear combination of predictions [12; 18; 33; 97]) or even a rule-based approach (where a dominant modality is selected for each target class [17]).

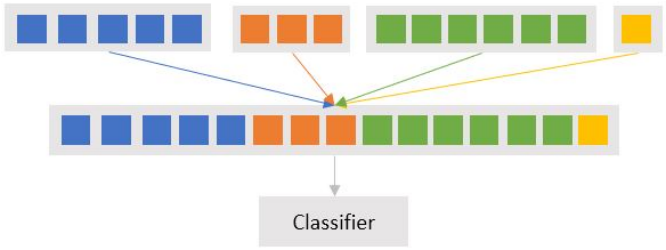


Figure 2.8: Early fusion by concatenating the features extracted from different modalities, represented in the different colours, before applying the classifier model.

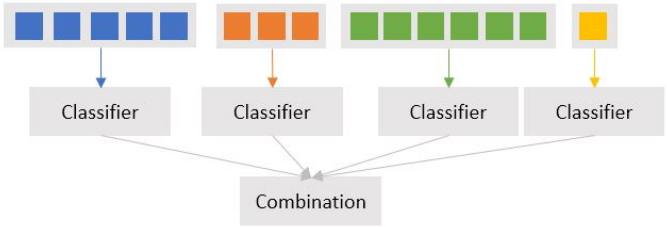


Figure 2.9: Late fusion by applying a classifier model to the features extracted from different modalities, represented in different colours, and combining the results of the models.

2.4 Conclusion

Overall, multimodal approaches have shown to bring great advantages in emotion recognition tasks, with different combinations of modalities and methods used to merge them.

Considering the different combinations of modalities, ECG and face information bring the best trade-off considering performance, usability, and robustness to adverse scenarios. Face information is a well-studied category that has proven to have good results in emotion recognition, although it can be easily faked by humans. ECG cannot, so it may bring the advantage of portraying the truly felt emotion. Moreover, in terms of data collection, both of them can be easily acquired: facial information can be acquired by a camera and ECG acquisition can be much less intrusive and simpler in comparison to other physiological signals like the EEG.

Chapter 3

Multimodal Databases

3.1 Introduction

Databases are fundamental for developing an accurate system, and many factors are involved in creating a good database.

The number of records is one of many factors to consider in a well-structured database because machine learning and, in particular, deep learning algorithms require a large amount of data available. One method to ensure this is to have a large number of subjects, which also ensures subject diversity. Small datasets lead to a variety of problems, including overfitting and signal and subject dependence. Since there is not a representative amount of data for each class, the model may learn patterns in the training data that are not representative and are considered noise, causing it to generalize poorly and leading to overfitting. Besides, with small databases, it is occasionally necessary to use the same subject or segments of the same signal for the training and testing phases of modeling which is not ideal for real-world applications.

Furthermore, aspects like the acquisition settings are also relevant. Regarding face video, illumination is one of the biggest concerns as well as ensuring that the participant's face is always caught on camera even if they slightly move. This leads to a division in the databases into those which are recorded in a laboratory-controlled environment and those recorded in-the-wild, which brings additional challenges but is more similar to real-world scenarios.

Finally, it is crucial that the data is recorded from people who are exhibiting real emotions so that it can be used in daily applications. This is usually achieved by eliciting the emotions of the participants by having them watch videos or listen to music clips. Therefore, naturalistic data should be preferred to acted data - *i.e.*, having an actor demonstrate a specific emotion - whenever possible.

3.2 Available Databases

There are already different databases that contain face expression information and physiological signals, namely ECG, that are further presented and analysed. A summary of these databases may

be found in Table 3.1.

AMIGOS

AMIGOS - A dataset for Multimodal research of affect, personality traits and mood on Individuals and GrOupS - was presented by Miranda-Correa et al. [61] in 2015. Three main neural and peripheral physiological signals were recorded using wearable sensors: EEG, ECG, and GSR. Besides, frontal, full-body, and depth videos were also recorded. The 2.5D nomenclature in Table 3.1 is due to the depth features added to the 2D videos. The recordings were performed in a laboratory environment with controlled illumination.

Different from other databases, emotions were elicited using both short and long videos in two configurations, one with individual viewers and one with groups of viewers. Experimentally speaking, this translates into two distinct settings: in the first one, 40 healthy volunteers watched 16 short emotional videos, and, in the second one, 37 participants watched 4 long videos, 17 alone and the other 20 in a group (5 groups of 4 people). This way, the database allows the study of the affective responses of individuals in relation to their personality and mood, and the analysis of how these responses are affected by the individual/group configuration, and the duration of the videos.

Regarding the emotion assessment, both internal and external annotations were performed, being both present in the dataset. In the internal annotation, the volunteers self-assessed their own affective states in terms of valence, arousal, control, familiarity, like/dislike, and selection of a basic emotion (sadness, happiness, fear, anger, disgust, and surprise plus neutral). In the external annotation, the annotators evaluated the emotion each subject was experiencing through valence and arousal levels.

BioVid Emo DB

BioVid Emo DB was presented by Zhang et al. [99] in 2016. Three psychophysiological signals were recorded: Skin Conductance Level (SCL), ECG, and trapezius EMG (tEMG). Besides, four video signals were also recorded: frontal, left-sided, right-sided, and frontal depth maps with a Kinect Sensor. The experiment was performed in a quiet laboratory environment with controlled illumination.

A total of 94 healthy subjects participated in the experiment but only the data of 86 subjects are available in the database due to missing and/or corrupted recordings. Each participant watched 15 standardized film clips, 3 per each discrete emotion, and only the data of the video for which they felt the strongest emotion is available - *e.g.* if a participant felt most disgusted by the first 'disgust' clip, only the video signals and psychophysiological parameters of this particular clip is provided in the database. The length of the clips varies from 32 to 245 seconds, with an average length of 68 seconds.

Regarding the emotion assessment, the five discrete emotions (amusement, sadness, anger, disgust, and fear) that are portrayed by the videos are the labels of the database.

DEAP

DEAP was presented by Koelstra et al. [46] in 2012. The EEG signal was recorded along with several peripheral physiological signals: GSR, respiration amplitude, skin temperature (T), blood volume, EMG, and Electrooculogram (EOG). Additionally, face video was also recorded. The experiments were performed in laboratory environments with controlled illumination.

The experiment involved 32 volunteers; however, face video was only recorded for 22 of the participants. The emotion elicitation was made with a subset of 40 available one-minute long excerpts of music videos.

Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity.

DECAF

DECAF - a multimodal database for DECoding user physiological responses to AFfective multimedia content was presented by Abadi et al. [1] in 2015. Different modalities were recorded: near-infrared (NIR) facial videos, horizontal Electrooculogram (hEOG), ECG, tEMG, EEG, and Magnetoencephalogram (MEG), which also acquires brain signals - enabling a comparison between the EEG and MEG modalities.

A total of 30 volunteers participated in the experiment, which consisted of listening to 40 one-minute music records and 36 movie clips. This allows for a comparison between music and movie stimuli for emotion recognition. All MEG recordings were performed in a shielded room with controlled illumination. Due to the sensitivity of the MEG equipment, all other devices used for data acquisition were placed in an adjacent room and controlled by the experimenter.

Regarding the emotion assessment, the participants were asked to rate each stimulus for arousal, valence, and dominance, from 1 (low) to 9 (high). In addition, a thorough data labeling was performed by 42 annotators.

MANHOB-HCI

MAHNOB-HCI was presented by Soleymani et al. [81] in 2012. Different modalities were recorded: EEG, eye gaze, audio, face and body videos, and different physiological signals - ECG, GSR, skin temperature (T), and respiration (RSP) amplitude.

A total of 30 healthy volunteers participated in a total of two experiments for the development of this database; however, only the data of 27 of them are available in the database due to technical problems and unfinished data collection of the remaining 3 participants. The first experiment consisted of watching 20 emotional videos; while the second one consisted of watching 14 video fragments and visualizing 28 images. The experiment environment in the laboratory was isolated from the outside environment, and the light and temperature were controlled to avoid variation in physiological reactions due to uncontrolled parameters.

Regarding emotional assessment, in the first experiment, the participants had to self-report their felt emotions using arousal, valence, dominance, and predictability as well as emotional keywords; while in the second experiment they were asked to agree or disagree with the tag that

was displayed along with the emotion elicitation.

RECOLA

RECOLA - REmote COLlaborative and Affective interactions - was presented by Ringeval et al. [72] in 2013. Different multimodal data were recorded continuously and synchronously: face video, audio, ECG, and electrodermal activity (EDA), also known as GSR.

In total, 46 participants took part in the experiment and were recruited as 23 dyadic teams - *i.e.* groups of 2 participants. The experiment consisted of a survival task that required the collaboration of the members of the group. It was performed in two controlled rooms in terms of lighting and isolation from external noises, one for each participant.

Regarding annotation, although the full length of the recording is available in the database, only the first five minutes of each recording were annotated, which was believed to be the most interesting part of the survival task, in which the participants spent more time discussing their strategies. The annotators measured emotion continuously on two dimensions: arousal and valence, as well as social behaviour labels on five dimensions. Concerning the annotation of the social behaviour, it consisted of answering each of these five questions: (i) agreement: does the person seem to agree with his/her partner? (ii) dominance: does the person appear to be dominant? (iii) engagement: does the person seem to be engaged? (iv) performance: does the person's speech appear to be clear and relevant for the task? (v) rapport: do the person and his/her partner seem to be or could become friends?

3.3 Conclusion

Bearing in mind the discussed requirements regarding a good database and the information gathered from each database previously presented, some conclusions may be taken regarding the most suitable databases for the work.

For instance, due to the high number of subjects in BioVid Emo DB, substantially higher than in the remaining ones, this database is ideal for the task at hand, which requires a high amount of data. However, since its annotations are only regarding discrete models, it presents the limitations of these types of models.

Considering only databases with annotations regarding dimensional models, RECOLA is the one that presents the highest number of subjects. However, it only has annotations regarding valence and arousal which consist of the most simple dimensional model, the 2D one, which besides being very much used is still incomplete in comparison to more recent models.

Finally, AMIGOS seems to be a good database as well, since it allows to establish the comparison between annotations following discrete and dimensional models with several dimensions. However, in both RECOLA and AMIGOS, the number of subjects is considerably smaller than in BioVid Emo DB.

Following these observations, it can be concluded that there is not one perfect database. Researchers should aim at developing larger databases, either in terms of number of subjects

or number of records per subject. This would allow dividing different subjects into training and testing sets, therefore having a test set of subjects "unseen" by the algorithm, which is ideal for real-world applications but hard to perform with a low number of subjects. Moreover, having a representative amount of data for each class would allow quality learning and avoid the overfitting phenomena. The elicitation of emotions found in the databases are all very similar, but the annotation of emotions varies. This is also a factor that researchers need to focus on carefully since it is crucial to determine the quality of a database. For instance, having untrained volunteers perform an annotation can be very susceptible to errors, especially considering dimensional emotional models.

Table 3.1: Summary of current available databases for emotion recognition containing data on facial expression and ECG.
*On DEAP, only 22 subjects have frontal face video.

Collection	Traits	Subjects	Data Sources	Emotion Elicitation	Annotations
AMIGOS [61]	ECG, EEG, GSR, Face (2.5D), Full-body	40	Lab	Short vs Long Videos	Valence, Arousal, Control, Familiarity, Like/Dislike, Six Basic Emotions + Neutral
BioVid Emo DB [99]	SCL, ECG, tEMG, Face (2.5D)	86	Lab	Film Clips	Five Discrete Emotions
DEAP [46]	EEG, ECG, EMG, GSR, RSP, T, Blood Volume, Face (2D)	32*	Lab	Music Clips	Arousal, Valence, Like/Dislike, Dominance and Familiarity
DECAF [1]	ECG, EEG, MEG, hEOG, tEMG, Face (NIR)	30	Lab	Movies and Music Clips	Valence, Arousal, Dominance
MAHNOB-HCI [81]	ECG, EEG, RSP, T, Eye Gaze Audio, Face (2D), Body Video	27	Lab	Videos and Images	Valence, Arousal, Dominance, Predictability
RECOLA [72]	ECG, EDA, Face (2D), Audio	46	Lab	Team Survival Tasks	Valence, Arousal

Chapter 4

Prior Art

4.1 Introduction

A multimodal emotional recognition model is composed of four main parts or processes: pre-processing, feature extraction, classification, and fusion. Note that the fusion step is not necessarily the last one since it can occur in different phases of the process, as presented in subsection 2.3.3. The present section aims to present and discuss the various methods used in the literature for the different steps of the process and both modalities (face video and ECG signals).

4.2 Pre-processing

The aim of pre-processing is to prepare the raw inputs for the future steps of the process by addressing problems related to the inputs' acquisition and other challenges intrinsically related to the modality. Therefore, each modality has its specific pre-processing pipeline.

4.2.1 ECG Pre-processing

The acquisition of the ECG signals is affected by different types of noise which makes difficult a reliable extraction of features.

One type of noise very much present is the baseline wander (BW), which is a low-frequency artifact caused by breathing and body movement [45]. It translates into the signal as an oscillation of the baseline, as seen in Figure 4.1 - *i.e.* the signal does not appear to be aligned on the x-axis.

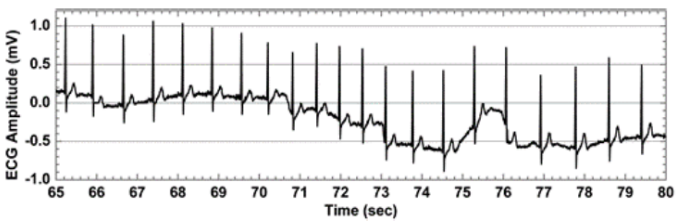


Figure 4.1: ECG baseline wander due to breathing and body movement, from [83].

To remove the baseline wander, different approaches are found in the literature. Since this artifact is a low-frequency noise, one common example of its removal is by using a high-pass filter, with cut-off frequencies usually in the range of 0.5 to 1 Hz. However, some consider that this approach can distort the ECG waveform. Therefore, other possible alternatives to filtering are transforms that may also cancel the ECG noise. Wavelet transforms can accurately capture both time and frequency information of a signal and, in [14], different wavelet transforms are compared in terms of performance.

Another type of common noise is called powerline interference. Such noise is characterized by a 50/60 Hz interference that arises from electrical interference - *e.g.* light bulbs, electronic devices like mobile phones and computers, etc. [45]. Figure 4.2 shows the effect of this noise on an ECG signal, in which it can be observed that it superimposes the low-frequency ECG waves like P and T waves.

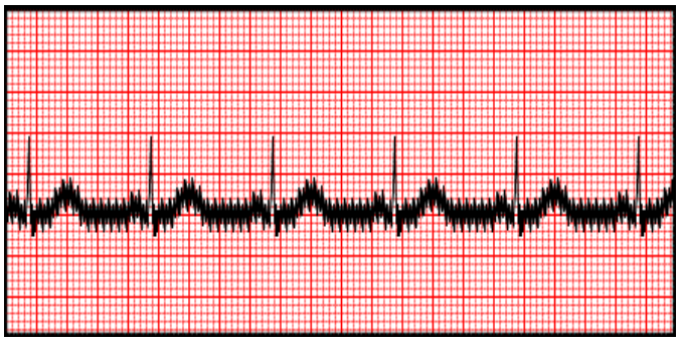


Figure 4.2: ECG affected by powerline interference, from [45].

To remove the powerline interference, a notch filter is normally used - *i.e.* a bandstop filter that eliminates the components of the signal within a specific frequency range and allows frequencies outside of that range.

In addition, electrode motion artefacts are mainly caused by small movements of the electrodes in contact with the skin. This artefact is more problematic since its spectral content overlaps with the QRS complexes making it more difficult to filter out. It is most problematic in uncontrolled environments with an increased amount of movement from the person leading to the detection of false heartbeats [45].

Finally, some artifact components can be removed through visual detection, requiring more expertise and knowledge [78].

4.2.2 Face Video Pre-processing

Considering face video, one of the first steps of pre-processing is usually the extraction of the frames from the video. While some consider that only some key frames are necessary, others say that all frames are essential to reduce the loss of information.

Face detection and alignment are traditional pre-processing steps in many face-related recognition tasks. Detection approaches locate the faces present in the data, obtaining their

bounding box or geometry. The Viola-Jones (V&J) face detector [86] is a classic and widely employed implementation for face detection, which is computationally simple for detecting near-frontal faces. However, it has problems with occlusions and large pose variations. The MultiTask Cascaded Convolutional Neural Network (MTCNN) [98] is a modern alternative for the V&J algorithm, leveraging a 3-stage neural network detector. This method uses facial landmarks - *i.e.* eyes, nose, and corners of a mouth - for the detection, so these can also be used for face alignment without extra computation steps. In fact, the authors in [51] have compared the existing methods for face alignment and concluded that deep learning methods like the MTCNN are the state-of-the-art when it comes to facial alignment since it achieves a very good performance in a fast way. Classic models such as the Active Appearance Model (AAM), the mixtures of trees (MoT) structured models, and the discriminative response map fitting (DRMF) lack in both speed and performance in comparison to new deep learning methods. In 2019, Deng et al. [20] presented RetinaFace, a robust single-stage face detector that performs pixel-wise face localisation on various scales and which has proven to be able to perform better than the previous state-of-the-art methods, namely the MTCNN.

Variations in illumination and contrast can introduce large changes in images, especially in unconstrained environments, and hence impair the FER performance. Therefore, illumination normalization is usually part of pre-processing when the lighting conditions are not controlled. Several methods have been found in the literature and in [68], the authors compared three different methods: global contrast normalization (GCN), local normalization, and histogram equalization. GCN and histogram equalization were reported to achieve the best accuracy for the training and testing steps, respectively. Figure 4.3 shows the pre-processing pipeline used in this comparison study [68], showing an example of face detection and illumination normalization.

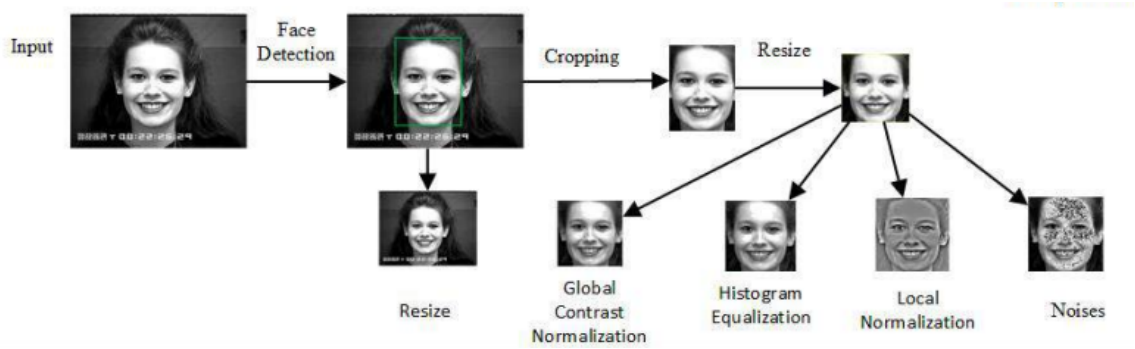


Figure 4.3: Face detection and illumination normalization example, from [68].

4.3 Feature extraction and selection

Feature extraction is able to transform the input into a set of important and relevant features for the following classification. The found methods for feature extraction may be divided into feature extraction from handcrafted features, in which the researcher manually computes and extracts the

features before feeding them to the classifier, and feature extraction from deep learning methods, in which patterns are automatically detected and recognized from the data and features are obtained for further classification. The performance of the latter usually surpasses that of the former but it requires more data to be able to extract useful information.

4.3.1 Handcrafted methods

4.3.1.1 ECG

Looking into the handcrafted features, these can be categorized according to the domain they are extracted as frequency or time-domain features. Furthermore, features can also be distinguished as fiducial or non-fiducial. Fiducial Features are retrieved from specific points of a given signal, such as the P wave, the QRS complex, and the T wave. Contrarily, non-fiducial features consider the totality of a signal or segments of it. This section divides approaches concerning the extracted features domain (time or frequency) and further analyses the fiducial and non-fiducial approaches in each. Fiducial approaches are usually present in time-domain methods, whilst non-fiducial approaches may be present in both domains. In each domain, statistical features calculated from the signal were also considered. Table 4.1 summarizes the techniques and features that will be further analysed.

Time-domain features

As already mentioned, the intuition behind the use of ECG signals is the variation in heart rate due to emotional states. Therefore, time-domain features are most times associated with heart rate (HR) and heart rate variability (HRV).

Zhongmin et al. [88] took segments of the ECG signals to extract statistical features: mean, median, maximum, range, variance, and standard-deviation. This is a non-fiducial approach since it does not focus on specific points of the ECG signal. However, these statistical measures can also be calculated as fiducial features as performed by many different authors. For instance, in the work of Wei et al. [90], the mean and the standard-deviation were taken from the amplitude of the P, R, and T peaks and also from the HRV.

In the work of Valstar et al. [57], both similar non-fiducial and fiducial approaches were considered by, on the one hand, taking the mean and variance of the ECG signal and, on the other hand, taking the mean, variance, mean of rising and falling slope and percentage of rising values of HR and HRV. Additionally, the authors extracted the zero-crossing rate (ZCR) - that, as the name indicates, is the rate at which a signal transitions from positive to negative or vice-versa -, the normalized length density (NLD), the skewness, and the kurtosis values. Skewness is a measure of the asymmetry of a distribution while kurtosis defines how heavily the tails of a distribution differ from the tails of a normal distribution. In this case, these were computed from the time-domain signal, but can also be computed from the frequency-domain signal, being called spectral skewness and spectral kurtosis. The formulas for the calculus of these two measures are the following:

$$skewnessX = \frac{\mathscr{P}X - \mu^3}{\sigma^3}, \tag{4.1}$$

$$kurtosisX = \frac{\mathscr{P}X - \mu^4}{\sigma^4}, \tag{4.2}$$

where X represents the input signal, μ is the mean value, and σ is the standard deviation.

Guo et al. [34], besides skewness and kurtosis, also calculates the entropy, which characterizes the randomness of data set, from the HRV by the formula:

$$entropyX = -sump * log_2 p, \tag{4.3}$$

where p contains the histogram counts of the data set. In the same work, additional fiducial features were introduced, including the average of RR intervals (mean-RRI), standard-deviation of the RR intervals (SDRR), the coefficient of variation of RR intervals (CVRR) - corresponding to the ratio of the SDRR and mean-RRI -, and the standard deviation of the successive differences of the RR intervals (SDSD).

Finally, Hsu et al. [41], besides the SDSD and the standard-deviation of RR intervals - here named SDNN -, calculated other fiducial features: root mean square of differences between adjacent RR intervals (RMSSD), number of successive RR intervals that differ more than 50 ms (NN50), percentage of successive RR intervals that differ more than 50 ms (pNN50), the number of beats per minute (BPM), the median value of RR intervals (median-RRI), the interquartile range of RR intervals (IQR-RRI), mean absolute deviation of RR intervals (MAD-RRI), mean of the difference between adjacent RR intervals (diff-RRI), coefficient of variation of RR intervals (CV-RRI), and the difference between the maximum and the minimum RR interval (range).

Frequency-domain features

In order to extract features from the frequency domain, signals need to be transformed from their natural time domain to this one. Some of the most used techniques are the Fast Fourier Transforms (FFT), the Power Spectral Density (PSD) and the Discrete Cosine Transform (DCT) [34; 41; 62].

The FFT is an algorithm that computes the discrete Fourier transform (DFT) of a sequence in a rapid way in comparison to the direct application of the formula 4.4:

$$X_k = \mathop{\mathscr{P}}_{n=0}^{N-1} x_n e^{-i2\pi kn \Uparrow N}, \tag{4.4}$$

with x_0, ..., x_{N-1} be complex numbers and k = 0, ..., N-1. The difference between a DFT and a DCT is that the DCT uses only real numbers, while a Fourier transform can use complex numbers. The DFT shows the spectral content of the signal: the amplitude and phase of harmonics in the signal whilst the PSD describes how the power of the signal is distributed over frequency, applying to signals existing over all time, or over a large enough time period.

In the works of Ménard et al. [62] and Sun et al. [84], the features used for classification are the actual coefficients from the frequency-domain representation of the signals. However, from these representations, a vast number of features can be extracted.

In many pieces of research [34; 41; 57], the power of the signal in different frequency bands is extracted from the frequency-domain representation, being these bands divided into very-low-frequency range (VLF) (0.0033–0.04 Hz), low-frequency range (LF) (0.04–0.15 Hz), and high-frequency range (HF) (0.15–0.4 Hz). These are frequently used to calculate the power ratio between the low-frequency range and the high-frequency range (LF/HF). In Hsu et al. [41], the authors went further on to calculate extra features related to these such as the LF and HF powers normalized to the sum of the LF and HF power (LF_{norm} and HF_{norm}); the power of each one of the three bands expressed as a percentage of the total power (pVLF, pLF, and pHF); the frequency of the highest peak in each of the three bands (VLF_{fr}, LF_{fr}, HF_{fr}).

Alike the time-domain features, researchers usually exploit statistical features from the frequency-domain representations such as the maximum, mean, standard-deviation, kurtosis, skewness, and entropy as seen in different papers [57; 88; 90].

Non-linear features

Table 4.1 shows a different category inside feature extraction, regarding non-linear analysis. The Poincaré plot of RR intervals is used to evaluate HRV. The resulting point cloud is usually characterized by its length along the line of identity (SD_2) and its breadth across this line (SD_1). These two values, together with the ratio between SD_1 and SD_2, represent valuable measures, having been used in Guo et al. [34] and Hsu et al. [41], and can be calculated by the following formulas:

$$SD_1 = \frac{1}{2SDSD_2} \tag{4.5}$$

$$SD_2 = 2SDRR_2 - \frac{1}{2SDSD_2} \tag{4.6}$$

$$SD_{12} = \frac{SD_1}{SD_2} \tag{4.7}$$

Hsu et al. [41] also calculated a set of non-linear features related to ECG-derived respiration (EDR) - namely the respiratory rate (RSPrate) and the coherence between final EDR signal and RR intervals (coherence) -, with autocorrelation of the QRS complexes through time - namely the maximum autocorrelation coefficient (ACF_{coef}) and the reciprocal of the lag time (ACF_{freq}) autocorrelation -, and with non-linear dynamics - namely the approximate entropy (ApEn) and the sample entropy (SampleEn), which are similar methods that quantify the randomness or predictability of RR interval dynamics.

Feature selection

As observed, after the feature extraction phase, many researchers end up with a large number of features to enter classification, some of which might be irrelevant and others redundant. It may easily lead to different consequences such as to propitiate a time-consuming process to analyse and train the classifier, and to cause overfitting problems and poor ability of generalization, which leads to a low recognition rate. It is also easy to encounter the problem of sparse features, also known as "curse of dimensionality", which results in the decrease of the model performance [78]. Feature selection allows to circumvent these issues.

There are several feature selection algorithms. In general, some algorithms reduce the dimensionality by taking out some redundant or irrelevant features like sequential backward selection (SBS) and sequential forward selection (SFS), while others transform the original one into a new set of features like principal component analysis (PCA) and independent component analysis (ICA) [78].

The surveyed research has led to two methods for feature selection: PCA, in Guo et al. [34], and the Generalized Discriminant Analysis (GDA), in Hsu et al. [41]. The GDA differs from the PCA in being a feature reduction approach that is used for dealing with non-linear discriminant problems using a kernel function operator.

Table 4.1: Summary of the handcrafted feature extraction and features selection methods used on the surveyed approaches for ECG emotion recognition.

Paper	Year	Dataset	Feature Extraction			Feature Selection
			Time-domain	Frequency-domain	Non-linear	
Ménard et al. [62]	2015	Private	-	FFT coefficients	-	
Guo et al. [34]	2016	Private	mean-RRI, CV-RR, SDRR, SDSD, kurtosis, skewness, entropy	LF, HF, LF/HF	SD_1, SD_2, SD_{12}	PCA
Valstar et al. [57]	2016	RECOLA	ZCR, mean, variance, skewness, kurtosis, NLD mean, variance, mean of rising and falling slope and percentage of rising values of HR and of HRV	Spectral entropy, slope and mean LF, HF, LF/HF	-	-
Wei et al. [90]	2018	MAHNOB-HCI	mean and standard--deviation of amplitude of P, R, T and HRV	maximum, mean and standard-deviation from PSD of HRV	-	-
Hsu et al. [41]	2020	Private	SDNN, RMSSD, NN50, pNN50, SDSD, BPM, median-RRI, IQR-RRI, MAD-RRI, diff-RRI, CV-RRI, range	VLF, LF, HL, TP, LF/HF, LF_{norm}, HF_{norm}, pVLF, pLF, pHF, VLF_{fr}, LF_{fr}, HF_{fr}	RSP_{rate}, coherence, SD_1, SD_2, SD_{12}, ApEn, SampleEn ACF_{coef}, ACF_{freq}	GDA
Zhongmin et al. [88]	2020	BioVid Emo DB, DEAP	mean, median, maximum, minimum, range, variance, standard-deviation	amplitude, skewness, kurtosis, PSD, DE	-	-
Sun et al. [84]	2022	Private	-	DCT coefficients	-	-

4.3.1.2 Face

Besides the division between handcrafted and deep learned, facial features can be further divided into global and local features. Global features extract information from the whole facial region, whilst local ones extract information from specific regions of interest, usually corresponding to AUs. Features can also be split into static and dynamic: static features describe a single frame or image, and dynamic ones include temporal information.

Handcrafted features can be further divided into appearance and geometrical. Appearance features use the intensity information of the image, while geometrical ones measure distances, deformations, curvatures, and other geometric properties. In the deep learned features, the nature of the extracted information is usually unknown. Although geometrical features are effective for describing FEs, they fail to detect subtler characteristics like wrinkles, furrows, or skin texture changes. Appearance features are more stable to noise, allowing for the detection of a more complete set of FEs, being particularly important for detecting microexpressions [15]. Table 4.2 summarizes the techniques and features that will be further analysed.

Appearance features

Global appearance features are based on feature descriptors extracted from the whole facial region. Gabor filters and Local-Binary Patterns (LBP) are the most popular global appearance features.

Gabor filter-based feature extraction consists of the application of a Gabor filter bank to the input image, defined by its parameters including frequency (f), orientations (θ), and smooth parameters of the Gaussian envelop (σ), which essentially means that it analyzes whether there is any specific frequency content in the image in specific directions in a localized region around the point or region of analysis [44]. In Gu et al. [32], each image is divided into equally-sized local blocks, without focusing on specific regions of interest. The Gabor filters are applied to each block, combining them after classification.

LBP was first presented in 1994 [63] to be used in texture description. This method labels the pixels of an image by thresholding the neighbourhood of each pixel and considers the result as a binary number: 1 if the neighbourhood pixel's intensity is equal to or higher than the one in the centre and 0 otherwise. After calculating with all neighbours, an eight-digit binary value is produced. When this binary value is converted to decimal, the LBP code of the pixel is generated and placed at the coordinates of that pixel in the LBP matrix. Figure 4.4 demonstrates the described operation, which brings the advantage of being computationally simple. Shan et al. [77] implemented LBP as feature descriptors, using an AdaBoost to learn the most discriminative LBP features.

Another well-established method to obtain global appearance features is through Scale-Invariant Feature Transform (SIFT). SIFT algorithm was first presented in 2004 [55] and not only is it rotation-invariant but it is also concerned with scaling. It is a feature extraction method where image content is transformed into local features that are invariant to translation,

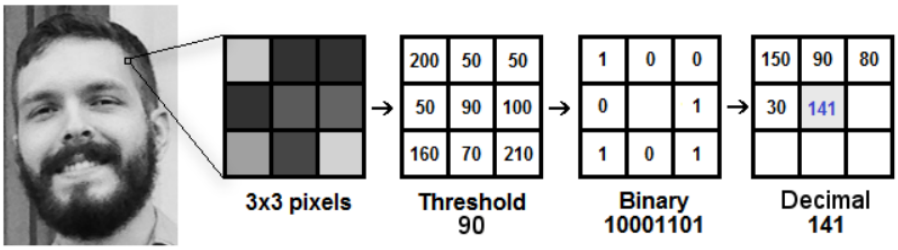


Figure 4.4: Example of LBP calculation, from [21].

scale, and other image transformations. Also, it is a very efficient method with a close to real-time performance. First, images are searched for local extrema over scale and space. Then, it finds potential key points and uses both a contrast and an edge threshold to eliminate any low-contrast key points and edge key points and what remains are the strong interest points. After, an orientation is assigned to each key point to achieve invariance to image rotation and, finally, a key point descriptor is created. In addition to this, several measures are taken to achieve robustness against illumination changes, rotation, etc [64]. In Zhongmin et al. [88], the authors have extracted SIFT features from facial images and then merged them with deep extracted features in order to propose an emotion recognition methodology.

Dynamic global appearance features are an extension to the temporal domain. Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) is an extension of LBP computed over three orthogonal planes at each bin of a 3D volume formed by stacking the frames [15]. In Zhao et al. [100], LBP-TOP is proposed, and in Valstar et al. [57], the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) was used, first by splitting the video into spatio-temporal video volumes, applying the Gabor filter and only then the LBP method.

Local appearance features require previous knowledge of regions of interest such as mouth, eyes, or eyebrows. Consequently, its performance is dependent on the localization and tracking of these regions. In Geetha et al. [30], the appearance of gray-scale frames is described by spreading an array of cells across the mouth and extracting the mean intensity from each cell.

Geometric features

Global geometric features for RGB images usually describe the face geometry based on the location of specific fiducial points, by attempting to measure distances, deformations, curvatures, and other geometric properties [28]. In the work of Valstar et al. [57], 49 facial landmarks were tracked with the supervised descent method (SDM) [92] and aligned with a mean shape from stable points (located on the eye corners and the nose region). As features, the authors considered the difference between the coordinates of the aligned landmarks and those from the mean shape and between the aligned landmark locations in the previous and the current frame. The facial landmarks were then split into groups according to three different regions: i) the left eye and left eyebrow, ii) the right eye and right eyebrow, and iii) the mouth. For each of these groups, the Euclidean distances (L2-norm) and the angles (in radians) between the points were computed.

The Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame was also computed. Finally, the arithmetic mean and the standard-deviation were computed on all features.

Geometric features can provide useful information when tracked on the temporal axis. For RGB data, facial motions can be estimated through optical flow, as done by Wöllmer et al. [91]. Figure 4.5 shows an example present in this work that demonstrates how a movement in the mouth region translates itself into optical flow computation.

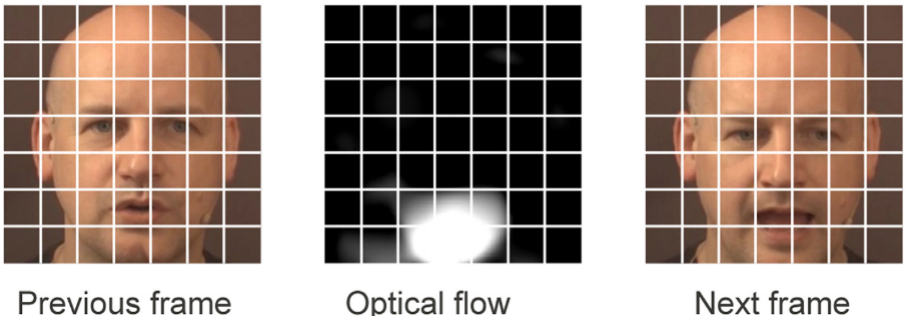


Figure 4.5: Example for optical flow computation: between the frames there is a substantial change in the mouth region, from [91].

Feature selection

As previously seen in ECG based approaches, many authors apply feature selection methods prior to classification in face based approaches as well. Both Gu et al. [32] and Valstar et al. [57] have applied PCA, which has been previously discussed. However, in the work of Gu et al. [32], another feature selection method is presented: Fisher's Linear Discriminant (FLD). FLD is used to seek a projection for each local feature such that it can be optimally separated from the others, taking into account the categories in the data, unlike PCA.

Table 4.2: Summary of the handcrafted feature extraction and features selection methods used on the surveyed approaches for facial expression/emotion recognition.

Paper	**Year**	**Dataset**	**Feature Extraction**		**Feature Selection**
			Appearance	**Geometric**	
Zhao et al. [100]	2007	Cohn-Kanade	LBP-TOP	-	-
Geetha et al. [30]	2009	-	intensity	-	-
Shan et al. [77]	2009	Cohn–Kanade	LBP	-	-
Gu et al. [32]	2012	JAFFE, CK	Gabor filters	-	FLD, PCA
Wöllmer et al. [91]	2013	SEMAINE	-	optical flow	-
Valstar et al. [57]	2016	RECOLA	mean and standard-deviation of LGBP-TOP features	mean and standard-deviation of coordinates differences, distances, and angles	PCA
Zhongmin et al. [88]	2020	BioVid Emo DB, DEAP	SIFT	-	-

4.3.2 Deep learning methods

Deep learning has recently gained popularity since it has attained cutting-edge performance for several applications. Deep learning attempts to capture high-level abstractions through hierarchical architectures of multiple non-linear transformations and representations. However, it has the downfall of requiring a large amount of data to be able to learn from.

Deep learning methods are transversal to ECG signals and face videos. Therefore, in this section, some deep learning techniques that have been applied for emotion recognition are briefly presented with examples of use for either ECG signals, face videos, or both.

Convolutional neural network (CNN)

Due to its translation invariance properties, CNN has been widely used in a variety of computer vision applications. Convolutional layers, pooling layers, and fully connected layers are three types of heterogeneous layers found in deep feed-forward artificial neural networks. To convolve through the entire input and produce activation feature maps, the convolutional layer has a set of learnable filters. Following the convolutional layer, the pooling layer is utilized to lower the spatial size of the feature maps and the network's computational cost. The two most frequent non-linear down-sampling procedures for translation invariance are average pooling and max pooling. These two types of layers are the part of the network responsible for feature extraction - *i.e.* turn the input into a meaningful representation -, whilst the fully connected layer, normally added at the end of the network, is responsible for classification, ensuring that all neurons in the layer are fully connected to activations in the previous layer and to enable the 2D feature maps to be converted into 1D feature maps for classification.

Regarding ECG, Santamaria-Granados et al. [76] used a convolutional neural network for the automatic feature extraction of the physiological signals, and through fully connected network layers, the emotion prediction is made. This is a fully-supervised learning task, which takes the labels of the training data to train and update the whole model in order to achieve the best results. In Sarkar et al. [65], a self-supervised learning is proposed, in which the network learns high-level abstract representations from unlabeled ECG data by applying transformations to the ECG signals and training the model on a transformation recognition task. This allows the model to learn spatiotemporal representations that generalize well across different datasets and different emotion categories. The weights of the convolutional layers in the self-supervised network are then transferred to an emotion recognition network, being kept frozen, and the fully connected layers are trained with the labeled ECG data.

Regarding face, Zhongmin et al. [88] followed an approach similar to the one from Sarkar et al. [65] by taking advantage of the weights of the convolutional layers from a pre-trained network to extract the features and only train the fully connected layers for classification. In this paper, they have used the Inception-ResNet-v2, which is a type of convolutional neural network that can accurately recognize similar objects and achieves very good results in image classification, pre-trained on ImageNet database. The structure for this network can be found in

Figure 4.6. Finally, Nguyen et al. [16] used a 3D convolutional neural (C3D) network to extract spatio-temporal features from the video stream.

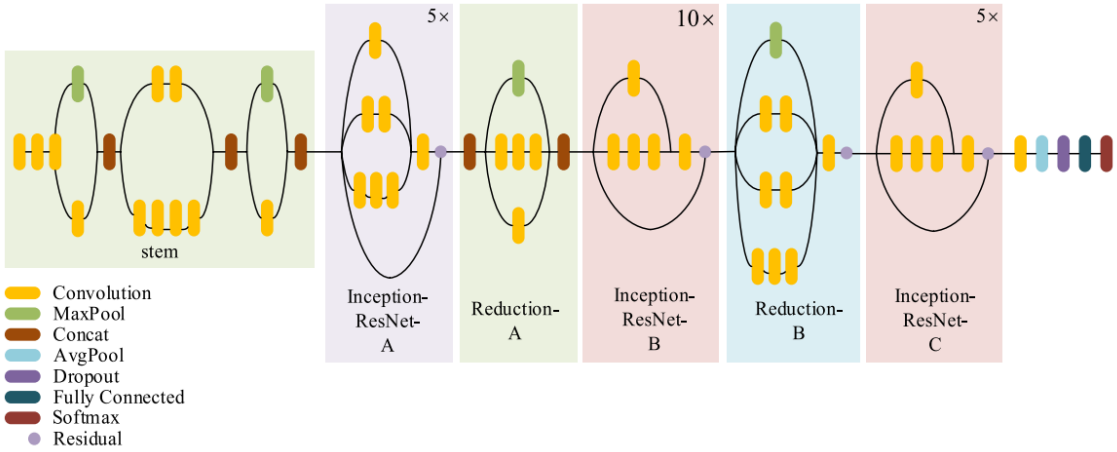


Figure 4.6: Schema of the Inception-ResNet-v2 network, from [88].

Deep belief network (DBN)

The DBN is a more complicated model proposed by Hinton et al. [38] that learns to extract a deep hierarchical representation of the training data. The traditional DBN is built with a stack of restricted Boltzmann machines (RBMs), which are two-layer generative stochastic models composed of a visible-unit layer and a hidden unit layer. In a DBN, the units in higher layers are trained to learn the conditional dependencies among the units in the adjacent lower layers, except the top two layers, which have undirected connections.

Liu et al. [53] presented a boosted DBN that iteratively performed feature representation, feature selection, and classifier construction in a unified loopy state from facial images.

4.4 Classification

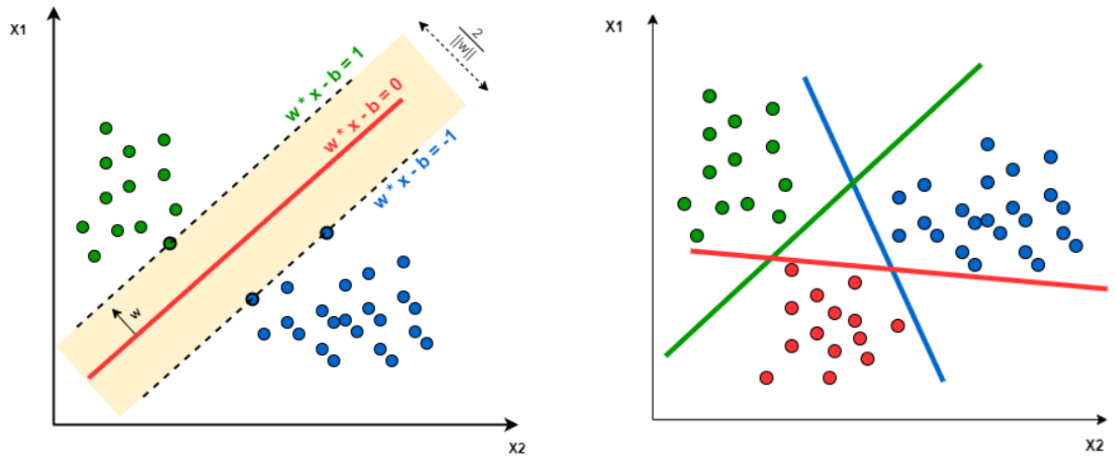
After the features are extracted, a decision method or classifier is needed, in order to do the proper recognition. As we have seen, handcrafted features necessarily need a classifier afterwards to execute the task. Despite the fact that deep learning models usually perform both feature extraction and classification in the same network, there is also the possibility of extracting features from a pre-trained model and then using them for classification. Therefore, this section intends to discuss the classification methods used in multimodal emotion recognition tasks, despite the way the features are obtained.

There exist several classification models suitable for emotion recognition, including K-nearest neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), CNNs, DBNs, and LSTMs [78]. In multimodal tasks, the most commonly found classifiers are the SVM, CNN, and

DBN as can be seen in Table 4.3.

SVM

A very common method found in the literature is by making use of a linear SVM. An SVM is a non-probabilistic binary linear classifier, which means that given a labeled training set, the SVM computes an optimal classification boundary that maximizes the width of the gap between the two categories, as can be seen in Figure 4.7(a). This can be broadened into a multiclass SVM, by using either a one-to-one approach - which breaks down the multiclass problem into multiple binary classification problems and a binary classifier is applied per pair of classes - or a one-to-rest approach - which breakdown is set to a binary classifier per each class [8]. Some data might not be correctly classified in low dimension space due to its non-linearity, and the sample space is mapped onto a high dimensional space with kernel functions, in which the sample can be linearly separable.



(a) (b)

Figure 4.7: Schematic representation of a) a binary and b) a multiclass SVM, from [8].

Valstar et al. [57] have used extracted features from face video, audio clips, ECG, and EDA signals and then used an SVM for classification followed by a late fusion strategy, having obtained a concordance correlation coefficient of 0.661. In the works of Liu et al. [87] and Tang et al. [36], an SVM was used to classify the features extracted from EEG and eye gaze after two different early fusion strategies, obtaining accuracies from 91% to 94% on SEED dataset and around 83% on DEAP dataset. Huang et al. [94] have also used SVM to obtain predictions from EEG signals, which then fused with the predictions from face information, lead to an accuracy from 81% to 83%, depending on the fusion strategy. Wei et al. [90] have used different SVMs to classify from EEG, ECG, RV, and GSR features and then merged them with a weighted strategy, achieving an accuracy of 84.62%. Finally, Zhongmin et al. [88] used SVM for classifying prediction from early fused features of ECG, SCL, tEMG and face video, reaching an accuracy of 80.89% on BioVid Emo DB and 84.09% on DEAP datasets. A summary of these works regarding modalities,

classification, fusion strategies, and results can be found in Table 4.3.

MKL

Multiple Kernel Learning (MKL), as used in [74], aims to learn the kernel in an SVM from training data in an optimized way. Using video, audio, and text modalities, this method achieved an accuracy of 76.85%.

CNN

CNNs, as previously described, allow for both feature extraction and classification. The work of Huang et al. [94] differs from the previously presented ones since it uses a set of fully connected layers, as seen in Figure 4.8, to classify from previously extracted handcrafted facial features, which are then fused with the predictions from the ECG signal, leading to an accuracy from 81% to 83%, depending on the fusion strategy.

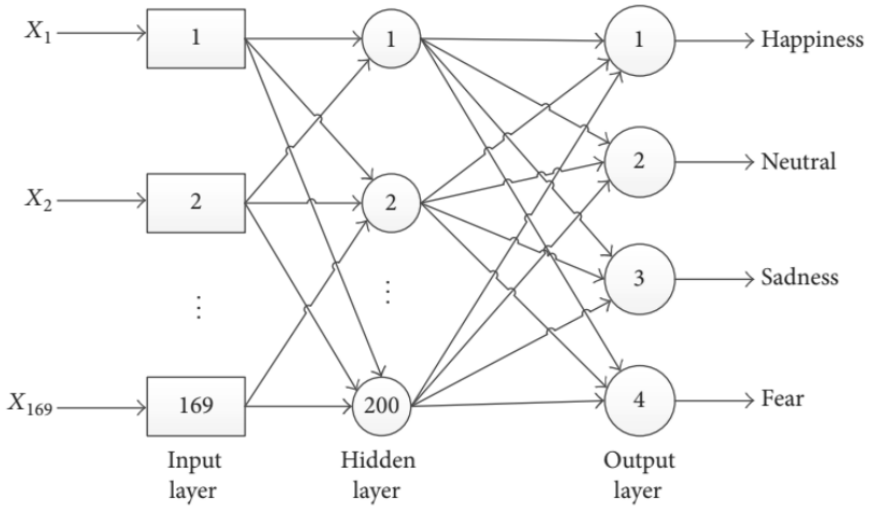


Figure 4.8: The architecture of the proposed system for face expression classification with one hidden layer with 200 neurons and whose output is the scores of four emotion states (happiness, neutral, sadness, and fear), from [94].

DBN

DBN was used by Nguyen et al. [16] in order to classify emotional state. Therefore, the DBN was treated as a multilayer perceptron (MLP), in which its hidden layers are built as RBMs with the principle as follows: the first layer RBM is treated as the input of the network; the hidden layer of one RBM represents the input of the following RBM and the last hidden layer RBM becomes the output of the network. A logistic regression classification was then added to the top of the network to make the prediction based on the output of the last hidden layer. This method has achieved an accuracy of 89.39%.

4.5 Fusion

As seen in Chapter 2.3.3, fusion can come either before or after classification, named early and late fusion, respectively. Early fusion has been considered a popular technique in order to merge all modalities into a single high-dimensional feature vector for the classifier since it joins all meaningful information together. Nevertheless, this fusion approach still has some problems, including the generation of high dimensional features on small datasets, leading to stress on the computational resources with respect to training and evaluating the classification model. For that reason, other researchers opt for late fusion strategies, which have also shown promising results.

The surveyed papers presented in Table 4.3 have shown very different fusion strategies that will be further discussed.

4.5.1 Early fusion

Starting with early fusion strategies, these go from a simple concatenation of the features as in Poria et al. [74] to complex deep learning methods such as DBNs, deep auto encoders (DAEs), and long-short term memory (LSTM) networks. Either one of these deep learning methods has shown that the increase in complexity resulted in better results in terms of accuracy in comparison to simple concatenation because the different modalities can be redundant, so concatenating features might not be efficient.

In Zhongmin et al. [88], the authors have used DBNs in order to merge the previously extracted features. More specifically, they have used three DBNs: the first one to merge the handcrafted features from the three physiological signals (multimodal DBN) obtaining high-level physiological features, and the second one to merge the handcrafted features with the deep features from face images (bimodal DBN) obtaining high-level facial features, and the final one to merge the high-level physiological and facial features (bimodal DBN) obtaining the final features to use for classification.

Liu et al. [87] and Tang et al. [36] are two subsequent works that have used different deep learning methods for early fusion not yet discussed: DAE and LSTM.

Deep autoencoder (DAE)

DAE was first introduced in [37] to learn efficient codings for dimensionality reduction. In contrast to the previously mentioned networks, which are trained to predict target values, the DAE is optimized to reconstruct its inputs by minimizing the reconstruction error. The structure of the bimodal DAE in Liu et al. [87] can be seen in Figure 4.9. The middle layer contains the high-level concatenated features to be used for classification.

Long-short term memory (LSTM)

Long-short term memory (LSTM), introduced by Hochreiter & Schmidhuber [39], is a special form of the traditional recurrent neural network (RNN) that is used to address the gradient vanishing and exploding problems that are common in training RNNs [51]. These can process

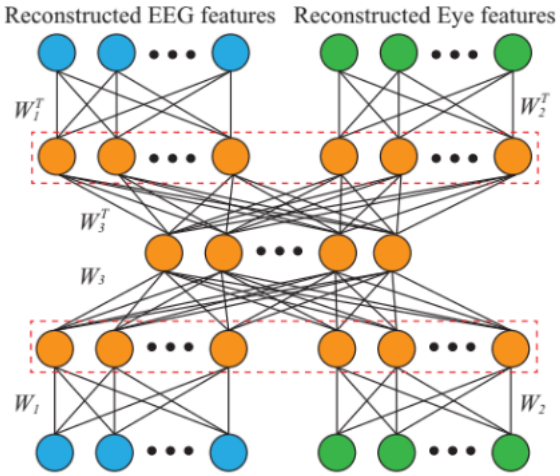


Figure 4.9: Structure of bimodal DAE, from [87].

entire sequences of data without treating each point in the sequence independently, but rather, retaining previous useful information to help with the processing of new data points.

Firstly, to understand the LSTM, it is important to define two concepts. The first is the cell state, corresponding to the current long-term memory of the network. The second is the hidden state, which represents the output at the previous point in time. In summary, LSTMs use three gates: forget gate, input gate and output gate. The forget and the input gate use the information from the previous hidden state and the new data point in the sequence to determine which pieces of the cell state should be forgotten and what new information should be added to it. Then, the output gate decides the new hidden state [40; 95]. The architecture of the typical LSTM block can be seen in Figure 4.10.

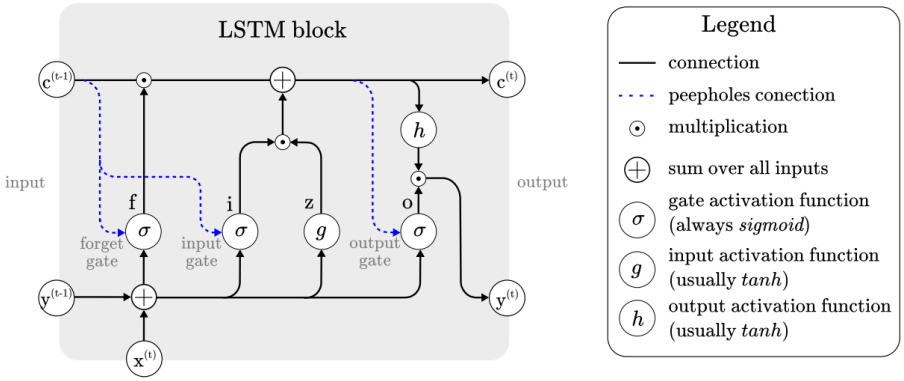


Figure 4.10: Architecture of the typical LSTM block, from [40].

Since the LSTM can model long-term dependencies in a sequence, it becomes very helpful for the task since there is either a sequence of ECG segments and/or video frames to be fused together. This advantage shows in the fact that Tang et al. [36] not only achieved the best accuracy out of the early fusion strategies but of the late fusion ones as well for the SEED dataset.

4.5.2 Late fusion

Moving to late fusion strategies, the ones observed in the literature were linear regression, sum and product rules, and score/weighted strategies.

As seen in section 2.3.3, the sum rule sums the confidences for a given class from each classifier and chooses the one with the highest confidence as output, whilst the product rule works like the sum rule but the confidences are multiplied instead. These two methods have achieved similar results in the work of Huang et al. [94].

Nguyen et al. [16] has used a score level late fusion which computes the likelihood in relation to the individual likelihoods from video and audio and weights regarding the number of audio and video clips. Wei et al. [90] also considers weights multiplied by the classifiers' probabilities. These weighted methods seem to surpass other late fusion strategies.

Valstar et al. [57] uses the probabilities of the different classifiers merged together using linear regression for classification. The difference between this work and the previous two is that while the latter estimates the values of the coefficients according to the set of presented inputs and output, the former previously calculates these coefficients.

Table 4.3: Summary of current available algorithms for multimodal emotion recognition regarding classification and fusion and the obtained results. Legend: CCC - concordance correlation coefficient; A - accuracy.

Paper	Year	Modalities	Dataset	Emotional Model	Classifier	Fusion Technique	Results
Valstar et al. [57]	2016	Video, audio, ECG, EDA	RECOLA	Discrete	SVM	Late (linear regression)	CCC = 0.661
Poria et al. [74]	2016	Video, audio, text	USC IEMOCAP	Discrete	MKL	Early (concatenation)	A = 76.85%
Liu et al. [87]	2016	EEG, eye signal	SEED, DEAP	Dimensional	SVM	Early (BDAE)	SEED: A = 91.01% DEAP: A = 83.25%
Tang et al. [36]	2017	EEG, eye signal	SEED, DEAP	Dimensional	SVM	Early (LSTM)	SEED: A = 93.97% DEAP: A = 83.53%
Nguyen et al. [16]	2017	Audio, video	eNTERFACE	Discrete	DBN	Late (weights)	A = 89.39%
Huang et al. [94]	2017	Face, EEG	Private	Discrete	CNN, SVM	Late (sum & product)	Sum: A = 81.25% Product: A = 82.75%
Wei et al. [90]	2018	EEG, ECG, RV, GSR	MAHNOB-HCI	Discrete	SVM	Late (weights)	A = 84.62 %
Zhongmin et al. [88]	2020	ECG, SCL, tEMG, face video	BioVid Emo DB, DEAP	Dimensional, Discrete	SVM	Early (MDBN, BDBN)	BioVid: A = 80.89% DEAP: A = 84.09%

4.6 Conclusion

In this chapter, various studies on the state-of-the-art have been analysed, allowing some conclusions to be drawn.

Regarding feature extraction, handcrafted methods are much more present in the literature, especially for the ECG signals. These handcrafted methods implicate expertise from the researchers since it is up to them to define the best features to extract for the task. With the ECG being a more recently explored modality, it is possible that the datasets containing this signal are smaller, therefore difficulting the learning ability of deep learning methods.

Nevertheless, deep learning methods seem to be the current state-of-the-art, with more recent works exploring this alternative since they are able to achieve better results. Also, since the features are automatically recognized from the data, it does not require such knowledge about the signals from the researchers. Moreover, it is shown that feature selection is a useful method by removing unnecessary or irrelevant features and enabling the high performance of the developed algorithms, both for handcrafted and deep extracted features.

Concerning classification, multimodal approaches show a clear preference for SVM classifiers.

Finally, looking into fusion methods, both early and late fusion strategies have shown promising results. However, since the surveyed papers not only present different fusion strategies but also modalities and classifiers, it is difficult to extract direct comparisons.

In conclusion, this dissertation will focus on the development of a robust and accurate emotion recognition model, making use of both handcrafted and deep learned features and tested on different classifiers and fusion techniques.

Chapter 5

Implementation of Prior Art Baselines

5.1 Introduction

After analysing the most common methods applied for emotion recognition using different modalities, the initial work developed focused on attempting to replicate some of those methods as the basis for the following work. Furthermore, besides replicating the approaches, some alterations needed to be made due to missing information or different programming languages used, which will also be described further.

In this chapter, the model of Zhongmin et al. [88] was chosen to obtain a baseline for emotion recognition both from ECG signals and face video due to its promising results while combining the desired modalities. Like the authors, the BioVid Emo DB [99] was used for the development of this work. In the article, an emotion recognition model was proposed using three different physiological signals together with facial videos which can be seen in Figure 5.1. For this work, both SCL and tEMG signals were disregarded, hence keeping a higher simplicity in comparison to the methodology proposed by the authors.

5.2 Methodology

5.2.1 ECG

Pre-processing

The authors used MATLAB to pre-process the psychophysiological signals but the same steps were recreated using Python, the programming language used throughout this work. A filtering process was implemented to filter out any noise or unwanted frequencies. First, a 2^{nd} order Butterworth band-stop filter was used with cut-off frequencies of 49 and 51 Hz to remove the powerline interference. Second, the authors proposed a 2^{nd} order Butterworth low-pass filter with a cut-off frequency of 35 Hz followed by wavelet packet decomposition to remove the baseline wander. These were combined and replaced by only one step, using a 2^{nd} order Butterworth band-pass filter with cut-off frequencies of 1 and 35 Hz, which achieved a similar effect since the

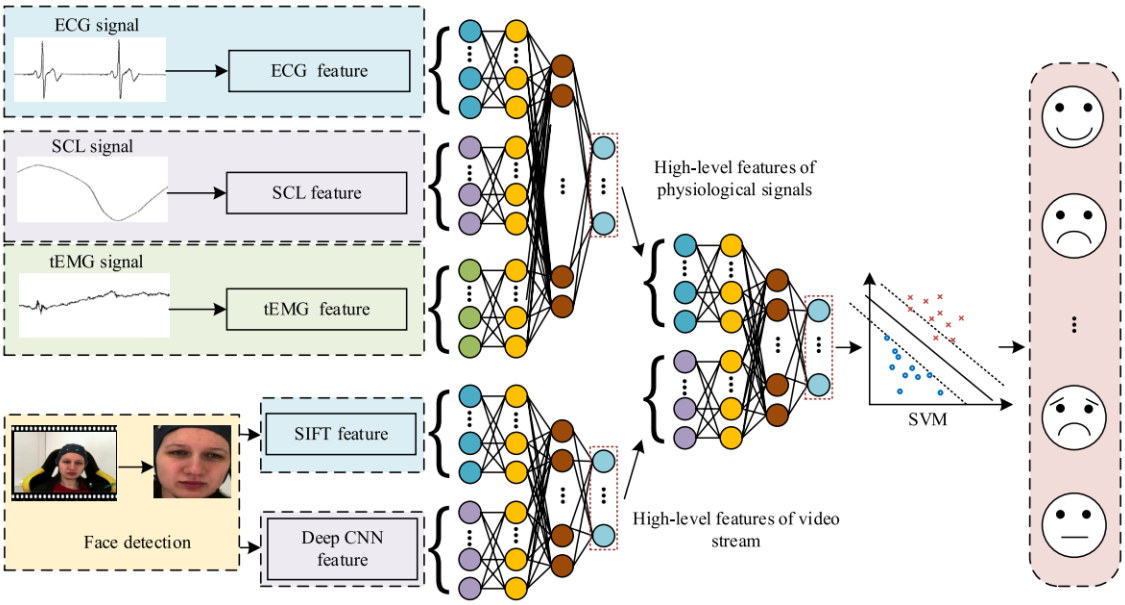


Figure 5.1: The MDBN-based model proposed by Zhongmin et al. that combines sentiment features from video and multiple psychophysiological signals, from [88].

baseline wander is a low-frequency noise.

Feature extraction

After pre-processing the signals, these were divided into different windows and the features were computed per window. There was no information regarding the size of the windows, only that the step size was equal to 40 ms. This indicated that the windows should be relatively small, otherwise the overlap would be so large that they would be identical. However, in the literature, some works have found larger windows to be most suitable, since the emotions portray themselves in the ECG signals as gradual in time. Therefore, four different combinations of window and step sizes were calculated and compared:

1. the step size of 40 ms from the article was respected with a 100 ms sized window, resulting in a 60 ms overlap;
2. three larger windows all with 50 % of overlap:
    2.1. 1 s window with 50 ms of overlap;
    2.2. 5 s window with 2.5 s of overlap;
    2.3. 10 s window with 5 s of overlap.

For each window, the time domain features computed were the mean, median, maximum, minimum, range, variance, and standard-deviation; and the frequency domain features computed were the amplitude, skewness, kurtosis, power spectral density (PSD), and differential entropy (DE). The amplitude feature was assumed as the maximum value for the magnitude of the fast

Fourier transform (FFT) of the input signal, while the skewness, kurtosis, and differential entropy were calculated from the magnitude of the FFT as well. Finally, these features were standardized.

Classification

The classifier used by the authors was the SVM, which is widely used in emotion recognition tasks as seen in Chapter 4. A C-Support Vector Classification (C-SVM) is the type of SVM that was used by the authors and therefore recreated in this work, with a linear kernel type and a cost (C) value of 1.

The SVM with these parameters was used throughout the different experiments in this chapter and the overall pipeline for the ECG part can be seen in Figure 5.2, which was adapted from Figure 5.1 to obtain only the steps involved with the ECG.

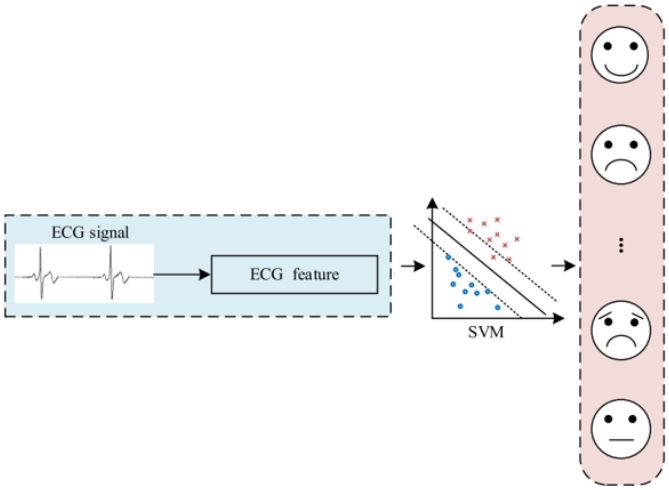


Figure 5.2: ECG pipeline adapted from the model proposed by Zhongmin et al. seen in figure 5.1.

Since no information was given regarding how to aggregate the features or predictions for the different windows, the classification was also done in four different ways:

1. classifying each window individually, assuming the label of each window equal to the label of the original full-length signal;
2. aggregating the features of the different windows from the same signal using the mean value of each feature prior to classification;
3. aggregating the features of the different windows from the same signal using the mean and the standard-deviation values of each feature prior to classification;
4. classifying each window individually and then aggregating the predictions to generate one final prediction per full-length signal.

5.2.2 Face

Pre-processing

In the video pre-processing, the authors extracted all frames from the video stream. These frames are sequential and temporally continuous in the video data. However, that turned out to be very computationally expensive for further steps, besides resulting in many identical frames. Therefore, only 5 frames per second were extracted from the videos, instead of 25 frames per second.

Afterward, the face in each frame was detected and cropped to match the input size of the Inception-ResNet-v2 network. This was achieved using the pre-trained MTCNN from facenet-pytorch [1].

Feature extraction

To extract features from each frame previously obtained, two different approaches were used: the first one was handcrafted and consisted of extracting scale-invariant feature transform (SIFT) features; the second one was to use the Inception-ResNet-v2 network to extract deep features.

The SIFT features are widely used in computer vision and consist of a set of key points and respective descriptors that are calculated from the image. Here an issue aroused: since the frames turned out to have a very different number of key points, classification could not handle them. A well established solution for this problem without disregarding any of the data is called Bag of Visual Words (BoVW), which consists of aggregating the key points' descriptors in clusters and counting the number of key points in each cluster in order to make a histogram of the frame. The clustering operation was done using the k-means algorithm, with k equal to 50 (number of classes multiplied by 10).

The deep features resulting from the Inception-ResNet-v2 network were considered to express emotion-relevant information in the video. The Inception-ResNet-v2 network used by the authors was mainly based on the TensorFlow network available pre-trained on ImageNet dataset. The authors fine-tuned it for a face recognition task and then extracted the deep features from different layers to test on an emotion recognition model. The features that achieved the best result were extracted from the fully connected layer. In this work, to avoid training the network for an additional task other than emotion recognition, the features from the same pre-trained model were extracted and used directly for the emotion recognition model. This was then compared with the features extracted from the Inception-ResNet-v1 from facenet-pytorch [1], which is already pre-trained on a face recognition dataset, VGGFace2. From the Inception-ResNet-v2, a total of 1536 features were extracted, while from the Inception-ResNet-v1, a total of 1792 features were obtained.

[1] `https://github.com/timesler/facenet-pytorch`

Classification

Due to the huge number of frames, using an SVM as a classifier was impractical. Therefore, a multilayer perceptron (MLP) was used for classification. For the deep extracted features, this consisted of a linear layer from Pytorch with 5 neurons, which received 1536 or 1792 input features, depending on if coming from the Inception-ResNet-V2 or Inception-ResNet-V1. The number of neurons correspond to the number of classes and are followed by a softmax activation. For the handcrafted features, the same structure was maintained but the classifier received 50 input features instead. However, to evaluate the effect of the number of layers in the classifier, an MLP with four linear layers was also used. The three hidden linear layers are followed by a ReLU activation function and a dropout layer with a probability of 0.2, and the last linear layer, also know as output layer, is followed by a softmax activation. The classifier was trained for 1000 epochs, with a batch size of 100 samples, and a learning rate of 10^{-4}, having Adam as optimiser and cross entropy loss to compute the loss between the inputs and the targets. An early stopping strategy was also implemented based on the validation loss, and the model which presented the lower validation loss was kept.

The overall pipeline for the face part can be seen in Figure 5.3.

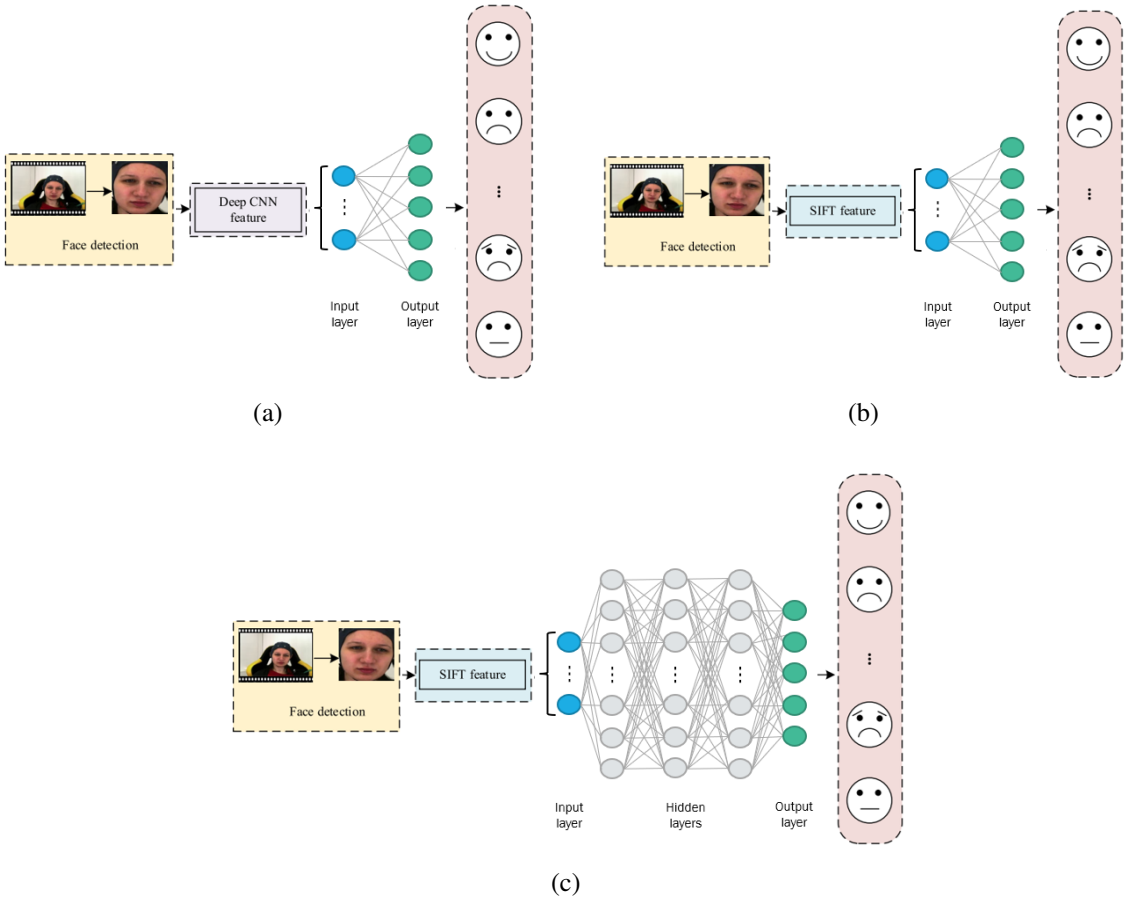


Figure 5.3: Video pipeline for a) deep features, and SIFT features having as classifier a b) one-layer MLP or a c) four-layers MLP.

5.2.3 Data division

Data division can be performed in several ways, such as:

1. **random:** many approaches in the literature perform a random division of the examples by considering the segments/frames as the inputs of the classifiers and shuffling the data before the division. This results in information of the same signal/video being present in both sets.
2. **signal-independent:** this works by first assigning each recording to their set and only then dividing it into segments/frames. This is a more realistic scenario since it ensures that information from the same recording is not used in both sets, but it is also more challenging.
3. **subject-independent:** this setting works similar to the previous one but also ensures that recordings from the same subject are not present in both partitions of the dataset, being the most strict scenario.

Alike the authors, a random division of the dataset was performed, by considering 80% of the data for training and 20% for testing. On the one hand, for the ECG segments, this division was enough since the parameters chosen for the classifier were the ones used by the authors. On the other hand, for face frames, a validation set was necessary to train the MLP, considering 10% of the training data for validation.

5.3 Results

5.3.1 Dataset overview

Prior to discussing the results obtained for each modality, it is important to understand the characteristics of the data in terms of the number of examples for each class both in the original dataset and after pre-processing - *i.e.* after the division in segments/frames. This information may be found in Table 5.1 and it was achieved by considering the duration of the ECG segments equal to 10 s.

Table 5.1: Number of examples for each class both in the original dataset and after pre-processing.

	Number of Records					
	Amusement	**Anger**	**Disgust**	**Sad**	**Fear**	**Total**
ECG signals	86	86	86	86	86	430
Face videos	86	86	86	85	86	429
ECG segments	976	1252	324	1657	1594	5803
Face frames	50561	63940	18870	81710	80580	295661

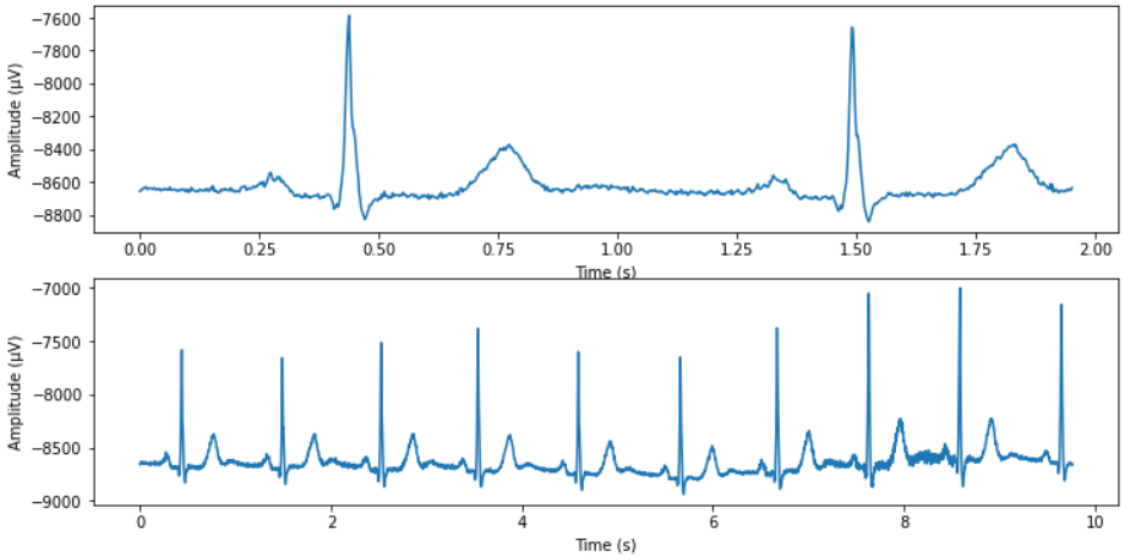
From the table above it can be observed that the original dataset contained the same number of recordings for each class with an exception on the face modality due to one corrupted and therefore unavailable video.

After pre-processing, the average length of recordings for the different classes highly influences the number of examples. On the extremes, the average length for the fear class is around

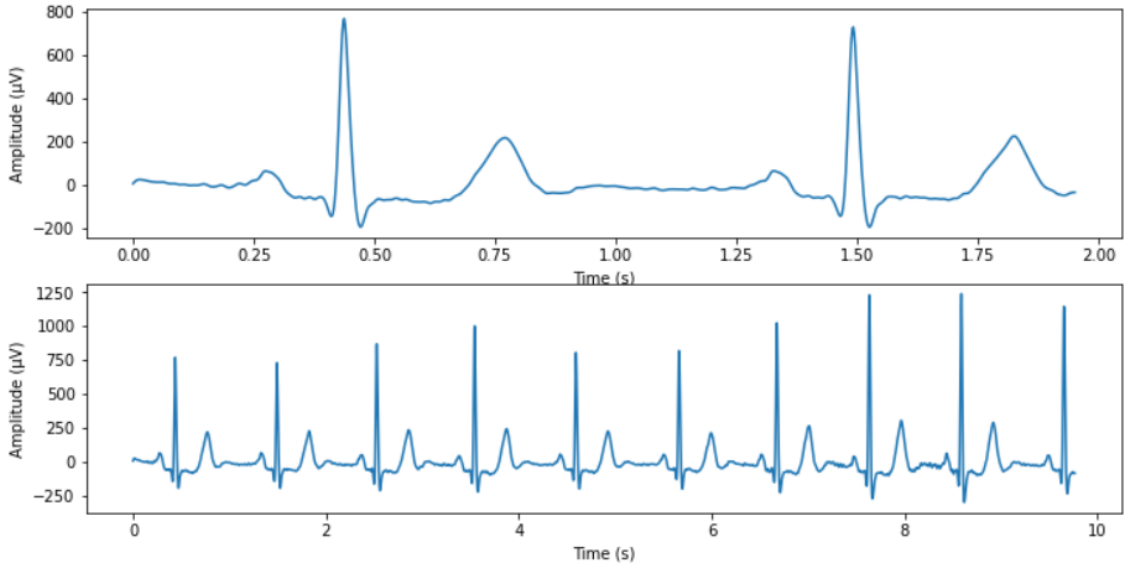
180 seconds and for the sad class is 143 seconds, while for the disgust class in only 56 seconds. This leads to class imbalance. Note also that for each 10 s segment there are approximately 50 corresponding frames, leading to a much higher amount of data for the face modality after pre-processing.

5.3.2 ECG

Visually, the filtering process was successful in pre-processing the ECG signal and an example can be seen in Figure 5.4.



(a)



(b)

Figure 5.4: Example of pre-processing of an ECG signal with a) representing a the same raw signal in two different scales and b) the result from filtering represented in both scales.

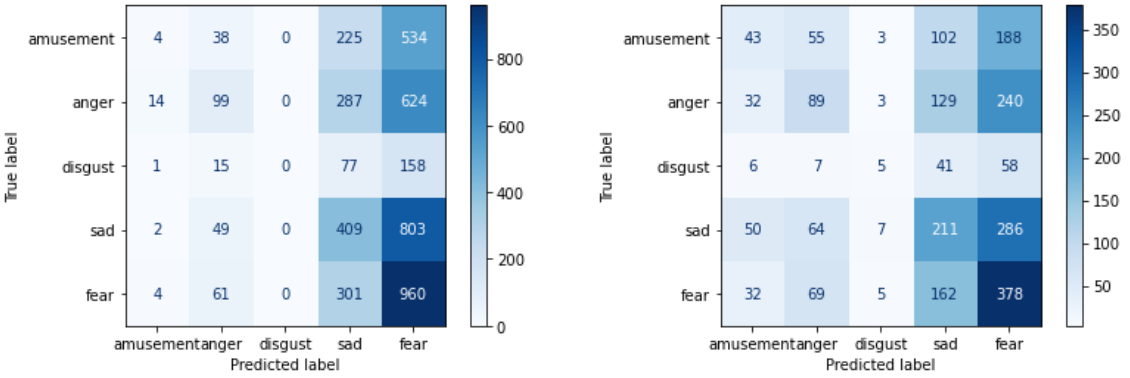
The classification results can be seen in Table 5.2 for the different window sizes, comparing them with the results obtained by the authors. Note that the results are only available for two of the four options of window sizes described in the methods due to the fact that the fit time of the SVM scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples, which is the case of the other two options.

Table 5.2: Classification results obtained using the different windows individually.

Window		**Accuracy**		**Individual Accuracies**				
Size	**Overlap**	**Train**	**Test**	**Amusement**	**Anger**	**Disgust**	**Sad**	**Fear**
5 s	2.5 s	33.61 %	31.55 %	0.50 %	9.68 %	0 %	32.38 %	72.40 %
10 s	5 s	41.38 %	32.05 %	11.00 %	18.05 %	4.27 %	34.14 %	58.51 %
Zhongmin et al. [88]		-	54.21 %	64.70 %	57.32 %	12.21 %	43.75 %	93.07 %

Table 5.2 shows that the obtained accuracies fall short in comparison to the ones obtained by the article. Nevertheless, it points out that the longest-sized windows allow for a better classification of all classes, while the other fails to represent both amusement and disgust classes.

Figure 5.5 shows the different confusion matrices obtained from predictions of the test dataset and corresponding labels, which allows to see which classes is the model predicting the most.



(a) 5 s window, 2.5 s overlap.

(b) 10 s window, 5 s overlap.

Figure 5.5: Confusion matrices obtained using the different windows individually.

Overall, we can observe that the majority of the predictions are lying either in the fear or on the sad classes. A possible explanation for this is the data imbalance seen in subsection 5.3.1, which results in the model penalizing more the errors in this two classes, ultimately leading to a better classification of these.

This phenomenon can be solved by instructing the SVM to balance the weights of each class - *i.e.* to use the values of the labels to automatically adjust weights inversely proportional to class frequencies in the input data. This was performed in the following experiment.

Effect of balancing class weights

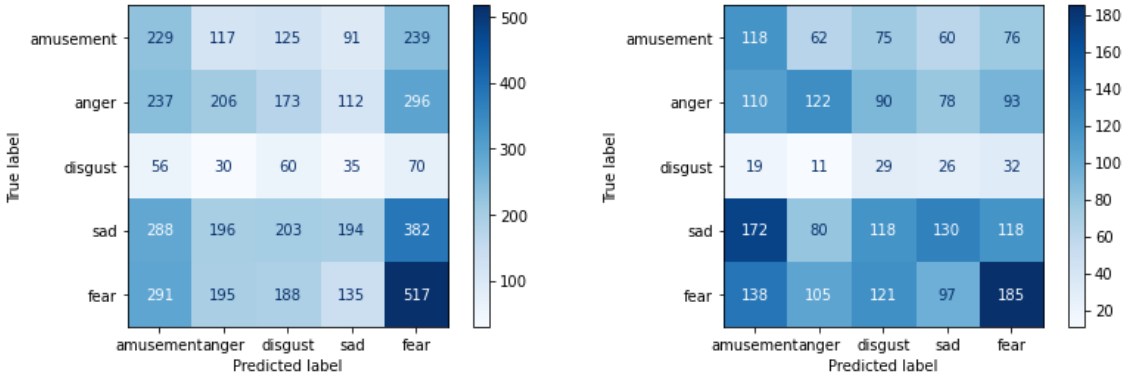
Table 5.3 and Figure 5.6 show the results in terms of accuracy and predictions distribution obtained with balancing the weights of each class.

It can be observed that, for both window sizes, all classes are identified. However, it has the drawback of not achieving such high overall accuracies since in the unbalanced experiment the high accuracies of the last two classes made up for the low accuracies in the others. Looking at the results from the authors, they seem to have chosen an unbalanced experiment as well, since there is such a high difference in performance between the different classes.

Given the results obtained with the windows separately, it brings the hypothesis that aggregating them may bring better results, since the emotion may not be present throughout all segments of each signal.

Table 5.3: Classification results obtained using the different windows individually, while balancing the weights of the classifier.

Window		**Accuracy**		**Individual Accuracies**				
Size	**Overlap**	**Train**	**Test**	**Amusement**	**Anger**	**Disgust**	**Sad**	**Fear**
5 s	2.5 s	29.31 %	25.85 %	28.59 %	20.12 %	23.90 %	15.36 %	38.99 %
10 s	5 s	38.05 %	25.78 %	30.18 %	24.75 %	24.79 %	21.04 %	28.64 %
Zhongmin et al. [88]		-	54.21 %	64.70 %	57.32 %	12.21 %	43.75 %	93.07 %



(a) 5 s window, 2.5 s overlap.

(b) 10 s window, 5 s overlap.

Figure 5.6: Confusion matrices obtained using the different windows individually, while balancing the weights of the classifier.

Effect of aggregating the predictions by majority voting

A post-model aggregation of the previously obtained predictions was performed by majority voting and the results may be found in Table 5.4. Note that this step is only reasonable when applied with a balanced classifier, otherwise the number of predictions from classes sad and fear would outshine those from the other classes.

Table 5.4: Classification results obtained using the different windows individually, while balancing the weights of the classifier, followed by majority voting of the predictions.

Window		**Accuracy**		**Individual Accuracies**				
Size	**Overlap**	**Train**	**Test**	**Amusement**	**Anger**	**Disgust**	**Sad**	**Fear**
5 s	2.5 s	33.26 %	26.24 %	31.76 %	22.09 %	22.50 %	16.28 %	38.37 %
10 s	5 s	56.74 %	28.57 %	34.15 %	34.89 %	25.00 %	16.27 %	31.76 %
Zhongmin et al. [88]		-	54.21 %	64.70 %	57.32 %	12.21 %	43.75 %	93.07 %

Looking at the results it can be concluded that this aggregation slightly benefits the predictions. However, it is not a significant increase since the model does not benefit from this aggregation to improve its predictions in the training phase.

Therefore, to attempt at obtaining better results, a pre-model aggregation was thought by merging the different segments using the mean and/or standard-deviation of each feature.

Effect of aggregating windows with mean and/or standard-deviation

Table 5.5 shows the results obtained by aggregating the different windows using both the mean value alone and the mean value together with the standard-deviation prior to classification.

Table 5.5: Classification results obtained using the mean and/or standard-deviation aggregation of features.

Window		**Accuracy**		**Individual Accuracies**				
Size	**Overlap**	**Train**	**Test**	**Amusement**	**Anger**	**Disgust**	**Sad**	**Fear**
Mean								
100 ms	60 ms	22.38 %	13.95 %	0 %	13.33 %	0 %	23.08 %	43.75 %
1 s	50 ms	29.65 %	4.65 %	0 %	0 %	0 %	15.38 %	12.50 %
5 s	2.5 s	83.14 %	13.95 %	13.04 %	6.67 %	10.53 %	30.77 %	12.50 %
10 s	5 s	97.09 %	12.79 %	13.04 %	20.00 %	10.53 %	7.69 %	12.50 %
Mean + Standard-Deviation								
100 ms	60 ms	24.13 %	8.14 %	4.35 %	13.33 %	0 %	15.38 %	12.50 %
1 s	50 ms	43.31 %	6.98 %	0 %	6.67 %	0 %	30.77 %	6.25 %
5 s	2.5 s	95.64 %	20.93 %	8.70 %	6.67 %	52.63 %	23.08 %	12.50 %
10 s	5 s	100 %	19.77 %	8.69 %	13.33 %	57.89 %	0 %	12.50 %
Zhongmin et al. [88]		-	54.21 %	64.70 %	57.32 %	12.21 %	43.75 %	93.07 %

These results point to large overfitting demonstrated by the high difference between train and test accuracies, meaning that the model is not generalizing well. Note that, while in the previous experiences the random data division lead to having segments of the same signal in both datasets, in this experience that does not happen since the division is performed after the aggregation with the mean and/or standard-deviation. Thus, the classifier is facing a signal-independent challenge,

which turns out to be quite difficult.

The C-SVM, as the name indicates, depends on the value of the C parameter, which is a regularization parameter whose aim is to balance overfitting and underfitting a model during training. The strength of the regularization is inversely proportional to C - *i.e.* a lower value of C will increase regularization, thus reducing the overfitting phenomena and vice-versa. Therefore, values of C ranging from 10^{-5} to 10^{2} were tested on the mean and standard-deviation features to observe the performance.

Effect of C-value

Figure 5.7 shows the results obtained by varying the C value for the 5 s windows. This window was chosen since it was the one which allowed best to visually represent the effect of the C-value.

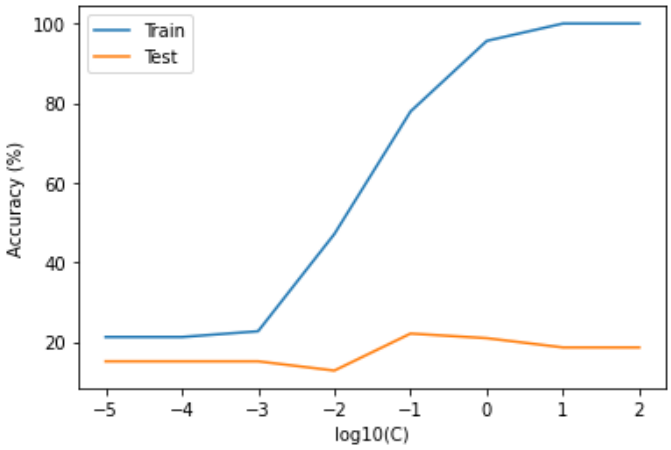


Figure 5.7: Train and test accuracies obtained by varying the value of C for 5 s windows.

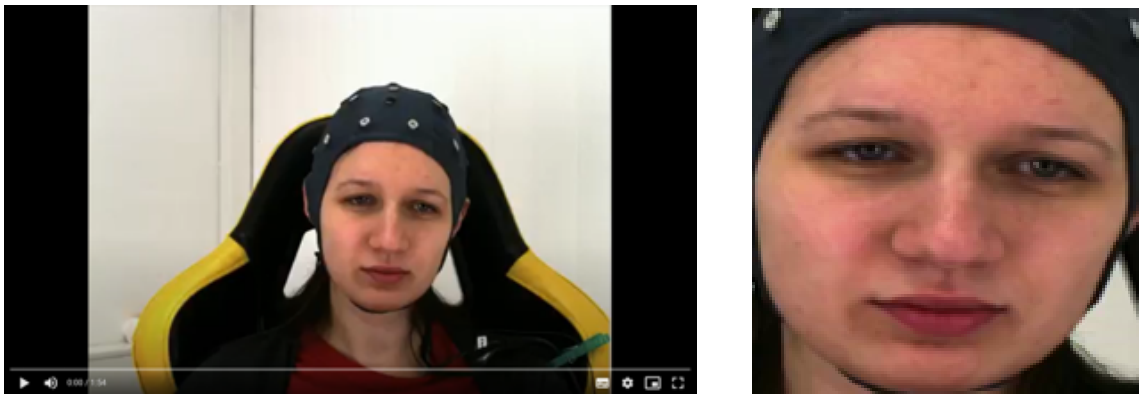
As expected, a lower C-value was able to reduce the overfitting phenomena, approximating the train and test accuracies. Also, C equal to 10^{-1} shows a slightly higher validation accuracy than the one observed before with C equal to 1. However, C equal to 10^{-1} is not enough to fight overfitting.

5.3.3 Face

The MTCNN turned out to be successful in detecting and cropping the faces in each frame and an example can be seen in Figure 5.8.

SIFT results and the effect of the number of layers

After extracting the SIFT features and applying the BoVW algorithm, these were evaluated using an MLP with one linear layer with a softmax activation. This MLP was chosen to allow a direct comparison with the results from the deep features, since the Inception-Resnet-V2 also has one fully connected layer in the end with a softmax activation for classification. Then, an

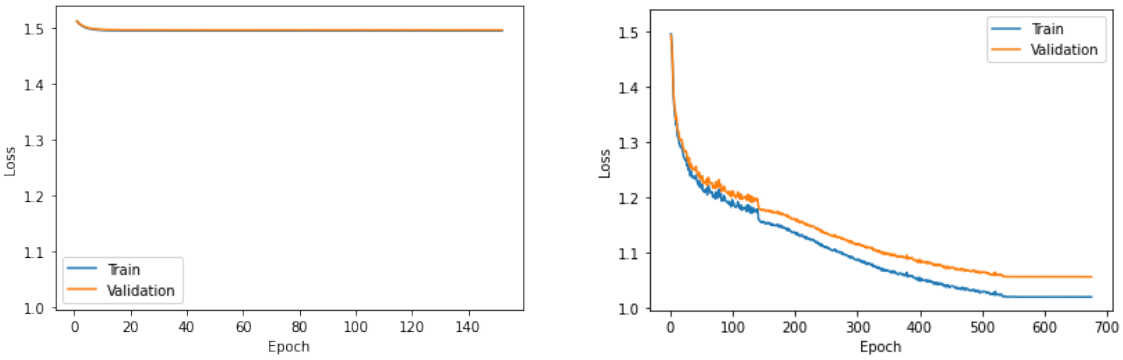


(a) (b)

Figure 5.8: Example of pre-processing face video with a) representing a full frame from a video with size 1038x1388x3 and b) the result from the MTCNN with size 299x299x3.

MLP with four fully connected layers as previously described in the methods section was used to evaluate the effect of the number of layers.

The results can be found in Figure 5.9 and Table 5.6. From these, it can be observed that the one-layer classifier shows a slight loss decrease during training. This is insignificant in comparison to the decrease observed in the four-layer model, which shows a higher capacity to learn from the training data, despite needing more training epochs to reach the best model possible due to its complexity. The latter also achieves a lower loss and higher accuracy both in the train, validation, and test datasets in comparison to the one-layer classifier.



(a) (b)

Figure 5.9: Train and validation loss evolution over the number of epochs using SIFT features for a) the one-layer MLP and b) the four-layers MLP.

Table 5.6: Classification results obtained using the SIFT features.

Number of layers	**Accuracy**			**Individual Accuracies**				
	Train	**Validation**	**Test**	**Amusement**	**Anger**	**Disgust**	**Sad**	**Fear**
1	31.17 %	31.03 %	31.15 %	2.68 %	3.08 %	0 %	57.83 %	51.45 %
4	63.39 %	61.07 %	61.19 %	56.67 %	59.70 %	6.01 %	69.39 %	69.54 %

In the work of Zhongmin et al. [88], no results are given for the SIFT-only classifier. Nevertheless, the obtained results are relatively satisfactory, with around half of the examples being classified correctly. However, these characteristics alone do not fit to the disgust category.

Deep results and the effect of pre-training

The results from the extracted deep features were obtained from a one-layer MLP as previously described and can be found in Table 5.7.

Table 5.7: Classification results obtained using the features extracted from Inception-Resnet-V2 and Inception-Resnet-v1, pre-trained on ImageNet and VGGFace2 datasets, respectivelly.

Network	Accuracy			Individual Accuracies				
	Train	Validation	Test	Amusement	Anger	Disgust	Sad	Fear
Inception-ResNet-V2	45.76 %	45.54 %	45.45 %	46.31 %	34.24 %	2.55 %	53.95 %	55.21 %
Inception-ResNet-V1	70.41 %	69.69 %	69.77 %	81.04 %	63.67 %	34.69 %	73.30 %	72.36 %

As previously described in the methods section, the authors extracted the features from the pre-trained Inception-Resnet-V2 and classified them with an emotion recognition classifier, obtaining an accuracy of 59.42%. So, these can be compared with the results indicated in the first row of Table 5.7. The accuracies obtained, although lower, are satisfactory since there is no information if the authors trained only the fully connected part of the network for the emotion classification, or fine-tuned the convolutional layers as well.

Then, the authors fine-tuned the Inception-ResNet-V2 for a face recognition task and extracted features from it to perform emotion recognition, obtaining an accuracy of 70.83%. In this work, the Inception-ResNet-V1 already pre-trained for a face recognition task was used in replacement to extract the features for emotion recognition. Therefore, these results are equivalent and it can be observed that similar results were obtained while saving an additional computational step, even though the Inception-ResNet-V1 is a prior version to the Inception-ResNet-V2. For that reason, the features extracted from the Inception-ResNet-V1 will be used for further improvements throughout this work.

5.4 Conclusion

This chapter focused on analysing the state-of-the-art approach of Zhongmin et al. [88] and replicating it, serving as a baseline for further work. As it can be concluded, missing information in the original article made it difficult to accurately replicate the reported results.

The results obtained from the ECG signals were the most different in comparison with the work of Zhongmin et al. [88]. This could be due to the number of assumptions that needed to be made. The face results, although a direct comparison with the paper could not be performed for the SIFT features, showed promising results, with good accuracy values by itself. Therefore, further efforts should be concentrated on the ECG signal prior to fusion.

Chapter 6

Improvements on Emotion Recognition from ECG Signals

6.1 Introduction

The initial work developed focused on attempting to replicate some state-of-the-art methods, which achieved weaker results regarding emotion recognition from ECG signals relative to the reported performance. Therefore, this chapter describes some improvements that were explored including deep feature extraction, merging these with the previously handcrafted features, and, finally, temporal aggregation of the segments of each signal by making use of an LSTM to obtain only one prediction per signal.

6.2 Methods

Deep feature extraction and classification using MLP

Sarkar et al. [65] has proposed a self-supervising learning methodology for emotion recognition from ECG signals in order to tackle the problems of fully-supervised methods - *i.e.* having a model trained from scratch for every classification task. The proposed methodology can be divided into two stages of learning:

1. First, ECG signals undergo six different transformations: noise addition, scaling, negation, temporal inversion, permutation, and time-warping. Then, a network is used to learn robust and generalized features from unlabelled ECG through recognition of these signal transformations.
2. Then, the original ECG signals and the emotion labels are used to train a network for emotion recognition, with the frozen convolution layers of the previous network and new fully connected layers to perform the emotion classification.

This methodology can be seen in Figure 6.1 and, in the end, it produces a total of 128 features.

Recently, Antunes [5] has improved this methodology by adding two additional transformations: baseline wander and magnitude warping. The self-supervised network she obtained was used in this experiment to extract deep features from the ECG signals.

To fit the inputs expected by the network, the ECG signals had to undergo a resampling process from 512 Hz to 256 Hz, a z-score normalization - *i.e.* normalizing the signal so that the mean of all its samples became equal to 0 and the standard-deviation equal to 1 - and were divided in 10 s windows with no overlap.

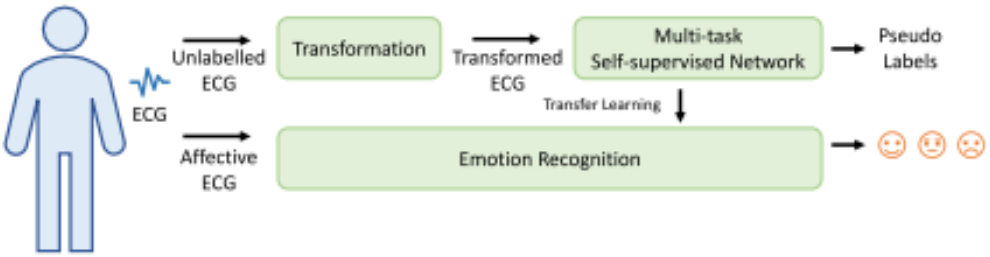


Figure 6.1: An overview of the proposed framework for self-supervised emotion recognition presented by Sarkar et al., from [65].

For classification, an MLP with four linear layers was used. The three hidden layers were followed by a ReLU activation function and a dropout layer, while the output layer was followed by a softmax activation. The MLP received 128 input features and all hidden layers had 512 neurons, which is four times the number of features it received as input. The output layer received the 512 values returned by the previous hidden layer and had 5 neurons, equal to the number of classes. This classifier was trained for 200 epochs, with a batch size of 100 samples, and a learning rate of 10^{-4}, having Adam as optimiser and cross entropy loss to compute the loss between the inputs and the targets. An early stopping strategy was adopted, and the model which presented the lower validation loss was saved. This training details remained the same throughout the following experiments of the present chapter.

Regarding data division, since this experiment considers the individual segments as inputs of the classifier, it was performed in a random way, by considering 80% of the data for training and 20% for testing as in section 5.2.3. Also, 10% of the training data was used for validation.

Classification from handcrafted features using MLP

Up to this point, two different sets of features were obtained: handcrafted ones, as described in Chapter 5, and deep ones, as just explained. To fairly compare the results obtained from deep features with handcrafted ones, the latter was classified using an MLP which received the 28 input features and presented the same structure as before: four linear layers, all three hidden layers had 112 neurons (four times the number features), and the output layer received 112 values and had 5 neurons, equal to the number of classes.

Note that for the calculation of these handcrafted features, the signals could not be normalized with a z-score normalization, otherwise the mean and standard-deviation features would be

pointless. Therefore, these features were calculated from the signals prior to this step of pre-processing and standardized just as in Chapter 5.

In this experiment, the data division was performed in the same way as the previous one.

Concatenation of deep and handcrafted features

The deep extracted features were merged with the handcrafted ones by means of a simple concatenation to evaluate which set of features performed the better: handcrafted only, deep only or the concatenation of both. Before entering the classifier, the concatenation of these features was standardized by removing the mean and scaling to unit variance so that differences in scales between the two would not interfere with the learning of the model. The mean and standard-deviation were extracted from the training data and used to standardize both validation and test data.

The concatenation of both types of features resulted in 156 input features. The classifier was developed following the same strategy as before with the hidden layers having 624 output neurons, and the output layer receiving 624 values and presenting 5 neurons.

In this experiment, the data division was performed in the same way as the previous two.

Temporal aggregation of segments with LSTM

Since most techniques rely on emotion segments rather than the complete signal, they are compelled to take the same label into account for each segment. This practice, however, can be misleading because there is no guarantee that the emotion is distributed evenly over the entire signal as opposed to being centered on specific areas. Having said that, there is a high motivation for methodologies that aggregate segments that are part of the same signal and predict emotions for the entire signal rather than for individual segments. In order to do this, heuristic methods such as the mean, median, maximum, among others, could have been used. However, these treat all the segments of the signal equally. Therefore, more complex methods can be studied to learn in a more intelligent way, ignoring some of the segments and giving more attention to the most relevant ones. That being said, the various signal segments were aggregated using a bidirectional LSTM, which is useful for learning sequential temporal information and may be of relevance concerning ECG data. In summary, the bidirectional variant of the LSTM consists of two layers: in the first layer, the input sequence flows forward; while in the second layer, the input sequence flows backward. The outputs from the two LSTM layers are then concatenated.

Being a bidirectional LSTM, there are several ways to use it to perform aggregation and even classification. By defining its hidden size as equal to the number of categories, it outputs double the number of classes. Therefore, either the first five (forward) or the last five (backward) outputs can be used followed by a softmax activation to serve as a classifier as well. Alternatively, all 10 outputs may be used, followed by a linear layer that will map them to the number of categories in the data using 5 neurons. To prevent overfitting, dropout can be used between the LSTM outputs and the linear layer. The schematic representation of this last method can be seen in

Figure 6.2. Note that this step was performed with the set of features that allowed to obtain a better performance in the previous experiments.

Also note that in this experiment, since the division of the data is applied to the full-length signals instead of the individual segments, the train, validation and test datasets are signal-independent as described in section 5.2.3, which brings an additional challenge to the model. The percentages of data for each partition remained the same.

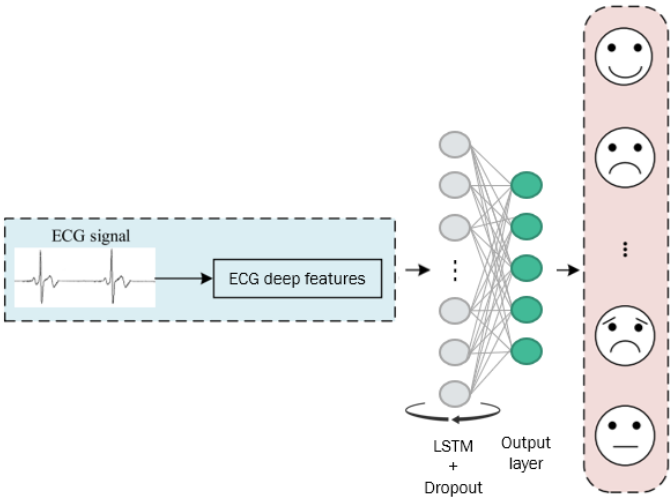


Figure 6.2: Schematic representation of the ECG pipeline adopted to aggregate the features of the different segments using an LSTM, followed by dropout and a linear layer for final classification.

6.3 Results

Table 6.1 shows the results obtained using a four-layer MLP to classify emotions from deep features only, handcrafted features only, and the concatenation of both.

Table 6.1: Classification results with four-layer MLP for ECG improvements.

Features	**Accuracy**			**Individual Accuracies**				
	Train	**Validation**	**Test**	**Amusement**	**Anger**	**Disgust**	**Sad**	**Fear**
Handcrafted	31.94 %	30.74 %	29.11 %	0 %	1.56 %	0 %	60.62 %	44.63 %
Deep	70.24 %	43.07 %	47.03 %	30.23 %	37.45 %	8.06 %	66.57 %	52.81 %
Deep + Handcrafted	55.74 %	35.28 %	28.55 %	16.28 %	22.55 %	0 %	48.63 %	27.19 %

On the one hand, using the deep extracted features, the model was able to fit the training data well, reaching a training accuracy of 70.24%. In validation and test, despite achieving lower accuracies, these are still relatively satisfactory, with almost 50% of the segments being correctly classified. Looking at the individual accuracies obtained, the most challenging one is the disgust class, with only 8% of the segments being correctly identified. On the other hand, using the handcrafted features, the results are much lower, with classes amusement and disgust not being identified in any segment, and the anger class achieving an unsatisfactory result.

Combining both types of features harms the results obtained with deep features only. This leads to the conclusion that all of the assumptions made to obtain the handcrafted features harmed the results from the model.

In conclusion, despite the considerable overfitting in the deep self-learning approach, which could be fought with additional regularisation, it has shown the superior ability to pinpoint the appropriate features for the task at hand, attaining improved overall performance in the test data. Therefore, the deep extracted features were considered for the aggregation of the segments with the LSTM.

Table 6.2 shows the results obtained using an LSTM to aggregate the segments of the same signal and classify the different emotions, with and without the help of a linear layer at the end.

Table 6.2: Classification results with LSTM for segments aggregation.

Model	Accuracy			Individual Accuracies				
	Train	Validation	Test	Amusement	Anger	Disgust	Sad	Fear
LSTM (forward)	24.1 %	27.27 %	22.09 %	14.29 %	0 %	17.39 %	0 %	81.25 %
LSTM (backward)	21.22 %	18.18 %	16.28 %	0 %	0 %	0 %	68.75 %	18.75 %
LSTM + Linear	56.27 %	18.18 %	33.72 %	14.29 %	17.65 %	82.61 %	0 %	31.25 %
LSTM + Dropout (0.2) + Linear	57.56 %	36.36 %	32.56 %	21.43 %	35.29 %	73.91 %	6.25 %	6.25 %

It can be observed that both forward and backward approaches achieved low results. In fact, these were obtained within the first 10 epochs of training and the model was never able to improve them by finding a better fit to the training data. Therefore, there was the need to use the full potential of the bidirectional LSTM by combining both forward and backward outputs.

From the different experiments regarding temporal aggregation of the segments, the one which used an LSTM followed by a linear layer with dropout brought the best results. Even though some overfit is still present in this approach, the dropout was able to approximate the accuracy of the validation dataset to the training dataset, thus reducing the overfit as intended. Note that this last methodology is also the only one capable of identifying all classes. Despite achieving lower results in comparison to the non-aggregated deep approach, these results can be in part due to the signal-independent data partitions that are used to train and test the aggregation structures, which becomes more demanding for the model, thus obtaining lower accuracies in the test dataset.

6.4 Conclusion

The present chapter focused on the application of deep learning techniques to the handcrafted methodology presented in Chapter 5 regarding emotion recognition from ECG signals. These deep learning techniques focused not only on feature extraction but also on the classification and aggregation of segments.

Deep extracted features proved to be more efficient for the task at hand when compared to handcrafted features, both classified with a similar MLP. In fact, when concatenating both, it brought down the accuracy obtained with deep features only.

Regarding aggregation, different methods could have been used. For instance, the mean and standard-deviation values such as in Chapter 5. However, it would be expected that more complex aggregation methods like MLPs and LSTMs would be able to obtain higher accuracies by being more selective and potentially highlighting the parts of the signal more significant to the task. Since the signals have very different lengths, from 32 s to 245 s, an LSTM was chosen to deal with them due to its capacity to model both short-term and long-term dependencies in a sequence, followed by a linear layer with dropout. Although the results were not as high in comparison to the non-aggregated results, this can be justified by the fact that the scenario changes from completely random to signal-independent, which difficulties the training of the model. Nevertheless, it is still an improvement from the preliminary results obtained.

Chapter 7

Improvements on Facial Emotion Recognition

7.1 Introduction

The initial work developed focused on attempting to replicate some state-of-the-art methods already proven to be able to achieve good results regarding facial emotion recognition. However, some improvements can still be explored. Therefore, in this chapter, these improvements will be attempted by first evaluating if there is an advantage in joining the handcrafted features with the deep ones and second in merging the different frames by making use of an LSTM in order to obtain only one prediction per video.

7.2 Methods

Concatenation of deep and handcrafted features

So far, in Chapter 5, handcrafted features using the SIFT and BoWs methodologies and deep features from the pre-trained Inception-ResNet-v1 were obtained. Similarly to the improvements of the model from ECG signals in Chapter 6, the handcrafted and deep features were merged by a simple concatenation. This aimed at verifying which set of features worked best.

The concatenation lead to a total number of 1842 features. Then, they were scaled to avoid big differences between the two. Finally, they were evaluated using an MLP with one linear layer and 5 neurons as in subsection 5.2.2. This classifier was trained for 200 epochs, with a batch size of 100 samples, and a learning rate of 10^{-4}, having Adam as optimiser and cross entropy loss to compute the loss between the inputs and the targets. An early stopping strategy was adopted, and the model which presented the lower validation loss was saved. This training details remained the same throughout the following sections of the present chapter.

Alike the methodologies developed to obtain the baseline, since this experiment considers the individual frames as inputs of the classifier, the data division was performed in a random way, by considering 80% of the data for training and 20% for testing as in section 5.2.3. Also, 10% of the

training data was used for validation.

Temporal aggregation of frames with LSTM

At last, to obtain only one prediction per video, instead of per frame, the different frames of the same video were aggregated with a bidirectional LSTM. Following the best approach to aggregation seen using ECG signals in Chapter 6, the ten outputs from the LSTM were fed to a linear layer with 5 neurons which mapped them to the final 5 categories, followed by a softmax activation. A dropout layer was also added in between. A schematic representation of this model can be seen in Figure 7.1.

It is important to remind that, in this experiment, the data division is applied to the full-length videos instead of the individual frames. Therefore, the train, validation and test datasets are signal-independent as described in section 5.2.3, which brings an additional challenge to the model.

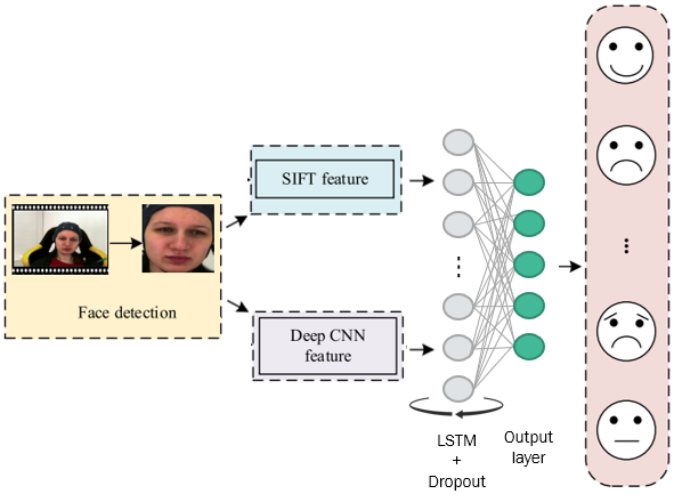


Figure 7.1: Schematic representation of the pipeline adopted to concatenate both SIFT and deep facial features and to aggregate the features of the different frames using an LSTM, followed by dropout and a linear layer for final classification.

7.3 Results

Table 7.1 shows the results obtained from both the preliminary experiments, with either SIFT features or deep features alone, and the results obtained when concatenating both attempting to improve the facial emotion recognition model.

When concatenating the handcrafted with the deep extracted features, an increase in performance was observed in all partitions of the dataset. Although the accuracies per class have all seen an increase, the class that benefited the most from this concatenation was the disgust class. It is curious that the two types of features alone were not able to successfully distinguish this class

Table 7.1: Best classification results obtained using the SIFT features and/or the InceptionResNet-V1 features.

Features	Accuracy			Individual Accuracies				
	Train	Validation	Test	Amusement	Anger	Disgust	Sad	Fear
SIFT	63.39 %	61.07 %	61.19 %	56.67 %	59.70 %	6.01 %	69.39 %	69.54 %
Deep	70.41 %	69.69 %	69.77 %	81.04 %	63.67 %	34.69 %	73.30 %	72.36 %
SIFT + Deep	76.32 %	75.19 %	74.95 %	85.37 %	69.82 %	57.53 %	77.45 %	74.13 %

from the others, but together the network is able to take advantage of both to hereby correctly identify more than half of the signals labeled as disgust.

Contrarily to what was observed with the ECG signals in Chapter 6, these results show that there is no overfitting in the model, and that it is able to generalize well to unseen data. This indicates that for the face modality more complex methods could be developed to attempt to achieve even better performances, while for the ECG modality, a simpler approach could be more beneficial. Nevertheless, comparing these two unimodal approaches, the face stands out as the strongest modality.

Table 7.2 shows the results obtained using an LSTM to aggregate the frames of the same video followed by a linear layer to classify the different emotions, with and without the use of dropout in between. From this table it can be observed that using an LSTM without dropout does not allow the model to correctly learn from the data. In fact, these results were obtained only in the second epoch of training and it represents the random model - *i.e.* only one class being predicted, leading to an accuracy of 100 % divided by the number of classes. Looking further into training, the model showed huge overfitting and no capability of generalizing, hence the importance of dropout.

Table 7.2: Classification results with LSTM for frames aggregation.

Model	Accuracy			Individual Accuracies				
	Train	Validation	Test	Amusement	Anger	Disgust	Sad	Fear
LSTM + Linear	22.02 %	21.21 %	13.95 %	100 %	0 %	0 %	0 %	0 %
LSTM + Dropout (0.2) + Linear	62.45 %	25.76 %	22.09 %	66.67 %	17.65 %	13.64 %	10.53 %	18.75 %
LSTM + Dropout (0.4) + Linear	70.04 %	25.76 %	10.47 %	33.33 %	0 %	0 %	0 %	31.25 %

The addition of dropout allows the model to learn well from the training data, reaching accuracy values similar to the ones observed before; however, these still show a weak capacity for generalization, with the validation and test accuracies both falling short of the ones obtained in the training phase.

Like in Chapter 6, this signal-independent data partition can be in part responsible for the results obtained with the temporal aggregation. However, other differences between the two can be discussed. Imagine a recording with 60 s - roughly equal to the average length of the recordings. Once the division in 10 s ECG segments occurs, it creates 6 segments and, with the percentages considered for data division, 5 segments would go to the train dataset and 1 to the test dataset. Following the same reasoning for face videos, it creates 300 frames, with 240 frames going to the train set and 60 to the test set. From this, it can be concluded that the model for emotion recognition

from face frames receives 50 times more samples than the model from ECG segments, which has been previously confirmed with the overview of the dataset present in subsection 5.3.1. Besides, many consecutive frames are similar and, being present in both train and test datasets enables the model to learn and adapt to all the data partitions much easier. In the temporal aggregated models, the number of examples becomes equal to the original ones present in the dataset and this drastic reduction in the number of samples, allied with the signal-independent partition, shows to be more significant and therefore harsher in the facial emotion recognition model. To fight this, more data is needed both by gathering more examples from the subjects already recorded and by recruiting more subjects. Finally, feature selection methods could come at hand here due to the high number of input features, some which may be irrelevant and others redundant, which can also be in be in part responsible for the overfitting observed.

Nevertheless, a dropout of 0.2 seems more fit for the problem since it allows to identify all classes, while a dropout of 0.4 does not.

In the temporal aggregation models, the results from the two unimodal models invert, with the face modality showing a huge overfitting in comparison to the ECG modality. Here, similar to the non-aggregated ECG model, the aggregated facial emotion recognition model could perhaps benefit from a simpler approach. For instance, instead of a linear layer in the end, the outputs of the forward and backward LSTMs could be merged with a simple sum, multiplication or average.

7.4 Conclusion

The present chapter focused on further improving the methodology presented in Chapter 5 regarding facial emotion recognition. It focused on joining together the previously handcrafted and deep extracted features, and also on the aggregation of the different frames to obtain only one prediction per video.

The concatenation of handcrafted and deep extracted features proved to be more efficient for the task at hand when compared to both of these features alone. It brought enough information to classify all categories adequately and therefore further experiments were conducted with this set of features.

Regarding aggregation, the results fall short of the non-aggregated results, which can be justified by not only the fact that the scenario changes from completely random to signal-independent, but also by the harsh reduction in the number of samples for the model to train and test on.

Chapter 8

Multimodal Emotion Recognition

8.1 Introduction

Having established two approaches for emotion recognition from face videos and ECG signals, this chapter intends to evaluate if the fusion of the two modalities is able to achieve a better performance. Therefore, one early fusion and two late fusion techniques were implemented and are here compared.

8.2 Methods

Early fusion

Early fusion techniques aim at combining the features of the different modalities prior to classification. In the previous two chapters, two different models were developed for emotion recognition from the full length of the recordings, both for the face videos and the ECG signals. The early fusion strategy developed aims at joining those two models. In summary, the features of a pair of ECG signal and face video enter the model and each go through their separate bidirectional LSTM, each with 5 hidden units. This results in 10 outputs from each LSTM, which are concatenated together before passing through a linear layer which will map them to the final 5 classes using 5 neurons. As before, a dropout layer with a probability of 0.6 is also added in between. This probability was chosen bearing in mind the results, since lower probabilities would not allow the model to keep training past the first 5 epochs without overfitting. All of these steps are trained together. A schematic representation of this model can be seen in Figure 8.1.

Late fusion

Late fusion techniques aim at combining the predictions of the classifiers of the different modalities. Therefore, the same previous pre-trained models were used and the probabilities from each signal and recording were extracted from the corresponding models. These probabilities were then fused by two different methods:

1. the sum rule, which sums the confidences/probabilities for a given class from each classifier and chooses the class with the overall highest confidence as output.
2. the maximum rule, which for each sample takes the classifier that is most confident in its output and chooses that one to make the final prediction.

A schematic representation of these models can be seen in Figure 8.2.

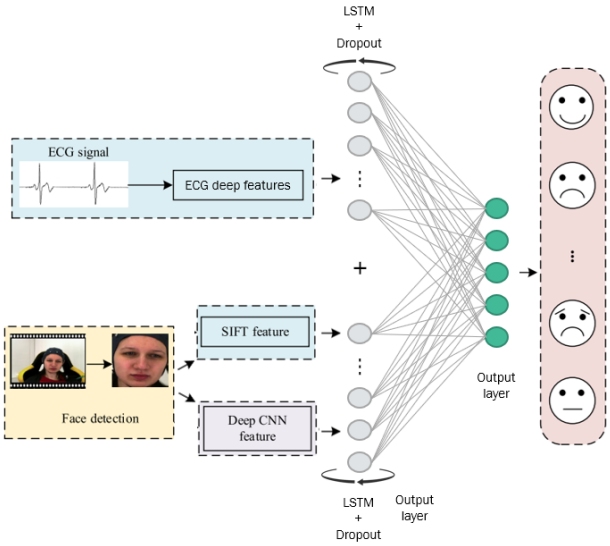


Figure 8.1: Schematic representation of the pipeline adopted for the early fusion model.

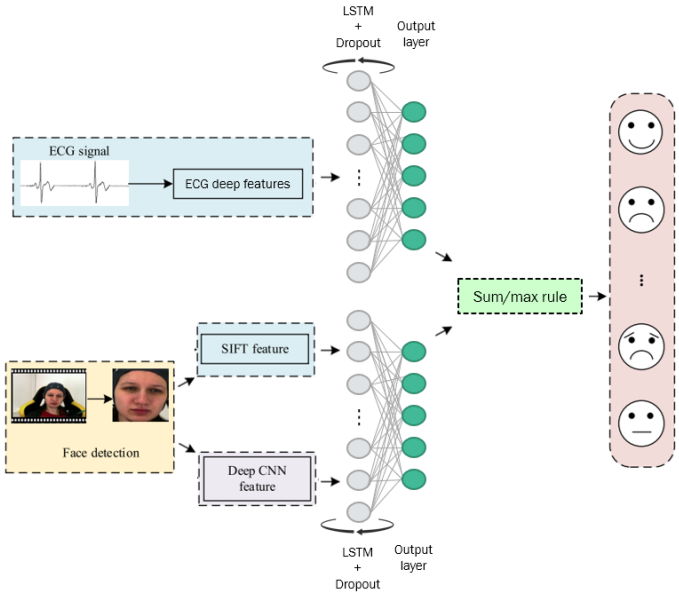


Figure 8.2: Schematic representation of the pipeline adopted for the late fusion model.

8.3 Results

The results obtained for each modality individually as well as for the fusion of the two with early and late fusion techniques can be seen in Table 8.1.

Table 8.1: Classification results with LSTM for frames aggregation.

Modality	Fusion	Accuracy			Individual Accuracies				
		Train	Validation	Test	Amusement	Anger	Disgust	Sad	Fear
ECG	-	57.56 %	36.36 %	32.56 %	21.43 %	35.29 %	73.91 %	6.25 %	6.25 %
Face	-	62.45 %	25.76 %	22.09 %	66.67 %	17.65 %	13.64 %	10.53 %	18.75 %
ECG + Face	Early	64.62 %	25.37 %	27.06 %	100 %	23.53 %	8.70 %	6.67 %	12.50 %
ECG + Face	Late (sum)	- %	- %	57.65 %	100 %	52.94 %	60.87 %	46.67 %	31.25 %
ECG + Face	Late (maximum)	- %	- %	60.00 %	92.86 %	47.06 %	60.57 %	53.33 %	37.50 %
Zhongmin et al. [88]	Early (MDBN)	-	-	80.89 %	89.25 %	90.74 %	52.60 %	86.45 %	85.42 %

On the one hand, the results for the early fusion model show that this suffers from the same problems observed with the face videos alone in Chapter 7. High overfitting is observed, with a big difference in performance between the different data partitions. More data could help the model learn more representative patterns that would allow it to better generalize to unseen data, feature selection methods could also help to eliminate irrelevant and redundant features, thus improving the performance of the model, and a simpler model could prevent it from adapting so well to the training data, leaving the other data partitions with such a different performance. On the other hand, the results from both late fusion models show that these are able to take the best out of the two individual models and combine them to obtain better predictions overall. Figure 8.3 visually represents the results obtained from ECG signals alone, face videos alone, and both late fusion techniques, for easier comparison.

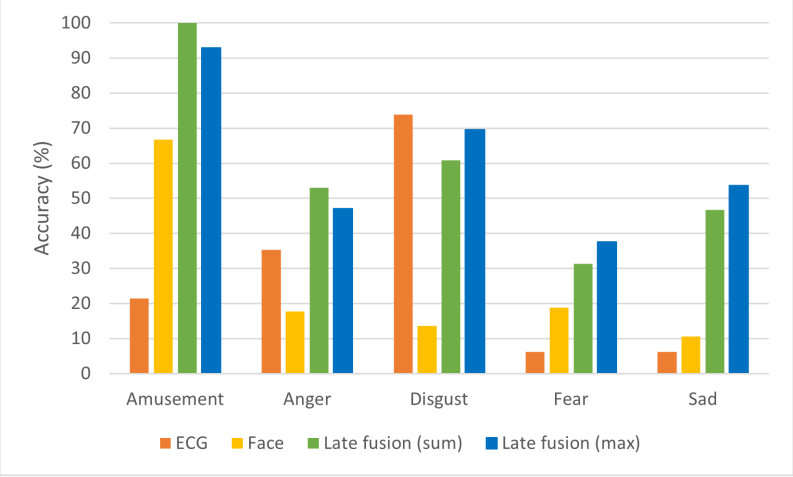


Figure 8.3: Comparison between accuracies obtained for the modalities, individually and combined.

From the figure above, it can be observed that the face videos are better at identifying the amusement class, while the ECG signals show a clear preference for the disgust class. Overall,

both the sum and maximum rule surpass the performances of the individual models for all the classes, except disgust. The advantage of fusion is most noticed for fear and sad classes, which despite the low accuracies obtained by the unimodal models, the confidences of these are able to complement each other and produce the right predictions when fused.

It is important to notice that, in the work of Zhongmin et al. [88], the disgust class was identified as the most complicated one to recognise; however, the developed model does not follow this behaviour, with disgust being the second most correctly identified class. Finally, despite the fact that the developed model could not achieve such high accuracies as the one obtained by Zhongmin et al. [88] with the same data, it still shows its worth by being in signal-independent settings, which was not guaranteed by the authors. Also note that the results are not directly comparable, since the authors consider four modalities, instead of two.

8.4 Conclusion

This chapter intended to develop multiple fusion methodologies in order to develop a bimodal model from the unimodal models developed in Chapter 6 and 7.

The early fusion approach showed the same problems already observed in the facial model alone, with huge overfitting. To fully understand the potential of this approach, more data is needed to provide the model with many diverse examples and enable it to generalize well even for unseen data.

Both late fusion approaches are simple and were able to take advantage of each unimodal model to produce better results, thus satisfying the hypothesis that fusing different modalities is beneficial for an emotion recognition task.

Chapter 9

Conclusion and Future Work

The main goal of this dissertation was to develop an emotion recognition model using both ECG signals and face videos, to explore the hypothesis that multi-modality methods can perform better than single-modality models. To achieve this, the fundamentals behind emotion modelling and the different possible modalities were explored, reaching the conclusion that the two chosen modalities are the most promising in terms of performance and robustness.

After getting acquainted with the several state-of-the-art methods for all the steps involved in an emotion recognition model, the promising methodology of Zhongmin et al. [88] was replicated to serve as a baseline for this work and to understand what improvements could be made. Missing information in the original article made it difficult to accurately replicate the reported results. Especially for ECG signals, the results obtained fell short of those reported by the authors, which could be due to the number of assumptions that needed to be made. Moreover, their methodology combined handcrafted and deep feature extraction from face videos, but only handcrafted feature extraction from ECG signals, which left a possible improvement since deep learning methods seem to be the current state-of-the-art due to their better results. Also, a limitation was encountered regarding the division of ECG signals in segments and of face videos in frames. This division allied to a random partition of the dataset lead to data from the same recording being present in both train and test splits, which is an unrealistic scenario.

After pointing out the limitations and possible improvements of the baseline, different improvements were explored and applied. Starting with the ECG, the methodology of Sarkar et al. [65], further improved by Antunes [5], was chosen to extract deep features from a pre-trained network. These features showed to be able to achieve the best performance in comparison to the handcrafted features alone and to the concatenation of both. Therefore, these were considered for further improvements. The results of the baseline regarding face information were already satisfying, so one improvement was made, which consisted of concatenating both handcrafted and deep extracted features, which in this modality proved to achieve even better results than when considered separately. Furthermore, the problems regarding random data partition and the consideration of ECG segments and video frames separately were tackled at once by the addition of an LSTM. The LSTM receives as input the characteristics of subsequent segments/frames and

returns a set of features representing the signal/video as a whole. This aggregation enabled the division of the data into train and test datasets without containing elements of the same recording. Despite being a much more realistic scenario, it proved to be much more difficult for the model, especially for the face modality. This model showed a high overfit, which could be reduced by giving the model more data to train on, which would allow it to learn representative patterns from the samples instead of memorizing the training data and not being able to perform in unseen data. Also, feature selection methods could be used in the future to reduce the number of features entering the model, keeping only the most relevant ones to the task.

Having established two unimodal models, different fusion scenarios were explored to achieve the final goal. To do so, one early fusion approach was first attempted which consisted of the features of a pair of ECG signal and face video entering the model and each going through their separate bidirectional LSTM. These results from each LSTM were then concatenated together before passing through a linear layer which mapped them to the final 5 classes. All of these steps were trained at once. The early fusion approach showed the same problems already observed in the facial model alone, with huge overfitting. To fully understand the potential of this approach, more data is needed to provide the model with many diverse examples and enable it to generalize well even for unseen data.

Two late fusion techniques were then developed by combining the predictions of the previously developed and pre-trained classifiers of the different modalities. The probabilities from each signal and video recording were extracted from the corresponding classifier and these probabilities were then fused by both the sum rule, which sums the confidences/probabilities for a given class from each classifier and chooses the class with the highest confidence as the prediction, and the maximum rule, which for each sample takes the classifier that is most confident in its output and chooses that one to make the final prediction. Both late fusion approaches are simple and were able to take advantage of each unimodal model to produce better results, thus achieving the final goal and satisfying the hypothesis that fusing different modalities is beneficial for an emotion recognition task.

In comparison to the initial work of Zhongmin et al. [88], this work was not able to achieve such high values of accuracy. Nevertheless, it was developed under a much more realistic scenario and was still able to achieve satisfying results despite only considering half of the modalities. Future efforts should be made on the collection of more data to be able to provide the model with realistic and diverse examples to train on, which would eventually allow it to generalize better to different and unseen data.

Bearing all this in mind, future work should focus both on the data and on the development of the models. Regarding the data, more recordings from the same subjects and more recordings from different subjects would allow to create a more diverse dataset. This would allow to divide the data into training and testing partitions not only in a signal-independent manner but also subject-independent, which is ideal for real-world applications but hard to perform with a low number of subjects. Also, more data is always needed to allow quality learning of the model and avoid the overfitting phenomena. Still focusing on the data, the annotations need to be made

carefully. Taking BioVid Emo DB [99] as an example, each recording is labeled as the emotion the video intends to elicit in the participants. A better assessment of the participant's emotional state should be done and the labeling could even be performed per segments of the recording, since there is no guarantee that the emotion is present throughout the whole recording. Regarding the models, data with missing annotations can be used for unsupervised/self-supervised learning, resulting in more generalized features rather than task-specific ones. Finally, other fusion techniques should be explored, namely late fusion techniques due to their simplicity and proven results, especially those which provide different weights to the classifiers.

References

[1] 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(3):209–222, 2015. doi: 10.1109/TAFFC.2015.2392932.

[2] S. M. Alarcão and M. J. Fonseca. Emotions Recognition Using EEG Signals: A Survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2019. doi: 10.1109/TAFFC.2017.2714671.

[3] Allied Market Research. Mental Health Market by Disorder (Schizophrenia, Alcohol Use Disorders, Bipolar Disorder, Depression, Anxiety, Post-Traumatic Stress Disorder, Substance Abuse Disorders, Eating Disorders, and Other Disorders), Service (Emergency Mental Health Services, Outpatient Counselling, Home-Based Treatment Services, Inpatient Hospital Treatment Services, and Other Services), and Age Group (Pediatric, Adult, and Geriatric): Global Opportunity Analysis and Industry Forecast, 2021–2030. `https://www.alliedmarketresearch.com/mental-health-market-A11770`, 2021. Accessed: 2022-01-09.

[4] S. A. Andrzejewski and J. S. Podoshen. The Influence of Cognitive Load on Nonverbal Accuracy of Caucasian and African-American Targets: Implications for Ad Processing. *Journal of International Consumer Marketing*, 29:1–8, 2016. doi: 10.1080/08961530.2016.1245120.

[5] I. Antunes. Feel My Heart: Emotion Recognition Using the Electrocardiogram. Master's thesis, University of Porto, https://repositorio-aberto.up.pt/handle/10216/135339, 2021.

[6] Apriorit. Challenges of Emotion Recognition in Images and Video. `https://www.apriorit.com/dev-blog/642-ai-emotion-recognition`, 2019. Accessed: 2022-03-31.

[7] M. Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44:572–587, 2011. doi: 10.1016/j.patcog.2010.09.020.

[8] Baeldung. Multiclass Classification Using Support Vector Machines. `https://www.baeldung.com/cs/svm-multiclass-classification`, 2021. Accessed: 2022-01-19.

[9] M. Bani, S. Russo, S. Ardenghi, G. Rampoldi, V. Wickline, S. N. Jr, and M. G. Strepparava. Behind the Mask: Emotion Recognition in Healthcare Students. *Medical Science Educator*, 31:1273–1277, 2021. doi: 10.1007/s40670-021-01317-8.

[10] B.P. Woolf, I. Arroyo, D. Cooper, W. Burleson and K. Muldner. Affective Tutors: Automatic Detection of and Response to Student Emotion. In *Advances in Intelligent Tutoring Systems*, pages 207–227. Springer Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-14363-2_10.

[11] S. Brás, J. H. Ferreira, S. C. Soares, and A. J. Pinho. Biometric and Emotion Identification: An ECG Compression Based Method. *Frontiers in Psychology*, 9, 2018. doi: 10.3389/fpsyg.2018.00467.

[12] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, page 205–211, 2004. doi: 10.1145/1027933.1027968.

[13] K. Byron, S. Terranova, and S. Nowicki. Nonverbal Emotion Recognition and Salespersons: Linking Ability to Perceived and Actual Success. *Journal of Applied Social Psychology*, 37 (11):2600–2619, 2007. doi: 10.1111/j.1559-1816.2007.00272.x.

[14] C.-C. Chen and F. R. Tsui. Comparing different wavelet transforms on removing electrocardiogram baseline wanders and special trends. *BMC Medical Informatics and Decision Making*, 20(343), 2020. doi: 10.1186/s12911-020-01349-x.

[15] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016. doi: 10.1109/TPAMI.2016.2515606.

[16] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, C. Fookes. Deep Spatio-Temporal Features for Multimodal Emotion Recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1223, 2017.

[17] L. De Silva and P. C. Ng. Bimodal emotion recognition. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 332–335, 2000. doi: 10.1109/AFGR.2000.840655.

[18] L. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications*, volume 1, pages 397–401, 1997. doi: 10.1109/ICICS.1997.647126.

[19] M. A. Delkhoon and F. Lotfizadeh. An investigation on the effect of gender on emotional responses and purchasing intention due to advertisements. *UCT Journal of Social Science and Humanities Research*, 2(1):6–11, 2014.

[20] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. RetinaFace: Single-stage Dense Face Localisation in the Wild. *arXiv*, 2019. doi: 10.48550/arxiv.1905.00641.

[21] K. S. do Prado. Face Recognition: Understanding LBPH Algorithm. `https://towardsdatascience.com/face-recognition-how-lbph-works-90ec258c3d6b`, 2017. Accessed: 2022-06-18.

[22] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. Al'Absi, S. Shah, S. Mitra, T. Kwon and J.W. Jeong. Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *9th ACM Conference on Embedded Networked Sensor Systems*, pages 274–287, 2011. doi: 10.1145/2070942.2070970.

[23] P. Ekman. Universals and cultural differences in facial expressions of emotion. *University of Nebraska Press*, 19:207–283, 1971.

[24] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

[25] P. Ekman and W. Friesen. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[26] R. el Kalioubyarchive. We need computers with empathy. *MIT Technology Review*, 120(6): 8, 2017.

[27] F. Guo, W.L. Liu, Y. Cao, F.T. Liu and M.L. Li. Optimization Design of a Webpage Based on Kansei Engineering. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 26(1):110–126, 2015. doi: 10.1002/hfm.20617.

[28] P. M. Ferreira, F. Marques, J. S. Cardoso, and A. Rebelo. Physiological Inspired Deep Neural Networks for Emotion Recognition. *IEEE Access*, 6:53930–53943, 2018. doi: 10.1109/ACCESS.2018.2870063.

[29] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. The World of Emotions is not Two-Dimensional. *Psychological Science*, 18(12):1050–1057, 2007. doi: 10.1111/j.1467-9280.2007.02024.x.

[30] A. Geetha, V. Ramalingam, S. Palanivel, and B. Palaniappan. Facial expression recognition – A real time approach. *Expert Systems with Applications*, 36:303–308, 2009. doi: 10.1016/J.ESWA.2007.09.002.

[31] GlobeNewswire. Global Emotion Detection and Recognition Market Research Report 2021: Focus on Law enforcement, Surveillance, & Monitoring, Entertainment & Consumer Electronics, and Marketing & Advertising. `https://www.globenewswire.com/news-release/2021/10/07/2310103/28124/en/Global-Emotion-Detection-and-Recognition-Market-Research-Report-2021-Focus-on-Law-enforcement-Surveillance-Monitoring-Entertainment-Consumer-Electronics-and-Marketing-Advertising.html`, 2021. Accessed: 2022-01-11.

[32] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, 45(1):80–91, 2012. doi: 10.1016/J.PATCOG.2011.05.006.

[33] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443, 2005. doi: 10.1109/ICSMC.2005.1571679.

[34] H. W. Guo, Y. S. Huang, C. H. Lin, J. C. Chien, K. Haraikawa, and J. S. Shieh. Heart Rate Variability Signal Features for Emotion Recognition by Using Principal Component Analysis and Support Vectors Machine. In *Proceedings - 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering, BIBE 2016*, pages 274–277, 2016. doi: 10.1109/BIBE.2016.40.

[35] H. Gray. *Anatomy of the human body*. Lea & Febiger, 1973. ISBN 0-81210377-7.

[36] H. Tang, W. Liu, W.L. Zheng, B.L. Lu. Multimodal Emotion Recognition Using Deep Neural Networks. In *ICONIP 2017: Neural Information Processing*, pages 811–819, 2017.

[37] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006. doi: 10.1126/SCIENCE.1127647/SUPPL_FILE/HINTON.SOM.PDF.

[38] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18:1527–1554, 2006. doi: 10.1162/NECO.2006.18.7.1527.

[39] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9: 1735–1780, 1997. doi: 10.1162/NECO.1997.9.8.1735.

[40] G. V. Houdt, C. Mosquera, and G. Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53:5929–5955, 2020. ISSN 15737462. doi: 10.1007/S10462-020-09838-1/TABLES/1.

[41] Y. L. Hsu, J. S. Wang, W. C. Chiang, and C. H. Hung. Automatic ECG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Affective Computing*, 11:85–99, 2020. doi: 10.1109/TAFFC.2017.2781732.

[42] C. Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3):260–280, 2007. doi: 10.1111/j.1745-6916.2007.00044.x.

[43] R. Jerath and C. Beveridge. Respiratory Rhythm, Autonomic Modulation, and the Spectrum of Emotions: The Future of Emotion Recognition and Modulation. *Frontiers in Psychology*, 11, 2020. doi: 10.3389/fpsyg.2020.01980.

[44] J. K. Kamarainen. Gabor features in image analysis. In *2012 3rd International Conference on Image Processing Theory, Tools and Applications, IPTA 2012*, pages 13–14, 2012. doi: 10.1109/IPTA.2012.6469502.

[45] R. Kher. Signal Processing Techniques for Removing Noise from ECG Signals. *Journal of Biomedical Engineering and Research*, 1:1–9, 2019.

[46] 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(1):18–31, 2012. doi: 10.1109/T-AFFC.2011.15.

[47] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. Emotion recognition and its application in software engineering. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 532–539, 2013. doi: 10.1109/HSI.2013.6577877.

[48] Y.-T. Lan, W. Liu, and B.-L. Lu. Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2020. doi: 10.1109/IJCNN48605.2020.9207625.

[49] P. Lang. The Emotion Probe: Studies of Motivation and Attention. *American Psychologist*, 50(5):372–385, 1995. doi: 10.1037/0003-066x.50.5.372.

[50] S. Langari, H. Marvi, and M. Zahedi. Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*, 20, 2011. doi: 10.1016/j.imu.2020.100424.

[51] S. Li and W. Deng. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. doi: 10.1109/TAFFC.2020.2981446.

[52] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 494–501, 2014. doi: 10.1145/2663204.2666274.

[53] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial Expression Recognition via a Boosted Deep Belief Network. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014. doi: 10.1109/CVPR.2014.233.

[54] L. D. Lopez, P. J. Reschke, J. M. Knothe, and E. A. Walle. Postural communication of emotion: Perception of distinct poses of five discrete emotions. *Frontiers in Psychology*, 8, 2017. doi: 10.3389/fpsyg.2017.00710.

[55] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 2004 60:2*, 60:91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.

[56] M. Elgendi, B.M. Eskofier, S. Dokos, and D. Abbott. Revisiting QRS Detection Methodologies for Portable, Wearable, Battery-Operated, and Wireless ECG Systems. *PLoS ONE*, 9(1), 2014. doi: 10.1371/journal.pone.0084018.

[57] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic. AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge. In *6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

[58] M. Malfaz and M. A. Salichs. A new architecture for autonomous robots based on emotions. *IFAC/EURON Symposium on Intelligent Autonomous Vehicles*, 37(8):805–809, 2004. doi: 10.1016/S1474-6670(1732079-7.

[59] Markets and Markets. Emotion Detection and Recognition Market Size, Share and Global Market Forecast to 2026. `https://www.marketsandmarkets.com/Market-Reports/emotion-detection-recognition-market-23376176.html?utm_source=PRNewswire&utm_medium=Referral&utm_campaign=PRNewswire&fbclid=IwAR3tEb0z3rJ73BjDedZqKBbV7Z8w81_Zm0KiBvpeONT1x4kTtsSj03Fn-Es`, 2020. Accessed: 2021-11-18.

[60] A. Mehrabian. 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):331–357, 1997. doi: 10.1007/BF02229025.

[61] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021. doi: 10.1109/TAFFC.2018.2884461.

[62] M. Ménard, P. Richard, H. Hamdi, B. Daucé, and T. Yamaguchi. Emotion Recognition based on Heart Rate and Skin Conductance. In *2nd International Conference on Physiological Computing Systems (PhyCS)*, 2015. doi: 10.5220/0005241100260032.

[63] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585, 1994. doi: 10.1109/ICPR.1994.576366.

[64] OpenCV. Introduction to SIFT (Scale-Invariant Feature Transform). `https://docs.opencv.org/4.x/da/df5/tutorial_py_sift_intro.html`. Accessed: 2022-06-18.

[65] P. Sarkar, A. Etemad. Self-supervised ECG representation learning for emotion recognition. *arXiv*, pages 1–13, 2020. doi: 10.1186/s12911-020-01349-x.

[66] N. Perveen and C. Mohan. Configural Representation of Facial Action Units for Spontaneous Facial Expression Recognition in the Wild. In *15th International Conference on Computer Vision Theory and Applications*, 2020. doi: 10.5220/0009099700930102.

[67] J. R. Pinto, J. S. Cardoso, and A. Lourenco. Evolution, current challenges, and future possibilities in ECG Biometrics. *IEEE Access*, 6:34746–34776, 2018. doi: 10.1109/ACCESS.2018.2849870.

[68] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana. Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition. *Procedia Computer Science*, 116: 523–529, 2017. doi: 10.1016/J.PROCS.2017.10.038.

[69] R. Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21 (4-5):529–553, 1982. doi: 10.1177/053901882021004003.

[70] A. Raij, P. Blitz, A. A. Ali, S. Fisk, B. French, S. Mitra, M. Nakajima, M. H. Nguyen, K. H. Plarre, M. Rahman, S. Shah, Y. Shi, N. Stohs, M. al'Absi, E. Ertin, T. P. Kamarck, S. Kumar, M. Scott, D. P. Siewiorek, and A. Smailagic. mStress: Supporting Continuous Collection of Objective and Subjective Measures of Psychosocial Stress on Mobile Devices. In *ACM Wireless Health 2010 San Diego, California USA*, 2010.

[71] K. Rattanyu, M. Ohkura, and M. Mizukawa. Emotion monitoring from physiological signals for service robots in the living space. In *ICCAS 2010*, pages 580–583, 2010. doi: 10.1109/ICCAS.2010.5669914.

[72] 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)*, pages 1–8, 2013. doi: 10.1109/FG.2013.6553805.

[73] R.L. Mandryk, M.S. Atkins and K.M. Inkpen. A continuous and objective evaluation of emotional experience with interactive play environments. In *Proceedings of the 2006 Conference on Human Factors in Computing Systems*, pages 22–27, 2006. doi: 10.1145/1124772.1124926.

[74] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, 2016.

[75] S. Scotti, M. Mauri, R. Barbieri, B. Jawad, S. Cerutti, L. Mainardi, E.N. Brown and M.A. Villamira. Automatic Quantitative Evaluation of Emotions in E-learning Applications. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1359–1362, 2006. doi: 10.1109/IEMBS.2006.260601.

[76] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access*, 7:57–67, 2019. doi: 10.1109/ACCESS.2018.2883213.

[77] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27:803–816, 2009. doi: 10.1016/J.IMAVIS.2008.08.005.

[78] 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(7), 2018. doi: 10.3390/s18072074.

[79] J. Singh, G. Goyal, and R. Gill. Use of neurometrics to choose optimal advertisement method for omnichannel business. *Enterprise Information Systems*, 14:243–265, 2019. doi: 10.1080/17517575.2019.1640392.

[80] S.K. D'Mello, S.D. Craig, B. Gholson, S. Franklin, R.W. Picard and A.C. Graesser. Integrating Affect Sensors in an Intelligent Tutoring System. In *2005 International Conference on Intelligent User Interfaces*, pages 10–13, 2005.

[81] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. doi: 10.1109/T-AFFC.2011.25.

[82] S. Sosnowski, A. Bittermann, K. Kuhnlenz, and M. Buss. Design and Evaluation of Emotion-Display EDDIE. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3113–3118, 2006. doi: 10.1109/IROS.2006.282330.

[83] E. Spanò, S. Di Pascoli, and G. Iannaccone. Low-Power Wearable ECG Monitoring System for Multiple-Patient Remote Monitoring. *IEEE Sensors Journal*, 16(13):5452–5462, 2016. doi: 10.1109/JSEN.2016.2564995.

[84] B. Sun and Z. Lin. Emotion Recognition using Machine Learning and ECG signals, 2022. URL `https://arxiv.org/abs/2203.08477`.

[85] S. Tivatansakul and M. Ohkura. Improvement of emotional healthcare system with stress detection from ECG signal. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6792–6795, 2015. doi: 10.1109/EMBC.2015.7319953.

[86] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990517.

[87] W. Liu, W.L. Zheng, B.L. Lu. Emotion Recognition Using Multimodal Deep Learning. In *ICONIP 2016: Neural Information Processing*, pages 521–529, 2016.

[88] W. Zhongmin, Z. Xiaoxiao, W. Wenlang, L. Chen. Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video. *International Journal of Machine Learning and Cybernetics*, 11(4), 2020.

[89] W.B. Cannon. The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, 39:106–124, 1927. doi: 10.2307/1415404.

[90] W. Wei, Q. Jia, Y. Feng, and G. Chen. Emotion Recognition Based on Weighted Fusion Strategy of Multichannel Physiological Signals. *Computational Intelligence and Neuroscience*, 2018, 2018. doi: 10.1155/2018/5296523.

[91] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31:153–163, 2013. doi: 10.1016/J.IMAVIS.2012.03.001.

[92] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. doi: 10.1109/CVPR.2013.75.

[93] Y. Ayzenberg, J. Hernandez and R.W. Picard. FEEL: Frequent EDA and Event Logging – A Mobile Social Interaction Stress Monitoring System. In *ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*, pages 2357–2362, 2012.

[94] Y. Huang, J. Yang, P. Liao, J. Pan. Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. In *Computational Intelligence and Neuroscience, Hindawi*, 2017.

[95] J. Yang, X. Huang, H. Wu, and X. Yang. EEG-based emotion classification based on Bidirectional Long Short-Term Memory Network. *Procedia Computer Science*, 174: 491–504, 2020. ISSN 1877-0509. doi: 10.1016/J.PROCS.2020.06.117.

[96] G. N. Yannakakis and J. Hallam. Real-Time Game Adaptation for Optimizing Player Satisfaction. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(2): 121–133, 2009. doi: 10.1109/TCIAIG.2009.2024533.

[97] Y. Yoshitomi, S.-I. Kim, T. Kawano, and T. Kilazoe. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No.00TH8499)*, pages 178–183, 2000. doi: 10.1109/ROMAN.2000.892491.

[98] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, 2016. doi: 10.1109/lsp.2016.2603342.

[99] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H. C. Traue, and S. Gruss. "BioVid Emo DB": A multimodal database for emotion analyses validated by subjective ratings. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6, 2016. doi: 10.1109/SSCI.2016.7849931.

[100] G. Zhao and M. Pietikainen. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. doi: 10.1109/TPAMI.2007.1110.