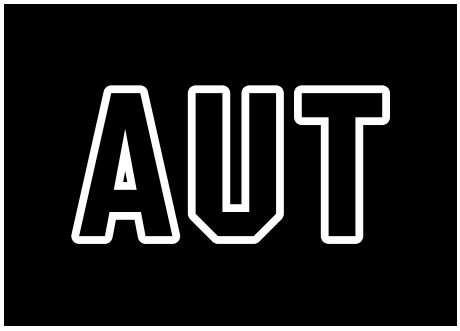


Design and Creative

School of Engineering, Computer and Mathematical Sciences

Knowledge Engineering and Discovery Research Institute (KEDRI)

Affective Computing using Brain-Inspired Spiking Neural Networks

C. K. Clarence, TAN

The thesis submitted to Auckland University of Technology in fulfilment of

requirements for the degree of Doctor of Philosophy

Supervisor Professor Nikolas Kasabov, Associate Professor Wei Qi Yan

C. K. Clarence, TAN

Affective Computing using Brain-Inspired Spiking Neural Networks

Computer and Information Sciences, 1 August, 2020

Supervisors: Professor Nikolas Kasabov, Associate Professor Wei Qi Yan

Auckland University of Technology

Knowledge Engineering and Discovery Research Institute (KEDRI)

School of Engineering, Computer and Mathematical Sciences

Design and Creative Technologies

1010 Auckland, New Zealand

Abstract

The interaction between humans and computational devices are becoming more and more common with the advent of personal digital devices, wearable systems, and other technological interventions. The field of affective computing combines veins of computer science and emotion investigation to essentially enable computational systems to identify the emotional state of users and to generate responses that humans are likely to perceive. It has also been argued that over time, it may be possible for systems to actually 'feel' emotions. This early work by Picard was then followed by much research bringing together diverse fields such as science, ethics, psychology, and engineering, among others. Today, the work on affective computing has resulted in systems that are capable of interpreting, identifying, and responding to the emotional states of users. For this purpose, affective computing makes use of various inputs such as facial images, voice data, biometric data, and body language of the user to identify the emotional state. Therefore, this work has aimed to answer three research questions:

1. What design architecture of an intelligent sensing machine can learn fast from a large amount of online information (emotion, expression etc.), with little prior knowledge and adapt in real-time with the accommodation of new data. The machine must also evolve with new connections of new input and output, learning with complexities of knowledge representation in a multimodal fashion? With this research question, the design algorithm of spiking neural networks (SNN) was studied, and the information processing techniques, learning algorithms, and applications of spiking neurons were discussed and analysed, focusing on feasibility and biological plausibility of the methods.

2. How to effectively emulate the 3D Spatio-temporal processing style associated with a human brain when recognising emotions? This research looked at the SNNs neural spike encoding on human's facial and physiological data and explored the use of the 3D structure to represent human emotions as timings of spikes. For the first time, we have applied SNNs to solve the facial emotion recognition (FER) problem, and the novel approach achieves classification accuracy compared with other state-of-art deep learning approaches that utilise data from facial expressions and physiological signals.

3. How to represent and manage different forms of memories function and integrate spatial, temporal and event-related experience within a multimodal spectrum, self-organise them in a rational and logical analysis of how a human being can sense them depending on the micro-event type? With this research question, the focus was on novel research for fusing temporal, spatial and event-related potential (ERP) information in a spiking neural network architecture. The ERP technique applies EEG signal segmentation based on detection of short-term changes in facial landmarks and relies on no handcrafted EEG features. The next research focus was to fuse the multimodal data consisting of facial expressions along with physiological signals such as ECG, skin temperature, skin conductance, respiration signal and pupil size using both feature-level and decision-level methods within the neural encoding algorithm.

Overall, through this work, the researcher has presented several novel approaches of how unimodal and multimodal affect are handled by the proposed Spatio-temporal 3D structure. It allows the researcher to generate data on which approach is efficient and whether the additional complexity of multimodal approach is sufficiently justified in terms of accuracy improvement obtained. This research also contributes to the existing literature by enhancing the understanding regarding SNNs, specifically three-dimensional Spatio-temporal structures that are constituted of SNNs.

Contents

List of Tables

List of Figures

List of Algorithms

Acknowledgement

I would like to take this opportunity to thank all the people without whom this study would never have been possible. Although it is just my name on the cover, it has been a collaborative effort from researchers from Venezuela, Croatia, India and Finland.

For my supervisors, Prof. Nikola Kasabov and Associate Prof. Wei Qi, Yan. Nik, you have encouraged and provided me with the opportunity to enrol into the doctoral programme after my last M.B.A studies nearly twenty years ago. It has not been easy for my family, to uproot to New Zealand, leaving our extended families and my career behind. I much appreciate your guidance and support you offered when needed and continuously pushing me to make this piece of work better. Wei Qi, thank you for being an incredibly valuable part of my unusual journey of mine and giving me encouragement and support, especially during the critical first year of my PhD studies.

I would also want to extend my gratitude towards Mrs Joyce D'mello, the administrative manager of KEDRI. Joyce has been a source of emotional stability, not only to me but to the whole of KEDRI's postgraduate students and researchers.

And finally, to my wife, Magdalene. I thank her for her selfless support, prayer and faith in God and believing that there is a higher purpose to all our sacrifices.

Declaration

I, Clarence Tan, declare that this thesis titled, Affective Computing using Brain-Inspired Spiking Neural Network and the work presented in it are my own and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Chapters 3 to 6 of this thesis represent separate papers that have either been published, submitted, or are in preparation for peer-reviewed journals. All co-authors have approved the indusion of the joint work in this doctoral thesis.

Auckland, August 2020

C. K. Clarence, TAN

School of Engineering, Computer and Mathematical Sciences

Contract Regarding Publication Intent (Co-authorship Declaration)

We hereby agree, as outlined below, regarding the manuscripts; Titled:

Spiking Neural Networks: Background, Recent development and the NeuCube Architecture – Chapter 3

FIRST/PRINCIPAL AUTHOR**:

Name : **C. K. Clarence, TAN (C.T.)** Signature:

SECOND AUTHOR:

Name : **Marko Sarlija (M.S.)** Signature:

THIRD AUTHOR:

Name : **Prof. Nikola Kasabov (N.K.)** Signature:

****Please see the agreed percentage contribution**

This is to confirm that the **first/principal author (C.T.)** of the manuscript, has a stated contribution of **82%; M.S. – 10% and N.K. – 8%.**

This contributions include:

- Conceptualizing a research idea
- Refining/ crystalizing a research idea
- Literature search: Summarizing literary pieces (e.g., articles, book chapters, etc.)
- Creating a research design (e.g., counterbalancing, randomization to conditions, survey design etc.)
- Selecting an Instrument/ a measure: Instrument construction
- Selection of statistical tests/analyses
- Performing statistical analyses and computations (including computer work)
- Interpretation of statistical analyses
- Manuscript:
  - Writing an introduction section
  - Writing a methods section
  - Writing results section
  - Writing discussion section
  - Writing conclusive summaryWriting limitations of the study
  - Writing future directions of the study

The signed agreement fulfils the Format 2 Requirements stipulated in the Postgraduate Handbook 2019 – Page 104.

Status of the manuscript for publication:

Accepted to Springer's Neural Letter Processing Journal (Status as of 31/03/2021).

Publications

1. Tan, C., Sarlija, M., & Kasabov, N. (2020). Spiking Neural Networks: Background, Recent Development and the NeuCube Architecture. *Neural Processing Letters*, 52(2), 1675-1701. https://doi.org/10.1007/s11063-020-10322-8

Neural Processing Letters

Spiking neural networks: background, recent development and the NeuCube architecture

--Manuscript Draft--

Manuscript Number:	
Full Title:	Spiking neural networks: background, recent development and the NeuCube architecture
Article Type:	Review Article
Keywords:	Artificial neural networks; Spiking neural networks; Spike encoding; Spike-timing dependent plasticity; Spatio-temporal brain data; NeuCube
Corresponding Author:	Marko Šarlija Sveuciliste u Zagrebu Fakultet Elektrotehnike i Racunarstva Zagreb, Grad Zagreb CROATIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Sveuciliste u Zagrebu Fakultet Elektrotehnike i Racunarstva
Corresponding Author's Secondary Institution:	
First Author:	Clarence Tan, MSc
First Author Secondary Information:	
Order of Authors:	Clarence Tan, MSc
	Marko Šarlija, MSc
	Nikola Kasabov, PhD
Order of Authors Secondary Information:	
Funding Information:	

School of Engineering, Computer and Mathematical Sciences

Contract Regarding Publication Intent (Co-authorship Declaration)

We agree, as outlined below, regarding the manuscripts; Titled:

FacialSense: Emotional Valence Recognition Using Brain-inspired Spiking Neural Networks – Chapter 4

FIRST/PRINCIPAL AUTHOR**:

Name: **C. K. Clarence, TAN (C.T.)** Signature:

SECOND AUTHOR:

Name: **Dr. Ceballos, G. A. (C.G.A)** Signature:

THIRD AUTHOR:

Name: **Prof Nikola Kasabov (N.K.)** Signature:

FOURTH AUTHOR:

Name: **Dr. Narayan P. S (N.P.S.)** Signature:

****Please see the agreed percentage contribution**

This is to confirm that the **first/principal author (C.T.)** of the manuscript, has a stated contribution of **85% ; C.G.A. – 7% ; N.K. – 5% and N.P.S. – 3%**

This contributions include:

- Conceptualizing a research idea
- Refining/ crystalizing a research idea
- Literature search: Summarizing literary pieces (e.g., articles, book chapters, etc.)
- Creating a research design (e.g., counterbalancing, randomization to conditions, survey design etc.)
- Selecting an Instrument/ a measure: Instrument construction
- Selection of statistical tests/analyses
- Performing statistical analyses and computations (including computer work)
- Interpretation of statistical analyses
- Manuscript:
    - Writing an introduction section
    - Writing a methods section
    - Writing results section
    - Writing discussion section
    - Writing conclusive summaryWriting limitations of the study
    - Writing future directions of the study

The signed agreement fulfils the Format 2 Requirements stipulated in the **Postgraduate Handbook 2019 – Page 104.**

Status/Date of the manuscript for publication:

Accepted pending revision Springer's Evolving Systems Journal (Status as of 31/03/2021).

Evolving Systems

FacialSense: Emotional Valence recognition using Brain-inspired Spiking Neural Network

--Manuscript Draft--

Manuscript Number:	EVOS-D-20-00109
Full Title:	FacialSense: Emotional Valence recognition using Brain-inspired Spiking Neural Network
Article Type:	S.I. : Sustainable Evolving Computing Systems and its Applications
Corresponding Author:	Clarence Tan Auckland University of Technology NEW ZEALAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Auckland University of Technology
Corresponding Author's Secondary Institution:	
First Author:	Clarence Tan
First Author Secondary Information:	
Order of Authors:	Clarence Tan
	Gerardo Ceballos
	Nikola Kasabov
	Narayan Puthanmadam Subramaniyam
Order of Authors Secondary Information:	

School of Engineering, Computer and Mathematical Sciences

Contract Regarding Publication Intent (Co-authorship Declaration)

We hereby agree, as outlined below, regarding the manuscripts; Titled:

NeuroSense: Short-Term Emotion Recognition Based on Spiking Neural Network Modelling of Spatio-Temporal EEG Patterns – Chapter 5

FIRST/PRINCIPAL AUTHOR**:

Name : **C. K. Clarence, TAN (C.T.)** Signature:

SECOND AUTHOR:

Name : **Marko Sarlija (M.S.)** Signature:

THIRD AUTHOR:

Name : **Prof. Nikola Kasabov (N.K.)** Signature:

****Please see the agreed percentage contribution**

This is to confirm that the **first/principal author (C.T.)** of the manuscript, has a stated contribution of **82%; M.S. - 10% and N.K. – 8%.**

This contributions include:

- Conceptualizing a research idea
- Refining/ crystalizing a research idea
- Literature search: Summarizing literary pieces (e.g., articles, book chapters, etc.)
- Creating a research design (e.g., counterbalancing, randomization to conditions, survey design etc.)
- Selecting an Instrument/ a measure: Instrument construction
- Selection of statistical tests/analyses
- Performing statistical analyses and computations (including computer work)
- Interpretation of statistical analyses
- Manuscript:
  - Writing an introduction section
  - Writing a methods section
  - Writing results section
  - Writing discussion section
  - Writing conclusive summaryWriting limitations of the study
  - Writing future directions of the study

The signed agreement fulfils the Format 2 Requirements stipulated in the Postgraduate Handbook 2019 – Page 104.

Status of the manuscript for publication:

Accepted to Elsiever's Neurocomputing Journal (Status as of 31/03/2021).

Publications

1. Tan, C., Sarlija, M., & Kasabov, N. (2021). NeuroSense: Short-Term Emotion Recognition and Understanding Based on Spiking Neural Network Modelling of Spatio-Temporal EEG Patterns. *Neurocomputing*, 434, 137-148. [23238]. https://doi.org/10.1016/j.neucom.2020.12.098

Elsevier Editorial System(tm) for
Neurocomputing

Manuscript Draft

Manuscript Number:

Title: NeuroSense: Short-Term Emotion Recognition and Understanding Based on Spiking Neural Network Modelling of Spatio-Temporal EEG Patterns

Article Type: Full Length Article (NN)

Keywords: Spiking Neural Networks;
Emotion Recognition;
Affective Computing;
EEG;
Event Detection

Corresponding Author: Mr. clarence tan, Ph.D

Corresponding Author's Institution: Auckland University of Technology

First Author: clarence tan, Ph.D

Order of Authors: clarence tan, Ph.D; Marko Sarlija, ; Nikola Kasabov

School of Engineering, Computer and Mathematical Sciences

Contract Regarding Publication Intent (Co-authorship Declaration)

We hereby agree, as outlined below, regarding the manuscripts; Titled:

FusionSense: Emotion Classification using Feature Fusion of Multimodal Data and Deep learning in a Brain-inspired Spiking Neural Network – Chapter 6

FIRST/PRINCIPAL AUTHOR**:

Name: **C. K. Clarence, TAN (C.T.)** Signature:

SECOND AUTHOR:

Name: **Dr. Ceballos, G. A. (C.G.A.)** Sig

THIRD AUTHOR:

Name: **Prof Nikola Kasabov (N.K.)** Signature:

FOURTH AUTHOR:

Name: **Dr. Narayan P. S (N.P.S.)** Signature:

****Please see the agreed percentage contribution**

This is to confirm that the **first/principal author (C.T.)** of the manuscript, has a stated contribution of **85%; C.G.A. – 7%; N.K. – 5% and N.P.S. – 3%.**

This contributions include:

- Conceptualizing a research idea
- Refining/ crystalizing a research idea
- Literature search: Summarizing literary pieces (e.g., articles, book chapters, etc.)
- Creating a research design (e.g., counterbalancing, randomization to conditions, survey design etc.)
- Selecting an Instrument/ a measure: Instrument construction
- Selection of statistical tests/analyses
- Performing statistical analyses and computations (including computer work)
- Interpretation of statistical analyses
- Manuscript:
  - Writing an introduction section
  - Writing a methods section
  - Writing results section
  - Writing discussion section
  - Writing conclusive summaryWriting limitations of the study
  - Writing future directions of the study

The signed agreement fulfils the Format 2 Requirements stipulated in the Postgraduate Handbook 2019 – Page 104.

Status/Date of the manuscript for publication:

Accepted to MDPI Sensor Journal (Status as of 31/03/2021).

Publications

1. Tan, C., Ceballos, G., Kasabov, N., & Subramaniyam, N. (2020). FusionSense: Emotion Classification using Feature Fusion of Multimodal Data and Deep learning in a Brain-inspired Spiking Neural Network. *Sensors*, 20(18), [5328]. https://doi.org/10.3390/s20185328

sensors

an Open Access Journal by MDPI

Indexed in: PubMed | Indexed in: Ei Compendex | CITESCORE 5.0 SCOPUS | IMPACT FACTOR 3.275

CERTIFICATE OF PUBLICATION

Certificate of publication for the article titled:

FusionSense: Emotion Classification Using Feature Fusion of Multimodal Data and Deep Learning in a Brain-Inspired Spiking Neural Network

Authored by:

Clarence Tan; Gerardo Ceballos; Nikola Kasabov; Narayan Puthanmadam Subramaniyam

Published in:

Sensors **2020**, Volume 20, Issue 18, 5328

MDPI Academic Open Access Publishing since 1996

Basel, March 2021

1 Introduction

The interaction between humans and machines are becoming more prevalent and sophisticated as technology advances. Human-machine interactions are no longer limited to specific and rare occasions. Multiple fields ranging from education to healthcare now incorporate synthetic systems in some capacity. This penetration of technology and human-computer interactions give rise to the need for computers to mimic human behaviour. For example, a telemarketing system that can judge the affective state of a potential customer has a clear advantage over one that does not. Such artificially intelligent systems with a certain degree of emotion recognition and the ability to respond with human-like emotions that are appropriate to the affective state of the consumer have significant applications. Therefore, today, researchers aim at finding ways to make human-machine interactions more organic, mimicking social interactions among human beings.

For machines to mimic human thought patterns, a few characteristics of the human brain must be emulated by an artificial system. It is well known that brain activity associated with emotion processing has Spatio-temporal dynamics. Therefore, an artificial system that can mimic this 3D processing structure is likely to be more successful at affective computing. The brain is also capable of both unimodal and multimodal analysis, recognising emotions from just one input such as facial expression to integrating multiple input data streams. A system that effectively mimics the brain should be able to achieve both unimodal and multimodal analysis. Ideally, such a system should be able to process changing human emotions depending on the complexities of feeling, memories, experience etc. Also, the structure developed for this purpose should be able to handle both cognitive and physiological information, much like the biological structures. The ability to do such affect

recognition continuously, while online, accounting for multisensory inputs is an added advantage.

1.1 Rationale and Motivation

Is the choice that human beings make a result of rationale selection available to them or their emotions that influence their decisions? Can machine intelligence sense human expression to recognise their perception, reaction, and behavioural tendencies through time and space? Can the results of the automatic recognition be used to influence their choice and their decision?

Emotion is a necessary component of all decisions (Vitay & Hamker, 2011). When faced with a choice, the sentiment from past experiences creates the preferences that likely contribute to that decision. A relevant conclusion from the study of a human who cannot connect thoughts and emotional aspects of the brain when it was damaged, showing they could make a rational decision but were not able to decide as they cannot sense how they felt about the options (Mosca, 2000). The following points highlight the influential role of emotion in consumer behaviour:

a. EGG imagery of the brain primarily shows consumers using their feelings instead of brand information or product knowledge, to make a purchasing decision (Vitay & Hamker, 2011). This evidence shows the direct relationship between emotions processing and cognitive processes.

b. Advertisement research has also revealed the direct correlation between emotional response to mass media, with consumers registering their intention to buy a product when an emotional appeal is provided in the television ads when comparing to print ads szirtes2017behavioral.

c. Positive sentiment towards a brand is an essential component of customer loyalty than other rationale judgements on the product attributes (Pang, Lee et al., 2006).

1.2 Aims of this thesis and Research Questions

Based on the rationale and motivations discussed previously, this work will focus on the development of novel methods and models for recognising human's emotion using all aspects of the cognitive, facial expression and physiological signals (body temperature, heart rate etc.).

Therefore, the research into the novel emotion-sensing or emotion Artificial-Intelligent (AI) framework must embody the characteristics of the followings:

a. Efficiently model the temporal, Spatio-temporal elements of the changing emotions or expressions through time and space.

b. Biological inspired machine intelligence that integrates brain-data with emotional information based on experience

c. From unimodal to multimodal, like how our brain can integrate different modalities, such as sound and vision into one integrated system.

d. Able to personalise the sensing of emotion that can learn and represent the sophisticated emotional understanding in a continuous online multisensory time-space.

e. Allows fast and massive information processing of changing human emotions depending on the complexities of feeling, memories, experience etc.

Neural Networks have provided a contribution to advances in the scientific study of the nervous system during the last few decades. An artificial neural network (ANN) refers to a system of interconnected neurons, which processes information by their activities in response to external inputs. ANN provides us with robust techniques in solving real-world problems relating to pattern recognition, time series prediction, data processing robotics etc. The third generation of ANN, Spiking Neural Networks (SNN) has attracted increasing interests over the last thirty years. The reason for that, SNN and the evolving class of the model allow spiking neurons to be created

and merged incrementally, capturing incoming data patterns connecting the new and the old, adapting to the fast trained while capturing meaningful information and turning into new knowledge. Therefore, SNN embodies all the essential motivation of the novel research into an intelligent sensing machine.

In that light, this research work hopes to address the following research questions:

1. **Research Question 1.** How to design architectures of spiking neural networks that are capable of efficiently model the temporal, Spatio-temporal elements of the changing human emotions or expressions? The research question focuses on software design components, including a review of the recent developments in the still-off-the-mainstream information and data processing area of SNN – the third generation of artificial neural networks.

2. **Research question 2**. How to perform neural encoding on facial expression and physiological data to represent human emotions as timings of spikes? The research question explores the use of the 3D structure obtained from research related to research question 1; study affects recognition in artificial systems. The initial approach involved unimodal data such as facial expressions or physiological data such as EEG alone.

3. **Research question 3.** How to integrate spatial, temporal and event-related potential present in unimodal and multimodal human physiological data using spiking neural network architecture? The research question explores the similar brain-inspired 3D structures, fusing temporal, spatial and event-related potential (ERP) information in a spiking neural network architecture

The first research motivation is to understand how SNNs work. This is done through an extensive literature review. There is also a need to understand the various mathematical models available for spiked neural networks. Also, the different coding mechanisms involved are to be studied in detail. The possibility of 3D Spatio-temporal structures is then investigated. Once the viability of SNNs as a neural

network building block in affective computing is established, the study then moves on to addressing other literature gaps. The ability of such structures to effectively recognise Affects from unimodal data (such as facial expressions or physiological signals or voice) is to be investigated. The reliability of such results can be further improved by using the same network structures for the event-related potential for multimodal affect recognition. For example, the possibility of using such 3D structures for multimodal affect recognition is yet to be fully understood. This is achieved by building an SNN based neural network that can use a combination of facial and physiological data to identify the Affect state.

It should be noted that the use of SNN structures for affect recognition, whether unimodal or multimodal, is not sufficiently understood from a review of existing literature. Given the rising importance of affective computing and artificial neural networks, the importance of developing such an understanding cannot be overstated. This is especially true with regards to 3D spatiotemporal structures that can effectively mimic a biological brain

By addressing the next two research questions, this research will help further the scientific understanding of affective computing and the role of neural networks in the same. It is also essential to develop an understanding of other areas related to spiked neural networks. For example, it is necessary to have a thorough knowledge of the history of neural networks before proceeding to model SNNs. This understanding is essential as it allows the researcher to understand the unique characteristics of SNN and the importance of such attributes in the context of affective computing.

Similarly, understanding how unimodal and multimodal Affects are handled by the proposed 3D structure allows the researcher to generate data on which approach is efficient and whether the additional complexity of multimodal approach is sufficiently justified in terms of accuracy improvement obtained. Similarly, it will also be interesting to note how various affective modes such as facial emotion recognition and emotion recognition using physiological signals fare against each other when

a similar neural structure is used for analysis. This research, therefore, seeks to contribute to the existing literature by enhancing the understanding regarding SNNs, specifically three-dimensional structures that are constituted of SNNs. There is also much to be learnt about how such structures process both unimodal and multimodal data, and this research thesis will, to some extent, address this specific literature gap. The work also offers the possibility of comparing accuracy data from unimodal and multimodal analysis using the same neural structure.

1.3 Thesis Structure

The presentation of this thesis is according to format two – the manuscript structure format stipulated in the latest AUT postgraduate handbook 2019/2020. Figure 1.1 depicts a visual explanation and a bird's eye view of the different components of the format two. The components of the thesis format come with different chapters, topics, research questions, four manuscripts chapters that are each preceded with an introduction/prelude and a final chapter that discussed the future direction of the research. Each manuscript chapter provides the status of the peer-reviewed publication in the following order:

a) Acc-Man – Paper accepted for publication,

b) Sub-Man – Paper submitted for publication

c) Appr-Man – Paper approved (by the PhD supervisor) for publication.

The four manuscripts describe the research outcome that answers the three research questions. The first manuscript is a research review paper on Spiking Neural Network that includes a technical study about KEDRI's NeuCube that is developed by Knowledge Engineering, Discovery and Research Institute (KEDRI) of the University.

The next three research manuscripts centred on the development of novel methods and processes using the Brain-inspired Spiking Neural Network methods and

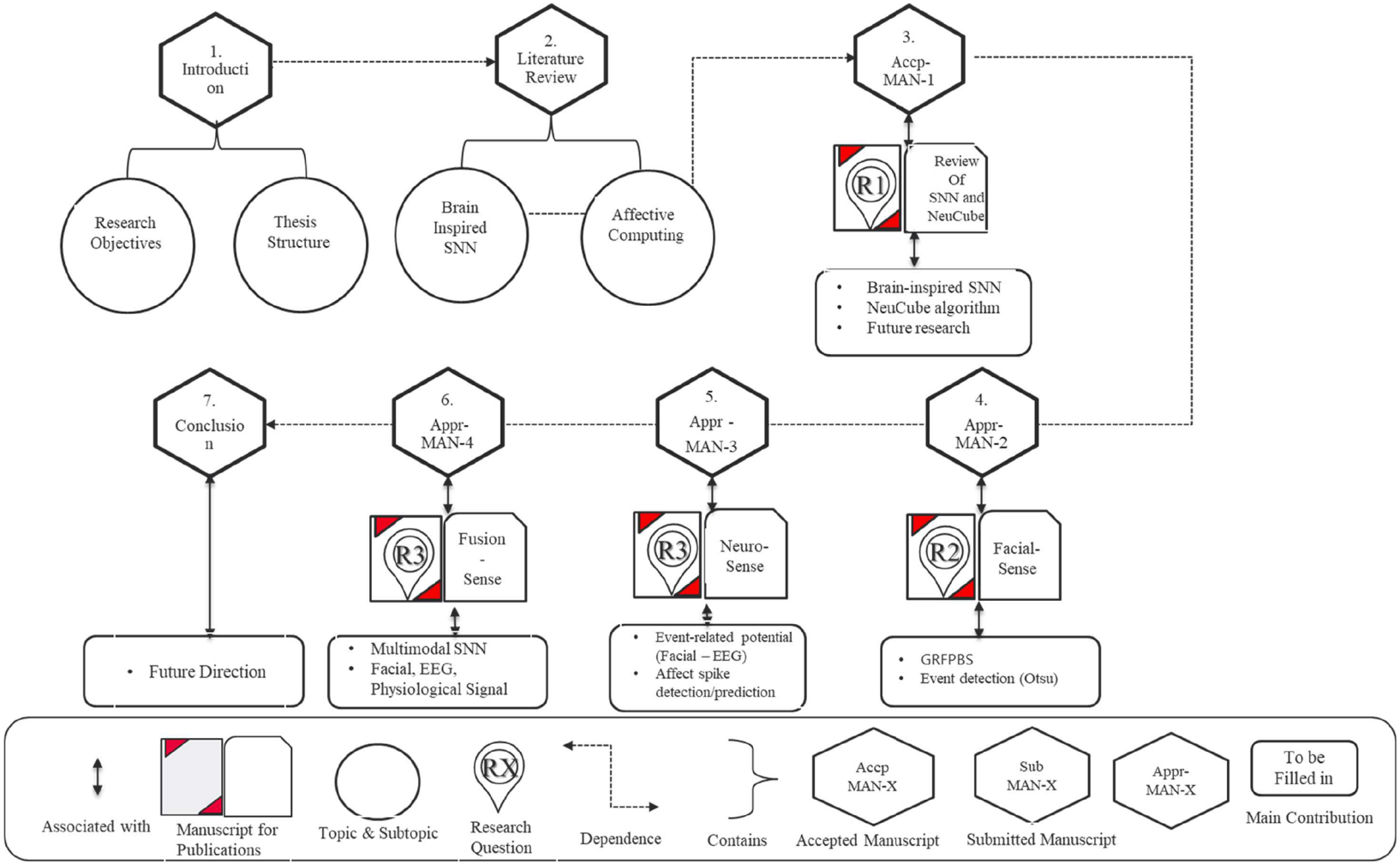


Figure 1.1: Bird's eye view of thesis (format 2) Components and their inter-depedence

infrastructure into emotion recognition using the facial expression and physiological process of human beings. The physiological process includes heart rate like electrocardiogram or ECG signal, respiratory rate and skin conductance like EDA signal and brain activity like electroencephalography or EEG signal. In the research thesis, all the physiological signals were used for emotions recognitions, either with unimodal or multimodal features selection signals. Each manuscript presents the answer to the research questions (R1 to R3), highlight the research outcome, explain the limitations and explore future research directions.

The structure of the thesis provides a continuous flow of the research questions that are centred on the emotion recognition using brain-inspired Spiking Neural Network. The chapter starts with the introduction, explaining the research objectives and the motivations. It is followed with a literature review of affective computing and SNN with a review of the methods and architecture of NeuCube, and it is potential for emotion recognition. Once the potential of the using SNN is established, the next three chapters include research into emotion recognition using Facial (FacialSense), EEG (NeuroSense) and multimodal fusing of facial and physiological (FusionSense) signals. Mainly, the three papers provide thorough research and novel application of SNN for emotion recognition.

2 A review of Affective Computing

2.1 Introduction

The interaction between humans and computational devices are becoming more and more common with the advent of personal digital devices, wearable systems, and other technological interventions. The field of affective computing combines veins of computer science and emotion investigation to essentially enable computational systems to identify the emotional state of users (affect recognition) and to generate responses that humans are likely to perceive as emotional (affect generation). It has also been argued that over time, it may be possible for systems to actually 'feel' emotions (Picard, 1995). This early work by Picard was then followed by much research bringing together diverse fields such as science, ethics, psychology, and engineering, among others. Today, the work on affective computing has resulted in systems that are capable of interpreting, identifying, and responding to the emotional states of users. For this purpose, affective computing makes use of various inputs such as facial images, voice data, biometric data, and body language of the user to identify the emotional state. Computational models termed as 'affect models' are then employed to make sense of these input parameters and identify the emotional state of the user (Tao & Tan, 2005). The present state of affective computing is much advanced, and its importance is on the rise given the increased degree of human-computer interactions. In fact, in 2016, it was reported that 77% of US citizens now own a smartphone, while about 15% have access to wearable technologies. Instead of the one-sided interactions that humans usually expect from machines, utilising affective computing can make these systems respond in more effective ways,

making the whole technology experience more satisfactory to the user (Brigham, 2017). Devices often come with built-in sensors that collect user data. The critical challenge in affective computing recognises the emotional state of a person based on the data available. This is achieved using machine learning techniques that can be employed to process speech, biometric data, posture information, etc. Affect models are usually trained on large datasets of relevant data so that they can them be employed for emotion recognition and affect generation.

Affective computing is finding applications in fields ranging from education (Yadegaridehkordi, Noor, Ayub, Affal & Hussin, 2019) to medicine (Luneski, Konstantinidis & Bamidis, 2010) and gaming (Luneski et al., 2010) and gaming (Guthier, Dörner & Martinez, 2016). For example, research shows that Affective computing systems may aid in the diagnosis of seemingly hidden medical conditions such as depression and chronic pain (Aung et al., 2015). Similarly, Affective systems can provide more empathetic, personalised feedback to students, making online learning more efficient (Grafsgaard, Wiggins, Boyer, Wiebe & Lester, 2013). In that light, this thesis explores the present state of affective computing research in the field. Various input streams or modalities such as voice, facial expression, physiological changes, body language are studied first, and the research on affect recognition in each of these modalities is discussed. The second stage of affective computing is giving computational systems the ability to generate 'affects'. Thereby simulating an appropriate emotional response to the user's present emotional state. This aspect is also discussed. Particular focus will be placed on the machine learning/artificial intelligence-based models that are employed for enabling affective computing. The shift from relying on a single modality (unimodal) to combining various input streams (multimodal) for recognising emotions will also be covered in detail.

This thesis reviews the work done in affective computing by discussing thesiss related to affect recognition, affect generation, and the recent trend of multimodal

analysis. The next section will discuss Affect recognition using various input modalities. A discussion on affect generation will then follow this before moving on to multimodal affect recognition. It is hoped that this review will contribute to existing literature and highlight the literature gaps that exist in the field to further research in those directions. Affective computing is based on understanding emotions and being able to quantify and classify them. Therefore, any review on affective computing is incomplete without a basic understanding of emotion models in Figure 1 that are commonly used in affective computing. The next section explores the models commonly used to study emotions and how they can be used to model emotions in an affective computing system.

2.2 Emotion Representation

Human emotions are often treated as abstract entities and quantifying them pose some unique challenges. However, some approaches and theories are available for the same, and these theories form the basis of affective computing. The first step to affective computing has the ability to identify and quantify emotions. It is, therefore, essential to discuss the various methods proposed in the literature to model emotions. The oldest and most extensively used model for quantifying emotions is the one proposed by Russell (1979). Russell defines the entire Affective space as bipolar. Russell, Lewicka and Niit (1989) presented a more detailed representation of the same model as a Circumplex of Affect, a circular pattern that allows every emotion to be placed in a two-dimensional Affect space, namely arousal and valence. Within affective computing, the two-dimensional models by Russell are popular given its simplicity, and it is the ability to represent diverse emotional states. However, there are other models of human emotions available as well. A model by Ekman (1992b) proposes six standard emotions anger, joy, fear, sadness, disgust and surprise. With regards to affective computing, this model of classifying emotions as discrete and specific entities have certain advantages. From here, these

emotions can further be divided into specific categories, namely boredom, confusion, joy, flow, and frustration. Of this affective computing can choose specific categories that are of interest to a specific application. For example, a student is likely to experience boredom, but less likely to experience disgust or fear. Therefore, an application in digital education that uses affective computing can concentrate more on the emotions required and ignore the rest. Another model not commonly used in the field of affective computing is the navarasa model, which classifies emotions into nine primary states (Sreeja & Mahalaksmi, 2015). This model, however, is not primarily explored in the field of affective computing, given its vagueness and non-numerical nature.

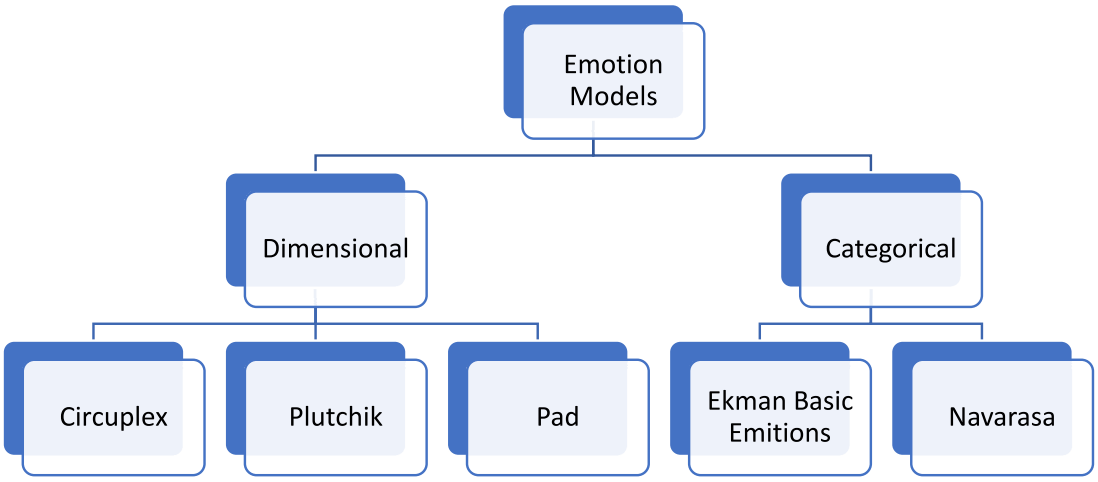


Figure 2.1: Emotion Models (Adapted from Sreeja and Mahalakshmi (2017))

Other dimensional models similar to the Circumplex model, which can be used for affective computing include Plutchik model and pad model. The Plutchik model also uses a two-dimensional space involving activation and evaluation. Similarly, Meharian (Sreeja & Mahalakshmi, 2017) shown in figure 2.1 presents a three-dimensional model that uses pleasure, arousal, and dominance to represent any emotional state. Given the dimensional nature of these models, they are better suited for affective modelling. This understanding of emotional states and the ways of representing them are essential for understanding affective computing models. These models form the basis of affective computing systems, as they provide a

methodology for standardising and quantifying emotional states. The next step is to understand how affective computing systems understand and recognise emotions. The various inputs that can be used to identify emotions and the methods for categorising the same are discussed in the next section.

2.3 Modalities and Affect Recognition

Different modalities can serve as input to an affect recognition engine. They can broadly be classified as facial emotion recognition, voice-based emotion recognition, gesture-based recognition, and emotion recognition based on physiological data such as ECG, EEG, body temperature, etc. These modalities will be discussed in detail in this section with specific stress on works using machine learning and artificial intelligence techniques for affect recognition

2.3.1 Facial Affect Recognition

Emotion recognition based on facial expressions is a keystone of human interactions. It is essential for maintaining human relationships, and poor facial recognition in a human being is usually associated with an inability to interact effectively in social situations (Wolfkühler et al., 2012). Naturally, when trying to make human compute interactions more humane, emotion recognition based on facial expressions play an essential role (Schiano, Ehrlich, Rahardja & Sheridan, 2000). Using machine learning and artificial intelligence, many models that are capable of recognising human facial expressions or “facial affects” have been proposed. Some of the earliest works in this field utilise neural networks to classify facial expressions into discrete, predefined human emotions. A system proposed by Avent, Ng and Neal (1994) successfully identified interest, happiness, sadness, surprise, anger, fear, contempt, and disgust with accuracies ranging from 68% to 89%. Gargesha, Kuchi and Torkkola (2002) reported the use of Artificial Neural Networks to identify and

classify emotions from facial expressions. The thesis also introduced a methodology to automatically capture relevant facial expression data from an image instead of relying on manual annotation of characteristic facial points. This study utilised the Japanese Female Facial Expression Database (JAFFE) alone to make the dataset more uniform for testing. The model utilised multilayer perception within the neural network and achieved an accuracy of 73%. By 2005, machine learning algorithms such as AdaBoost (Adaptive Boosting) allowed for average accuracies of close to 95% for facial affect recognition, even when working in real-time (Bartlett et al., 2005). At present, affect recognition accuracies close to 100% have been reported by employing artificial intelligence algorithms such as random forest (Wei, Jia & Chen, 2016) or advanced neural network structures (S.-H. Wang, Phillips, Dong & Zhang, 2018). It should be noted that the strides in facial affect recognition is mainly aided by the presence of large-scale datasets such as Karolinska Directed Emotional Faces dataset (KDEF) and Extended Cohn-Kanade (CK-Plus) that can be used to train the artificial intelligence-based algorithms to recognise emotions. In addition, regional data sets such as the JAFFE discussed earlier are also necessary when applications are to be tailored to a specific ethnic group or face type.

2.3.2 Voice Affect Recognition

Emotion recognition from the speech is widespread as speech is often more efficiently and economically obtained than modalities such as body language or physiological inputs. Besides, speech Affects are preferred for applications in e-learning, tele-marketing, etc., making it a viable area of research (Kerkeni et al., 2019). Much like with facial affects, voice affects identification is also reliant on having an excellent emotional speech database, and the ability to recognise and identify appropriate affects from a given speech input. A study by Balakrishnan & Rege, (ND) utilises the Sustained Emotionally coloured Machine-Human Interaction using Nonverbal Expression project (SEMAINE) to test two models, one utilising a basic neural net

and another based on a convolutional neural network to analyse speech emotions. In comparison, the convolutional neural network exhibited better accuracy. The thesis argues that while basic neural networks identify and remember emotional patterns at the beginning of each audio interval. CNN allows for learning throughout the operation, thus minimising the risk of 'forgetting' and thereby giving better performance. This has given rise to recent popularity of CNN algorithms in voice affect recognition. Another thesis by Alu, Zoltan and Stoica (2017) uses CNN for voice affect recognition, specifically targeting it for uses in robots. The model reported an accuracy of over 70% and is targeted at applications in assistant robots and intelligent driving assistance systems. For example, Tzirakis, Zhang and Schuller (2018) presents a CNN based approach which takes a raw speech waveform and identifies the emotional state of the individual with greater accuracy than other established methods. Other Machine learning algorithms and neural networks have been used extensively for this purpose. When machine learning techniques are concerned, long-short term memory recurrent neural network (LSTM RNN), and Restricted Boltzmann Machines (RBMs) have been discussed extensively in the literature for recognising voice affects (B. W. Schuller, 2018). It should be noted that unlike with facial affect recognition, voice affect recognition depends on the language and dialects of the speakers. Therefore, it is essential to train the models using datasets appropriate to the applications it is designed for (Sohn , (Shon, Ali & Glass, 2017).

2.3.3 Body Language/Gestures

Body language and gestures are universal indicators of emotions. The information conveyed through hand gestures, the position of the head and inclination of the body, in general, are all rich sources of data on emotion recognition (Noroozi et al., 2018). Body gesture-based emotion recognition, much like facial affect or voice affect recognition, utilises existing databases to identify specific affects by using

tools of machine learning and artificial intelligence. However, when compared to the literature available on facial and voice affect recognition, emotion recognition based on body language is yet to receive the same scientific attention. This area is further challenged by the degree of subjectivity involved in establishing benchmarks. A neural network-based approach using data from 50 actors across body types and ethnicities in different poses provided over 87% accuracy despite the varied input dataset. It should be noted that while the absolute accuracy is 87%, the accuracy compared to human emotion recognition was close to 95% in this case. Therefore, even with basic neural network structures, it is possible to assess human emotions from body language with near-human accuracy. Recently, machine learning techniques and crowdsourcing methods have recently been applied to gesture-based affect recognition (Zacharatos, Gatzoulis & Chrysanthou, 2014). A recent study by Santhoshkumar and Geetha (2019) uses a feedforward deep convolution neural network architecture to identify emotions based on body language from videos and achieves its objective with an accuracy of over 95%. It should be noted than an earlier attempt by Gavrilescu (2015) only resulted in an efficiency of over 86%, signifying the drastic improvement this field of affect recognition is currently undergoing. Such emotion recognition from body posture has additional applications in realistic gaming environments, and in allowing neuro-atypical individuals to better integrate into society (Piana, Stagliano, Odone, Verri & Camurri, 2014).

2.3.4 Physiological data-based affect recognition

It has been argued that as physical expressions of emotion such as facial expression and body language can be faked to a certain degree, physiological data such as EEG, body temperature, and ECG are more reliable indicators of affect recognition (Shu et al., 2018). A canonical correlation relating these signals to emotions alone could predict the correct emotions from physiological data, with over 85% efficiency (L. Li & Chen, 2006). By utilising more advanced methods such as artificial intelligence

approached and machine learning algorithms, much higher accuracy rates can be achieved. Harper and Southern (2019) for example, used Bayesian Neural Networks to analyse the heartbeat data obtained by a fitness tracker to identify the emotional state of the user wearing the tracker. Heartbeat data is obtained as a photoplethysmogram. While the method was effective, this study does point to the need of better physiological datasets for training the neural networks to identify emotions. Similarly, P. Sarkar and Etemad (2019) uses ECG data to determine emotions. The method proposed essentially consists of two neural networks, one for signal transformation recognition and another for identifying emotions. The first signal transformation recognition network is trained using an extensive dataset to identify specific transformations in the input signal. The next step is to use the second network to identify the emotions associated with each transformation successfully. Two publicly available datasets (AWELL and AMIGOS) were employed in this study, and the results show that this two-network model performed better than the traditional approaches, reporting accuracies close to 98%.

These studies clearly show that various modalities ranging from facial expressions to physiological data can be used to identify the emotional state of the human being during a human-machine interaction. Some algorithms can identify these affects based on appropriate input data. Besides, there are datasets that can be used to train these models in emotion recognition. However, this identification forms only the first step in making human-machine interactions more efficient. The next part necessitates that machines can respond appropriately once these emotions are identified.

2.4 Affect Generation

In order for affective computing to be useful, it is not only necessary for machines to recognise the emotions of the user. Once these emotions are identified, it is essential that the systems respond in a way that mimics the human response to

such emotions. In essence, the system should be able to generate affects that mimic human emotions.

2.4.1 Generating Facial Affects

Given that human facial expressions are the most significant factor in human interactions, much effort has been directed towards making robots, virtual characters, and similar synthetic systems capable of facial affect generation. Therefore, there are facial animation models that can replicate the facial affects of humans. An example is the MPEG4 facial standard which has a total of 68 defined Facial Animation Parameters (FAP) that are related to different parts of the human face. These can be used to animate realistic faces of any colour, size, and ethnicity (Gachery & Magnenat-Thalmann, 2001). However, accurate representations of human facial expressions are still challenging. It has been noted that extreme expressions such as outright laughter, screams, etc. which are relatively easy to replicate are rare in day to day interactions. Instead, more subtle indicators of emotion (such as blushing, sweating, etc.) are necessary to represent emotions accurately (Alkawaz, Basori, Mohamad & Mohamed, 2014). In their work, Alkawaz et al. (2014) simulate sweat, skin colour, and tears to allow avatars to have a realistic emotional range, be it with mild or extreme emotions. Recently, a lab in Cambridge reported that a robot named Charlie who was created to recognise human emotions has now been trained to generate facial affects mimicking humans as needed Laybounr (2018). Though realistic, the robot, however, cannot emulate the nuances of human facial expressions to perfection, and it is reported that humans find interacting with the robot strange. Similar is the response to Sophia, a robot made specifically for human interactions. Sophia uses a convoluted neural network to articulate information through multiple inputs such as a camera and hearing devices. Once the emotions of the human are recognised by the neural network, an appropriate facial affect is generated by triggering Sophia's facial muscles (Hanson Robotics, 2019). Therefore,

while facial affect generation is an advancing field, it needs further research before synthetic characters can realistically generate facial affects. While strides have been made in fields such as gaming and other virtual interactions, humans can still easily recognise human facial affects from generated ones (Hornyak, 2018).

2.4.2 Generation Voice Affects

There is significant commercial potential in voice affect generation. For example, such systems can aid in telemarketing, phone banking, and many such applications. Therefore, much attention has been focused on voice affect generation. Rank and Pirker (1998) discussed the possibility of generating emotional speech based on parameters such as articulatory precision, frequency, amplitude, energy distribution, etc. The study cannot be classified entirely as intelligent affect generation, as the system is not based on any artificial intelligence technique. However, it does highlight the potential for voice affect generation once a specific emotional state is identified. Rao (2011) explored the role of neural networks in speech synthesis and identified that neural networks could be employed to identify the positional, phonological, and contextual aspects of speech, allowing for speaker recognition and emotion identification. These neural networks can then be trained to generate an appropriate affective response based on the target speaker and his/her emotions. Later, L.-H. Chen, Ling and Dai (2014) proposed a method where deep neural networks were employed. A four-layer neural network is trained on various aspects of target speaker characteristics and source spectral envelope, thereby generating superior voice affects. In more recent times, commercial applications of such affect generation have become popular. A Chinese tech company recently claimed that it could clone a human voice based on just 3.7 seconds of recorder voice data. Wavenet uses deep learning and a high multilayer neural network to make google digital assistant sound human, thus bringing voice affects to the commercial space (Balaban, 2019). As speech generation becomes more mainstream, voice databases

that can be used for training artificial intelligence systems to recognise, identify and quantify emotions also become more focused, covering various languages, accents, and dialects (Adigwe, Tits, Haddad, Ostadabbas & Dutoit, 2018). It is safe to assume that in the near future, systems with artificial intelligence will be able to generate voice affects with much greater precision that they can generate facial expressions.

2.4.3 Generating Gestures and Body Language

Recently, significant advances have been made in allowing artificial systems (both robots and virtual characters) replicate human body language. In a thesis titled 'Everybody Dances Now' (Chan, Ginosar, Zhou & Efros, 2019), a team from University of Berkley presents a case where movements can be realistically transferred from a real 'source' person to a virtual character with great precision, allowing for virtual characters to have hyper realistic body language. The model proposed first trains a virtual stick figure to mimic the movements of a source subject. The frame can then be transferred to any virtual avatar, thus generating realistic body language affects. It should be noted that Chan et al. (2019) do not directly discuss emotions or how they are communicated through body language. However, the work does highlight the progress in replicating body language realistically. On work-related to affect generation, Xu, Broekens, Hindriks and Neerincx (2015) vary a set of physical movement parameters on robots to mimic positive and negative human body language. While interacting with humans, the humanoid robot NAO exhibited these traits and human beings participating in the interaction were able to identify the traits and responded accordingly. NAO expresses that body language affects by using an affect space, which consists of a set of key poses and combinations of these key poses. For example, it is possible to blend the posture-related to sadness with pride in a fixed ratio. It was seen that depending on the specific blend, the success of human participants in identifying the emotional state ranged from 100% to 35%. Given that the robot NAO does not have the ability to express facial affects,

these results are highly impressive (Beck, Hiolle, Mazel & Cañamero, 2010). Body language and gesture-based affects are, therefore, fast becoming a part of human-machine interactions. As robots and other virtual agents become more common, it is likely that realistic affect generation will be a part of regular human-machine interactions (Xu et al., 2015).

2.5 Multimodal approach to Affective Computing

The thesiss discussed so far, both in affect recognition and affect generation, concentrate on a single modality. For example, there are works related to facial affect recognition, or body language affect generation. However, in a normal social situation, when two human beings interact, emotions are conveyed in a multimodal manner, using multiple channels such as voice, facial expression, physiological responses (such as sweating), and body language. Therefore, there is now a shift towards replicating this multimodal approach in affective computing as well. Though complex, such as approach has many advantages and is likely to generate more realistic results. This section explores the recent shift from unimodal to multimodal affective computing, discussing the key advantages associated with a multimodal approach.

A multimodal approach studies any one aspect of human emotional expressions such as a facial expression or voice modulation. A multimodal approach shown in figure 2.2, is a framework built of various good performing unimodal systems. The significant new intervention in multimodal affective computing is the introduction of multimodal fusion techniques. Data from different modalities are combined for analysis, either at the feature level or at the decision level. In essence, data is processed as a combined entity in feature level fusion while it is split into modalities and then analysed in decision level fusion (Poria, Cambria, Bajpai & Hussain, 2017a). Some algorithms use a combination of both these approaches or rely on specified rules to fuse modalities. Much like with unimodal systems, multimodal datasets such

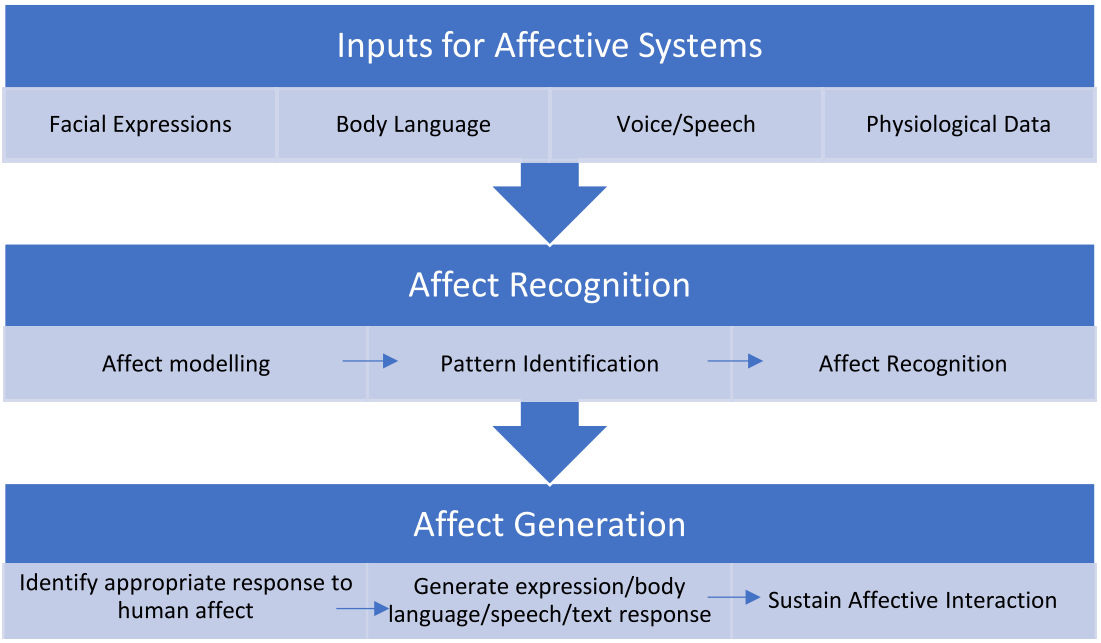


Figure 2.2: Overview of Affective Computing

as the Youtube dataset, SEMAINE dataset, HUMAINE dataset, and the ICT-MMMO dataset are now available to researchers (Poria, Cambria et al., 2017a). The survey presented by Poria, Peng, Hussain, Howard and Cambria (2017) clearly illustrates how multimodal approaches can outperform unimodal approaches. The study also highlights the need for more research in this area, especially when dealing with synchronisation of multiple modalities, the effect of affects in conversation, and the development of more agile processing approaches for real-time results. The real-time processing speeds are essential in affective computing if the hope is to achieve humans like interactions between humans and machines. Therefore, future research in the field of multimodal affect generation is of great significance. There is also very little information on combining physiological indicators of emotions with other modalities in affective computing. These reasons lead the researcher to conclude that there is still much to be achieved before real affective computing is made possible, where humans and machines that can detect and generate real multimodal affects can interact.

However, this is not to say that there is already a significant volume of work on the topic. citetlisetti2002maui proposes an adaptive multimodal system that

combines visual, auditory, and autonomous nervous system signals from a user to recognise affects of the user. A humanoid multimodal interactive agent then responds by generating appropriate affects in response. This study, though highly theoretical, shows promising results and acknowledges the possibility of real-time multimodal interactions between humans and computer-generated entities. Again, the challenges of real-time multimodal interactions are highlighted in this work. The applications of such affective systems are highlighted in work by C. Lisetti, Nasoz, LeRouge, Ozyer and Alvarez (2003). The thesis considers an example in telehealth, where the service providers usually have the challenge of assessing a person's emotional state from afar when only limited modalities (such as voice, or physiological data) are available to them. An affective computing system for telehealth using wireless, wearable computers that collect multimodal data is discussed. The system will also have empathetic digital avatars that are capable of affect generation to allow for effective patient interaction. The data thus collected will be more useful in assessing the patient's emotional state, thus leading to better health care access. Similar applications can be found in digital education, advertising, and multiple other fields.

While applications are diverse, much is yet to be learnt about multimodal information and how it can be processed effectively. Social media to a large extend has democratised the availability of multimodal content. However, multimodal analysis frameworks are still in the development stage. A work on multimodal affective computing by Gu et al. (2018) uses a hierarchical multimodal architecture to evaluate multimodal data. The modalities employed are textual and audio data which are synchronised. Using feature-level fusion, the method uses a convoluted neural network to analyse the combined data. The model was tested on YouTube dataset, IEMOCAP, EmotiV, MOSI, and exhibited an accuracy close to 77%. While not high, it should be noted that this study only handles two modalities, and therefore, limited data is available for analysis.

However, in recent times, the accuracy of multimodal systems have improved significantly still. A framework proposed by (Poria, Cambria, Hussain & Huang, 2015) using the eNTERFACE dataset achieved an accuracy of nearly 88%, thus outperforming most existing systems. This framework employs multiple kernel learning algorithms to analyse the data available and achieve this relatively high degree of accuracy. The proposed architecture combines auditory, visual, and textual modalities to achieve the reported accuracy, beating the existing systems by at least 10% (Poria, Cambria et al., 2017a). The study also highlights the importance of further exploring decision level fusion as it is less studied that feature level fusion.

The thesiss discussed clearly illustrate the advantages of multimodal affective computing over unimodal designs. With multilevel architecture and appropriate fusion methodology, the multimodal approach provides for the possibility of realistic human to machine interactions in real-time. However, much more work is needed in effectively integrating various modalities and handling the vast quantities of data involved. The next section discusses the potential applications of multimodal analysis, and the challenges to be handled before accurate multimodal analysis is achieved. Besides, the potential areas of future research will also be identified based on the thesiss surveyed discussed in this review.

2.5.1 Challenges, Applications and Future research

The thesiss so far in this review highlight the importance of affective computing for future technology interactions. It can safely be concluded that affective computing systems that are capable of identifying and generating affects will have a significant role in future technologies. It is also likely that future developments in the field will be concentrated on multimodal analysis as a great deal of current research focuses on the same. When compared to unimodal affective computing, the multimodal approach is more likely to generate realistic human-machine interactions. This is because, much like human-human interaction, this multimodal approach considers

multiple inputs such as facial expressions, physiological characteristics, and body language simultaneously for affect recognition. Similarly, as research progresses, affect generation systems are becoming multimodal as well as mimicking human-like facial expressions, body language, and voice. However, this is not to say that multimodal affective computing is without challenges. Significant research is needed in the fusion techniques that can be used to combine multimodal input data.

Similarly, multimodal datasets that are to be processed are highly data-intensive, making it challenging to handle real-time interactions. Therefore, a multimodal system that is computationally fast enough for real-time affect recognition and generation will have far-reaching applications. Finally, most multimodal systems ignore the possibility of including physiological data while recognising affects. Physiological data is becoming more and more accessible given the rise of wearable devices such as fitness trackers. Also, unlike body language or facial expressions which can be misleading, physiological indicators are always a predictable indicator of the mental state of the person involved. Therefore, including physiological characteristics, along with other modalities, can significantly improve the accuracy of multimodal affective computing.

Multimodal affective computing models are likely to find applications in fields ranging from education to geriatric care. A study by Rivera-Hernández, Stoyanov, Tsolaki and Ramón (2013) posits that affective computing systems can better the life of elderly patients by adding more realistic and varied interactions. Such systems may reduce the rates of depression and Alzheimer's disease among the elderly. Similarly, detecting the affective states of learners and presenting educational materials accordingly can enhance the learning outcomes in virtual classrooms (C.-H. Wu, Huang & Hwang, 2016). Such systems can also be used to better tailor advertisements to each specific user, and to enhance the chances of the right product reaching the right customer (Adibuzzaman et al., 2013). Another critical area where multimodal affective computing is finding application is in online gaming. With

massively multiplayer games getting more realistic, the industry works towards making avatars more affective, responding to the situation and emotional states of the players (Psaltis et al., 2016).

Therefore, it is safe to conclude that multimodal affective computing systems that are capable of recognising affects and generating appropriate responses have significant market potential in the coming years. Research in this area, including modalities that are less explored, such as physiological data can add to the existing knowledge in the field while simultaneously finding applications in diverse fields.

In all the cited works above the main goal was to classify emotions from different input data. Very few of them tried to explain what happens in the human brain when experiencing different emotions. The presented thesis aims at both good classification and understanding the affect data and human brain activities, and that is why a brain-inspired SNN architecture is utilised for this purpose, as explained in the next chapter.

3 Spiking neural networks: background, recent development and the NeuCube architecture

3.1 Prelude to Chapter 3 Manuscript

Artificial intelligence is at the forefront of human development in the present century. Intelligent systems presently engage in activities ranging from telemarketing to remote healthcare. Therefore, there is a constant need to emulate the biological ability to process information and learn in computer networks. In that light, neural networks are systems that endeavour to mimic the brain and have the ability not just to process information like a standard computational system, but also to 'learn' to perform specific new operations without being programmed explicitly for that task. The idea of neural networks has been in existence since the early 1940s, and multiple generations of neural networks have been developed over time. The first generation involved neurons that used relatively simple methods such as weighted sum method. These were followed by the second generation that utilised Sigmoid neurons that utilise a smooth differentiable sigmoid function to the weighted neuron input. These neurons can then be stacked to generate deep learning systems. While such artificial neural networks (ANNs) have impressive capabilities, they fail at emulating the functioning of a biological neural network. They do not exhibit the nonlinearity usually associated with biological neurons. Besides, neurons communicate using spikes of energy (action potential), while ANNs use continuous signals as in ANNs. These fundamental differences between biological systems and ANNs gave rise to the third generation of neural networks termed as spiked neural networks or SNNs.

It should be noted that even with these advancements, artificial neurons are outperformed by biological neurons in efficiency and energy requirements. These third-generation neurons communicate using discrete spikes, thus mimicking the biological networks to a certain degree. Neurons transmit information using a series of excitation and inhibition signals. These signals generate electrical potential, which then translates to spikes of energy which primarily contains the information being transmitted. This process is simulated in spiked neural networks.

As the name implies, SNNs utilise signal spikes to encode data. The information is passed in discrete spikes, much like in biological systems. When compared to a continuous signal (as used in ANNs), the spikes allow for temporal information encoded into the system.

One of the objectives of this review is to understand these spiked neurons and their functioning better. In order to do so, a detailed review of the computational methods, models, and algorithms that are commonly employed in SNNs for information processing and learning. The models discussed in this review are Leaky integrate-and-fire model, the Izhikevich model and Spike response model (SRM). In addition, other models such as Wilson model and FitzHugh-Nagumo model are discussed in brief, and the complexity and plausibility of these models functioning like a biological system are studied taking the number of variables also involved into account. The discussion focuses on highlighting the novelty of each model and the basic mathematics involved. Information processing and coding are also discussed, and the types of coding covered are rate coding, temporal spike coding, and synchronous coding.

It should be noted that there is still a lack of understanding of whether biological system employs rate coding or spike coding. Therefore, all three methods mentioned are of significance as more search is needed to understand which best emulates a biological system. Finally, two learning approaches, rate-based learning and spike-based learning are discussed as well. The advantages and challenges of each of

these models, methods, and learning approaches are covered in the review. These discussions will illustrate why SNNs are more successful in mimicking biological intelligence. However, much is yet to be achieved before these SNNs can perform at par with biological neurons. One such advancement that makes SNNs more capable is the ability to configure in three-dimensional structures that allow for functioning that mimics brain behaviour.

This understanding of SNNs is essential for exploring the primary focus of this review, a relatively new structure made of SNNs termed as NeuCube. These structures try to emulate the 3D spatial structure of a biological brain and are critical steps in the evolution of neural networks. These three-dimensional structures, with the process of creating and merging spiked neurons, can understand new patterns in multidimensional data. They are commonly applied to brain data, such as EEG data and fMRI patterns. These structures form a crucial step in allowing SNNs to emulate the human brain and can bridge the gap between human and machine intelligence. Therefore, this paper covers the unique general architecture of these NeuCubes, starting with the input module and its encoding algorithm. This is then followed by a discussion on initialisation of the module, and finally, a detailed analysis of how these structures implement specific output functions to generate the necessary functionality. This discussion will highlight why these SNN structures are better able to simulate and even predict brain activities. It should be noted that this paper does not present a new model of spiked neural networks. Instead, it merely discusses the existing scientific literature on the topic. Such structures are commonly used to understand better the functioning of the human brain, including brain functioning, specific behaviours, and decision-making processes. The similarity to human neural network patterns makes NeuCubes especially significant. Being a new step in the evolution of neural networks, these SNNs will play a significant role in advancing artificially intelligent systems and allowing them to mimic biological neural networks. Therefore, a study of such nature is of critical importance.

In summary, this paper gives a brief description of various neural network generations and the unique characteristics that identify each. The focus of this paper is on third-generation spiked neural networks. Once the models and approaches related to SNN and its compatibility to biological systems are described, the review concentrates on Neu Cubes, a 3D Spatio-temporal SNN-based data machine framework. The study discusses the use of NeuCubes to analyse spectro-temporal brain data in detail. A discussion then presents the key findings obtained from the literature review including the various possible applications of SNNs, the potential upcoming work in this field, and the open research problems that exist before SNNs can be used for mimicking biological systems with greater accuracy. By reviewing literature and identifying the open research problems, it is hoped that this work will lead to more considerable scientific attention on SNNs, thus advancing the field of artificial intelligence.

3.1.1 Contributions and Publications

Contributions

1. *Discusses the development and implementation aspects of SNNs and their similarity to biological systems*
2. *Review of the computational methods, models, and algorithms that are commonly employed in SNNs for information processing and learning*
3. *Analyses why these SNN structures are better able to simulate and even predict brain activities*
4. *Overview of the NeuCube, 3D Spatio-temporal SNN-based data machine framework*
5. *Describe the software design framework of NeuCube as a prototyping and testing environment framework.*

Publications

1. Tan, C., Sarlija, M., & Kasabov, N. (2020). Spiking Neural Networks: Background, Recent Development and the NeuCube Architecture. *Neural Processing Letters*, 52(2), 1675-1701. https://doi.org/10.1007/s11063-020-10322-8

3.2 Introduction

In the parlance of machine learning and artificial intelligence (AI), a neural network can be defined as a network of neurons that are able to perform computations and solve problems. Neural networks for learning as seen today, have come a long way since their discovery in the late 50 s by Hubel and Wiesel (1959) followed by the development of Neocognitron, a neural network composed of multiple layers, by Fukushima in the early 80 s Fukushima (1979). Depending on the type of neurons used, artificial neural networks (ANNs), as they are referred to, can be thought to belong to three generations.

First-generation ANNs are composed of neurons that compute a weighted sum of binary inputs and produce an output of 1 if the sum crosses a pre-defined threshold, else the output is zero (see Figure 3.1a). These neurons are also known as perceptrons Rosenblatt (1958) or threshold gates. Mathematically, a perceptron can be expressed as:

$$y = \begin{cases} 0, & \sum_j w_j x_j \leq \theta \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

where θ is a threshold parameter.

The second-generation ANNs are composed of artificial neurons, also known as Sigmoid neurons, which apply a nonlinear activation function f to the sum of weighted neuron inputs, as shown in Figure 3.1b. The function f is a sigmoid function, which is smooth and differentiable. The primary motivation behind using such functions is to enable the application of backpropagation algorithms, that are based on error function gradient computation, to train the ANNs.

By stacking more than one layer of neurons, as shown in Figure 3.1c, and applying backpropagation to learn multiple layers of representation, deep learning neural networks can be constructed. Deep-learning neural networks are today capable of solving problems in diverse areas ranging from speech recognition Hannun et al.

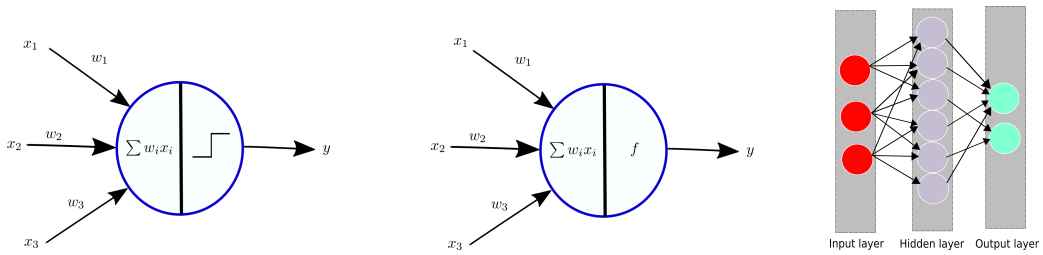


Figure 3.1: Model of a perceptron that uses (a) a step-function to give an output of 1 if the weighted sum of inputs cross a pre-defined threshold and (b) a sigmoid function to give a continuous output based on the weighted sum of continuous inputs. A neural network model with one hidden layer is shown in (c)

(2014), visual recognition Russakovsky et al. (2015), pedestrian detection Ouyang and Wang (2013), recognition of traffic signs Cireşan, Meier and Schmidhuber (2012) to playing GO Silver et al. (2016) and biomedical signal processing Ganapathy, Swaminathan and Deserno (2018); Šarlija, Jurišić and Popović (2017), and can in some cases outperform humans.

As impressive as this feat may be, the fact remains that biological neurons in humans and other animals still outperform ANNs in terms of energy and efficiency. Many deep learning algorithms rely on hundreds of graphical processing units (GPUs) and central processing (CPUs) to solve problems which a human brain can solve in a fraction of these resources in terms of energy. To give an example, it required 1 megawatt (MW) of energy to solve GO challenge using deep learning, whereas the best human players could achieve similar results in about 20 Watts Drubach (2000), which is the power rating of a human brain.

Furthermore, deep learning networks are not biologically plausible. Although initially inspired by biological neurons, the neurons in deep-learning networks solve the learning problem in a fundamentally different manner. Bengio et al. in their recent work discuss problems that arise regarding the biological plausibility of backpropagation Bengio, Lee, Bornschein, Mesnard and Lin (2015). For instance, backpropagation algorithms are not biologically plausible as biological neurons also exhibit nonlinearity, whereas backpropagation algorithm is purely a linear

operation. Another crucial issue is the biological implausibility of ANNs themselves, as biological neurons communicate with each other using discrete spikes known as action potentials and not by continuous values as seen in the implementation of deep learning algorithms with second-generation ANNs. These issues gave rise to the present or third generation of ANNs known as the spiking neural networks (SNNs) Maass (1997b), which use spiking neurons as their computational unit. In the remainder of the paper, ANNs will be used to refer to second-generation ANNs. Accordantly with what is previously said, off-the-mainstream approaches in neural networks and machine learning for pattern recognition have been recently encouraged Trentin, Schwenker, El Gayar and Abbas (2018) in Neural Processing Letters, with an entire special issue being devoted to various off-the-mainstream fields in pattern recognition (associative memories, density-related algorithms, etc.). We believe that the SNNs, being the third-generation of ANNs, as an alternative or a complement to traditional second-generation ANNs, definitely deserve attention when considering off-the-mainstream approaches in neural processing.

The paper is organized as follows. Section 3.3 formally introduces SNNs and section 3.4 briefly reviews some of the popular computational models used in SNNs. Sections 3.5 and 3.6 describe information processing and learning algorithms used in SNNs. Finally, section 3.7 presents a recently developed 3D SNN architecture for mapping, learning and understanding of spatio-temporal brain data called NeuCube N. Kasabov (2014). In the conclusion (section 3.8) we discuss potential applications and challenges in SNNs/NeuCube, and encourage it's further use and development while not staying limited to just brain-data-specific applications.

3.3 Spiking neural networks

The human brain is composed of 10^{12} neurons and each neuron makes about 10,000 connections, known as synapses. The structure of a neuron can be divided into three basic parts: 1) dendrites, 2) cell body or soma, and 3) axons. The dendrites are

nerve cell extensions that process input signals to a neuron via synapses and axon, the thin projection of a neuron, can be thought of as a long process that carries the output signal away from the neuron. The cell body or soma of a neuron contains the nucleus and cytoplasm, as with any other cell in the body.

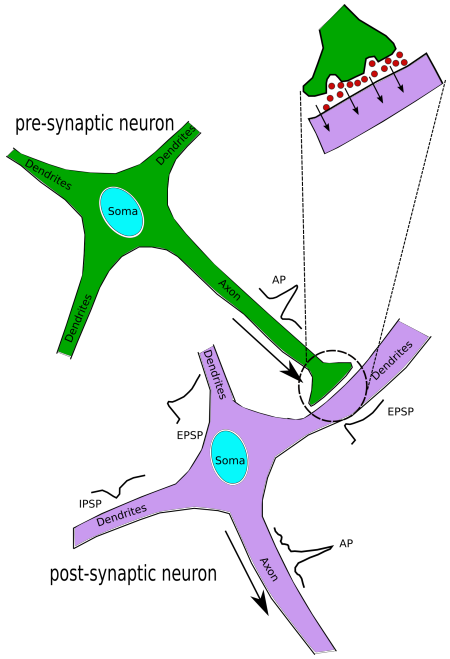


Figure 3.2: Illustration of a pre-synaptic neuron (green) and a post-synaptic neuron (purple) connected through a synapse. Neurotransmitters (red circles) are released at the synapses of many dendrites of a post-synaptic neuron, giving rise to post-synaptic potentials, which are finally summed and a decision to send an action potential via the axon of post-synaptic neuron is made.

In Figure 3.2, what is known as a pre-synaptic neuron is carrying an action potential (AP) or spike through its axon, which causes the release of neurotransmitters at the synapse. The neurotransmitters come in two flavors - excitatory and inhibitory. If an excitatory neurotransmitter is released, positive ions are released into the dendrite of the post-synaptic neuron, causing an excitatory post-synaptic potential (EPSP). If an inhibitory neurotransmitter is released, negative ions flow into the dendrite of post-synaptic neuron causing an inhibitory post-synaptic potential (IPSP). At the soma of the post-synaptic neuron the IPSPs and EPSPs from all the pre-synaptic neurons are spatially and/or temporally summed. When this sum exceeds a threshold, the post-synaptic neuron fires an AP. Thus, the AP can be

thought of as a "currency" with which neurons exchange information, and using discrete spikes like APs rather than continuous signals is what makes the brain energy-efficient. Spiking neural networks (SNNs) are a special class of artificial neural networks (ANNs), also commonly referred to as the third generation of ANNs, where the neuronal units communicate using discrete spike sequences, as exemplified in Figure 3.3. Analogous to a biological neuron, the inputs to a spiking

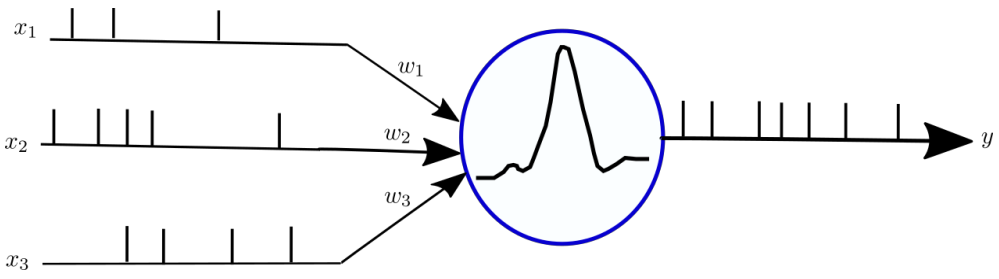


Figure 3.3: Model of a LIF spiking neuron.

neuron are discrete spikes, which are then combined to produce an output spike if a certain threshold is exceeded. Otherwise, the output is zero. Therefore, SNNs also contain temporal dynamics, which makes them suitable for real-time operation, making updates that are purely event- and data-driven, unlike the repeated and often redundant process of updating the weights in ANNs, which is the computational bottleneck for tasks that require real-time interaction with the environment.

3.4 Computational models of SNNs

As mentioned previously, SNNs use spiking neurons as their main computational unit. Numerous mathematical models of a spiking neuron have been proposed in the literature Herz, Gollisch, Machens and Jaeger (2006). Below we briefly review the three most widely used models – 1) Leaky integrate-and-fire (LIF), 2) Izhikevich model and 3) Spike response model (SRM).

3.4.1 Leaky integrate-and-fire model

In this type of model, illustrated by Figure 3.3 and first introduced by Lapicque (L. F. Abbott, 1999; Lapicque, 1907) over a hundred years ago, the membrane is characterised by a resistance R and capacitance C. Let $u(t)$ be the voltage across the membrane at time t and u_{rest} be the resting-state potential. Then, we have the following equation describing an RC-circuit:

$$i(t) = \frac{u(t) - u_{rest}}{R} + C\frac{du(t)}{dt} \tag{3.2}$$

where $i(t)$ is the membrane current. Rearranging the terms in the above equation, we can write:

$$\tau\frac{du(t)}{dt} = u(t) - u_{rest} + Ri(t) \tag{3.3}$$

where τ is the membrane time constant. A spike is generated at time t_f, when the membrane potential reaches the threshold, i.e., $u(t_f) \geq u_{thresh}$. After t_f, the membrane potential is reset to the resting state value u_{rest} and for $t > t_f$ the dynamics are again given by the equation above (3.3). To make this model biologically more plausible, an absolute refractory time δ_{ref} may be included such that the dynamics are restarted as per the equation 3.3 after $t_f + \delta_{ref}$ rather than immediately after t_f Gerstner and Kistler (2002).

3.4.2 Izhikevich model

The Izhikevich model Izhikevich (2003) basically has two variables, u and v that describe membrane voltage and the recovery rate of a neuron. The evolution of the membrane voltage u is described by a pair of ordinary differential equations (ODEs):

$$\frac{du}{dt} = 0.04u^2 + 5u + 140 - v - i \tag{3.4}$$

$$\tau\frac{dv}{dt} = a(bu - v) \tag{3.5}$$

After initiating the action potential, the following resetting scheme is used:

$$\begin{aligned} u &= c \\ v &= v + d \end{aligned} \tag{3.6}$$

when $u \geq 30mV$, which is the peak of the spike Izhikevich (2004). The model parameters, a, b, c and d can be tuned to produce various neural dynamics Izhikevich (2003, 2004). Briefly,

1. The parameter a describes the time-scale of v.
2. The parameter b describes sensitivity of v to the (subthreshold) fluctuations in u.
3. The parameter c describes the resetting of u after the initiation of a spike, which is caused by fast high-threshold K^+ conductance.
4. The parameter c describes the resetting of v after the initiation of a spike, which is caused by slow high-threshold K^+ and Na^+ conductance.

3.4.3 SRM

The spike-response model (SRM) is based on kernel function $K(\cdot)$ and described by a single variable u_i, which is the membrane voltage of the i-th neuron. To describe this model, let us assume that the neuron i is at a resting-state potential of 0 Volts. Incoming spikes from pre-synaptic neurons will affect $u_i(t)$ and after some time, $u_i(t)$ will return to its resting state value. If the net effect of all the incoming spikes is such that $u_i(t)$ crosses the threshold θ, then an output spike is triggered. Assuming that the neuron i has last fired its spike at time t_0, the dynamics of $u_i(t)$ is given by:

$$u_i(t) = \eta(t - t_i') + \sum_j w_{ij} \sum_f \epsilon_{ij}\big(t - t_i', t - t(f)_j\big) \int_0^\infty \kappa\big(t - t', s\big) I(t - s) ds \tag{3.7}$$

where the function $\eta(\cdot)$ describes the form of the action potential, i.e., the spike and I is the external driving current. The term $t_j^{(f)}$ describes the time of the input spike of neuron j and w_{ij} represents the synaptic weight between neurons i and j.

3.4.4 Other models

In addition to the three widely used deterministic models in SNNs described above, several other models of spiking neurons have been proposed and these include Hodgkin-Huxley model (HH) Hodgkin and Huxley (1952), Wilson model Wilson and Callaway (2000), FitzHugh-Nagumo model FitzHugh (1961), Hindmarsh-Rose model Hindmarsh and Rose (1984) and Morris-Lecar model Morris and Lecar (1981) (see Izhikevich (2004) for description). The table below shows a comparison of various spiking models in terms of the number of variables, complexity and biophysical plausibility, which is adapted and modified from Izhikevich (2004).

Table 3.1: Comparison of different models of spiking neurons

Models	No. of variables	Complexity	Biologically plausible
Leaky integrate-and-fire	1	Very low	No
Izhikevich	2	Very low	No
SRM	1	Low	No
Hodgkin-Huxley	4	Very high	Yes
FitzHugh-Nagumo	2	Medium	No
Wilson	2	Medium	No
Moris-Lecar	3	High	Yes

It is important to note that all widely used models discussed in this paper are deterministic, and as such might have limited applicability for solving more complex problems in the future. Spiking processes in biological neurons could also be viewed on as stochastic by nature. For example, spike time could also depend, besides on the deterministic input signals, on gene and protein expression Katsumata et al. (2008), on physical connection properties Huguenard (2000), on probabilities of spikes being received at the synapses, etc. These facts motivated for new ways to enhance the current SNN models with probabilistic parameters, yielding a recent

development of a probabilistic spiking neuron model (pSNM) N. Kasabov (2010), that may allow for a broader range of applications and understanding in the future, but exceeds the scope of this paper.

3.5 Information processing in SNNs

The real world data and signals are analogue, continuous. Efficient and successful encoding of such inputs into discretely timed spike trains is the crucial and initial step of information processing in SNNs Sengupta and Kasabov (2017). For this cause, we turn to prevailing neural coding theories in neuroscience.

Almost a century ago, Adrian and Zoltermann demonstrated that the firing rate of the stretch receptor in a muscle spindle of a frog increased with the strength of the stimulus Adrian (1926), leading to the conclusion that firing rate is the primary "currency" of information exchange. This coding scheme is known as rate coding. Since this seminal study, there have been many other studies that have challenged this simplistic view of neural coding Bohte (2004); Gautrais and Thorpe (1998); Lestienne (2001). In particular, S. J. Thorpe (1990) argued that since neural processing is so fast (for example visual processing is completed under 100 ms), it must rely on temporal coordination of spikes rather than the firing rate, which requires more time.

Still, one of the fundamental questions in neuroscience is how neurons encode and decode information through spikes Brette (2015)? Is it through rate coding or spike coding? This remains a matter of intense debate and below we briefly review some of the popular theories.

3.5.1 Rate coding

The rate code model is one of the most commonly used models to describe information processing and can use different definitions for the firing rate. The simplest

definition of the firing rate is the temporal average, i.e., the number of spikes divided by the corresponding time interval duration. One can also define the firing rate in the context of spatial average, where spikes from a population of neurons within a certain time-interval are averaged. For the third definition of firing rate, one can consider spikes as random events and the firing rate is given as the average over several trials, also known as spike density, of the same stimulus for a single neuron. Although it has been argued that the rate coding scheme based on spike density is biologically not very accurate Gautrais and Thorpe (1998); S. J. Thorpe (1990), it is one of the most widely used/considered coding schemes.

3.5.2 Temporal spike coding

As opposed to rate coding, temporal coding encodes the information by employing the exact timing of individual spikes. Such approach to information processing in SNNs is biologically supported by evidence based on observation of different types of biological neurons Bohte (2004), with first successful application in SNN-based supervised learning demonstrated in Mohemmed, Schliebs, Matsuda and Kasabov (2011). In the following paragraphs we discuss some of the commonly used coding schemes based on temporal spike coding.

Time-to-first-spike

In this neural coding scheme, information is encoded in the time between the beginning of the stimulus and the appearance of the first spike (see Figure 3.4). Such a scheme is ideal for fast processing of information as it requires only a few milliseconds after the appearance of a stimulus to convey a decision on a stimulus.

Rank order coding

In the rank order (RO) coding scheme S. Thorpe and Gautrais (1998), a population of neurons is considered and the information is encoded in the order in which the

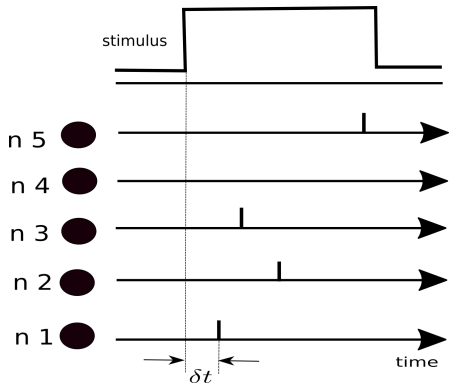


Figure 3.4: Time-to-first-spike: neuron $n1$ is the first to spike at δt after the stimulus onset.

neurons spike, as shown in Figure 3.5a. This scheme assumes that each neuron in the population fires only once after the presentation of a stimulus.

Latency coding

The latency coding scheme depends on the relative timing of spikes (see Figure 3.5b), which also has implications on whether a synapse is potentiated or depressed. For instance, if the relative timing of spikes between the pre-synaptic and post-synaptic neuron is less than 20 ms, then the long-term potentiation (LTP) occurs. Otherwise, long term depression (LTD) occurs.

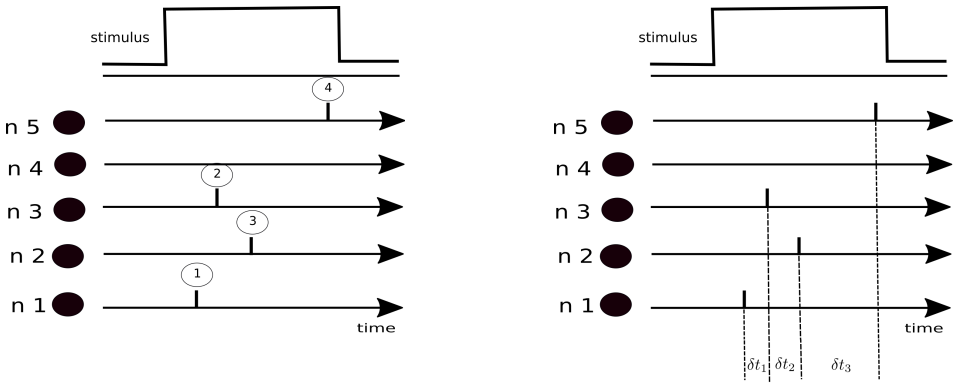
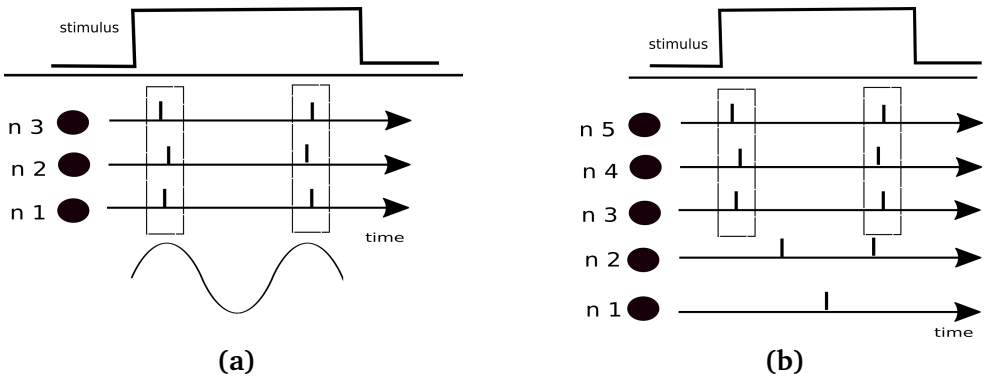


Figure 3.5: Rank order coding (a): information is encoded in the order in which neurons spike. In this example the order is $n1$-$n3$-$n2$-$n5$. Latency coding (b): information is encoded in the spike timing δt_1, δt_2, δt_3 (neurons $n3$, $n2$ and $n5$) relative to $n1$ which spikes first.

Phase coding

In this coding scheme (see Figure 3.6a) it is assumed that neurons spike at different phases with respect to some referent oscillation and thus the phase of the pulse concerning the referent oscillation codes the information. Experimental studies in rats have shown that the information about the spatial location is encoded in the phase of a spike with respect to the hippocampal oscillation. Such oscillations, that are due to a population of neurons, are quite common in several areas of the brain Buzsaki (2006).



(a) **(b)**

Figure 3.6: Phase coding (a): The internal reference oscillation is depicted as a sinusoidal signal and the neurons $n1$, $n2$ and $n3$ spike at the same phase relative to this oscillation. Synchrony coding (b): neurons $n3$, $n4$ and $n5$ spike almost at the same time, as opposed to neurons $n1$ and $n2$ that are not synchronized in their spikes.

Synchrony coding

In synchrony coding model, neurons that spike synchronously are believed to encode the information. Based on several experimental studies, it has been shown that neurons tend to fire synchronously (see Figure 3.6b) when they represent the different bits of information on the same object Buzsaki (2006). For example, it was shown that when the neurons in the visual cortex are activated with a single contour, they tend to synchronise their discharges, but not when the contour is moving in different directions Buzsaki (2006).

3.5.3 Encoding method selection criteria

Similar to the brain, where different spike encoding methods are applied in different sensory areas, such as retina, cochlea, olfactory, gustatory, etc., different encoding methods can be selected in the NeuCube architecture (see subsection 3.7.2) for different types of continuous value input time-series data using different optimisation criteria. Such criteria can be:

a) Minimizing the loss of information after decoding (recovering) the signal back from spike representation to continuous value. In Petro, Kasabov and Kiss (2019) and N. K. Kasabov (2018b) four types of encoding algorithms have been studied: Ben's Spiker algorithm (BSA); threshold-based encoding; step-forward (SF) encoding; moving-window (MW) encoding. BSA can follow smoothly changing signals if filter coefficients are scaled appropriately. SW encoding was most effective for all types of signals as it proved to be robust and easy to optimize. Signal-to-noise ratio (SNR) can be recommended as the error metric for parameter optimization. Different encoding algorithms can be suited to different EEG and fMRI data for example, the most commonly used one being SF N. K. Kasabov (2018b). Free software for finding the optimal encoding of time series for SNN is available[1].

b) Maximazing the classification/prediction accuracy at the output classification/prediction module of a NeuCube model Sengupta and Kasabov (2017). Paper Sengupta and Kasabov (2017) also evaluates how much compression of the input data can be achieved through spike encoding without loss of classification accuracy at the output of an SNN system. It was demonstrated on several benchmark data that input information passed to a NeuCube classifier can be compressed by orders of magnitude without compromising the outcome.

[1] https://github.com/KEDRI-AUT/snn-encoder-tools

3.6 Learning in SNNs

The learning methods for SNNs fall into two categories: 1) Rate-based learning and 2) Spike-based learning. This naturally arises from the previous section (section 3.5), where we have divided information processing approaches in SNNs into rate-based and spike-based. The following subsections (3.6.1 and 3.6.2) provide descriptions of the two learning categories.

3.6.1 Rate-based learning

Training deep-SNNs directly using backpropagation is not possible as gradients cannot be computed for spike inputs. Thus, an indirect approach of training an ANN with backpropagation and then converting it into an equivalent SNN by relating the activations of ANN units to firing rates of spiking neurons is used. The relation between the transfer function (i.e., the input-to-output relationship) of a spiking neuron and the activation of the rectified linear unit (ReLU) of ANNs, have been thoroughly described in Diehl and Cook (2015); Rueckauer, Lungu, Hu and Pfeiffer (2016). Given a network of L layers, with weights W_l, $l \in 1, \ldots, L$ connecting layer $l-1$ to l with b_i^l being bias for neuron i in layer l, and assuming that the number of units in each layer is Nl, the activation in ANN is given by:

$$a_i^l := \max\left(0, \sum_{j=1}^{N_l-1} W_{ij}^l a_j^{l-1} + b_i^l\right) \tag{3.8}$$

with initial condition $a^0 = x$ as input, which is assumed to be normalized so that $x_i \in [0,1]$. In case of a SNN, the membrane potential $u_i^l(t)$, is driven by the input current $z_i^l(t)$:

$$z_i^l := \tau\left(\sum_{j=1}^{N_l-1} W_{ij}^l \Theta_j^{l-1} + b_i^l\right) \tag{3.9}$$

where $\Theta_j^{l-1} = \Theta(u_i^l(t-1) + z_i^l(t) - \theta)$ is a step function. The spiking neuron integrates the inputs $z_i^l(t)$ until the membrane potential $u_i^l(t)$ exceeds the threshold θ.

Given that an input pattern is presented with time step dt, the maximum firing rate is constrained by $1/dt$ and the firing rate is given by $r_i^l(T) = \sum_{t=1}^{T} \Theta_{t,i}^l / T$ which is simply the number of spikes generated divided by the total time the input is presented. The ANN-to-SNN conversion proceeds such that the firing rates r_i^l correlate with activations of ReLU units, a_i^l, such that $r_i^l(T) \rightarrow a_i^l / dt$ Cao, Chen and Khosla (2015); Diehl et al. (2015). In the following paragraph, we briefly review some of the work that has employed ANN-to-SNN conversion.

In a study by P. Merolla et al. (2011), a two-layer (484 visible units and 256 hidden units) restricted Boltzmann machine (RBM) was trained on handwritten digits from MNIST database and the weights were learned using contrastive divergence algorithm, achieving an accuracy of 94% when tested on out-of-sample data. After learning the weights, the 256 hidden units were converted to integrate and fire neurons and two axons (for excitatory and inhibitory synapse) represented each of the 484 visible unit. Following a thresholding procedure, continuous weight matrices were converted to binary matrices and spiking RBM neurons, achieving an accuracy of 89%, which was less than what was achieved by RBM before conversion to spiking version. Using Siegert neurons as computational units, O'Connor et al. O'Connor, Neil, Liu, Delbruck and Pfeiffer (2013) trained each RBM in a time-stepped mode, using CD algorithm for training, with modification to encourage sparse and receptive fields in the hidden layer. For the conversion of RBMs to equivalent spiking network with LIF neurons, the firing rates of the Siegert neurons in RBMs are normalised and converted to activation probabilities. Hunsberger and Eliasmith Hunsberger and Eliasmith (2015) proposed to transfer a static convolutional neural network (CNN), modified from Krizhevsky, Sutskever and Hinton (2012) to spiking neurons. They trained the static CNN by using a smoothed version of LIF rate equation, so that the gradient during backpropagation remains bounded. Furthermore, in order to simulate variability in filtered spike trains the static CNNs were trained with noise added to the output of each neuron for each training example. The static CNN was

then converted to an SNN by replacing back the soft-LIF model with the normal LIF spiking model and removing the noise added during training. Esser et al. Esser, Appuswamy, Merolla, Arthur and Modha (2015) used a training network sharing the same network topology as the deployment network, TrueNorth neurosynaptic chip (P. A. Merolla et al. (2014)). The training network takes data (input and output of neurons, synaptic connections) in a continuous format, with the constraint that the values are within the range $[0, 1]$, which can be obtained by re-scaling the pixels of input images. These continuous values can then be interpreted as a probability of a spike occurring or not when mapping to the corresponding hardware. The network was trained using backpropagation methodology and the probability whether a neuron will spike or not was derived using a complementary cumulative distribution of a Gaussian function. The performance of SNNs obtained after conversion of ANNs on benchmark tasks is still lower than what is achieved by state-of-the-art ANNs and to overcome this, Diehl et al. Diehl et al. (2015) used ReLUs and mapped the weights from the ReLU network to network with integrate-and-fire units. Furthermore, bias was fixed to zero throughout the training with backpropagation. Finally, the key step of weight normalisation was employed, which helped in reducing the errors due to replacing ReLUs with IF neurons.

The major drawback of rate-based learning, as an indirect learning approach in SNNs is the fact that, since traditional rate-coding is used in ANNs, processing times are longer and many spikes are needed to encode the input. Finally, we summarise some of the main challenges in rate-based learning:

1. *Negative values* Negative values can arise when using ANNs like CNNs for several reasons:

    - The sigmoid function used in CNNs can result in negative outputs.
    - Since weights and biases can both be negative, the output which is given as the sum of weights and biases can be negative.
    - Preprocessing step of input image or pattern can produce negative values.

Although neurons with negative values can be represented as inhibitory neurons in SNNs, this would require a doubling of the neurons. An undesirable consequence of this would be a direct increase in hardware resources and power consumption. Also, adding inhibitory neurons to SNNs can lead to complicated interconnections.

2. *Bias* Representing biases in spiking networks is not a straightforward task, and the biases in each layer can be positive or negative.

3. *Max-pooling* Max-pooling in SNNs requires more neurons as we need a two-layer neural network with lateral inhibition to implement spatial max-pooling analogous to what is done in convolutional neural networks.

3.6.2 Spike-based learning

Spike-based learning is based on spike-timing dependent plasticity (STDP), which is believed to be the primary form of synaptic change in neurons Gerstner, Kempter, van Hemmen and Wagner (1996) and involved in learning and memory formation in mammalian visual cortices. Bi and Poo provided the first evidence based on biological observations, on the importance of relative timing of the pre- and post-synaptic spikes in modulating the synaptic efficacy Bi and Poo (1998). Interestingly, one of the first spike-timing-dependent learning algorithms was described by Gerstner already in 1993, where it was shown using SRM neurons that coding by spatio-temporal spike patterns was beneficial over mean firing rate Gerstner, Ritz and Van Hemmen (1993). Over the last few decades, evidence of STDP as a learning mechanism has been established in both in vivo and in vitro studies Cassenaer and Laurent (2007); Jacob, Brasier, Erchova, Feldman and Shulz (2007); Mu and Poo (2006) and STDP has been widely used as learning algorithm in the field of AI. In this section we briefly review the concept of STDP and some works using STDP learning rule for training SNNs.

According to the STDP learning rule S. Song, Miller and Abbott (2000), the strength of the synaptic weight is proportional to the degree of correlation between the spikes in the pre-synaptic and post-synaptic neuron. As shown in the Figure 3.7, if a spike occurs in the post-synaptic neuron shortly (under 20 ms) after the occurrence of a spike in the pre-synaptic neuron, the synaptic connection is strengthened, and this is known as long-term potentiation (LTP). On the contrary, if a post-synaptic neuron fires before the pre-synaptic neuron, long-term depression (LTD) occurs which decreases the synaptic weight. The STDP rule describing the change in weights Δw is generally given as:

$$\Delta w = \begin{cases} Ae^{\frac{|t_{\text{pre}} - t_{\text{post}}|}{\tau}}, t_{pre} - t_{\text{post}} < 0, A > 0 \\ Be^{\frac{|t_{\text{pre}} - t_{\text{post}}|}{\tau}}, t_{\text{pre}} - t_{\text{post}} > 0, B < 0 \end{cases} \tag{3.10}$$

where A and B are the learning rates for LTP and LTD, with τ being the time constant. Generally, we can say that the strength of a synapse is increased at the moment of post-synaptic firing by an amount that depends on the value of the trace left by the pre-synaptic spike. Similarly, the weight is depressed at the moment of pre-synaptic spikes by an amount proportional to the trace left by previous post-synaptic spikes.

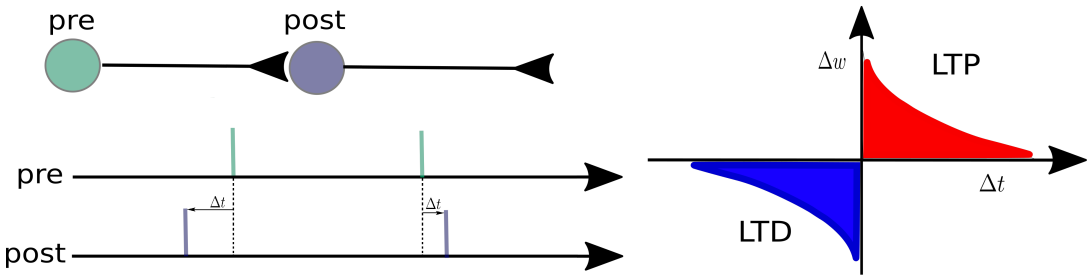


Figure 3.7: The concept of STDP: the function on the right shows the change of the synaptic connection weight Δw as a function of the time difference Δt between a pre- and a post-synaptic spike arrival time. Positive Δt (pre-synaptic spike before the post-synaptic) leads to Long-Term Potentiation (LTP) of the synapse, with negative Δt (post-synaptic spike before the pre-synaptic) leads to Long-Term Depression (LTD) of the same synapse. Spike timings within the two pre- and post-synaptic spike pairings are shown on the left: first pairing results with a negative Δt value and the second pairing with a positive Δt value.

It has been demonstrated that by selecting inputs which have low time jitter, STDP increases the post-synaptic spike time precision Gerstner and Kistler (2002). Several strategies for training SNNs based on STDP learning rule have been proposed. Kheradpisheh et al. Kheradpisheh, Ganjtabesh, Thorpe and Masquelier (2018) proposed an STDP-based learning approach for SNNs, whose architecture consisted of three convolutional layers comprising nonleaky integrate-and-fire neurons equipped with STDP and three pooling layers. The first layer uses a difference of Gaussian (DoG) filter which encodes the contrast strengths in the input image to latencies, leading to earlier firing of neurons for higher contrasts and vice-versa, thus implementing the rank-order coding scheme (see subsection 3.5.2). The neurons in convolutional layers integrate these spikes and fire if a threshold is reached, whereas the pooling layers provide translational invariance using nonlinear max pooling. Their results showed that SNNs achieved accuracies of 99.1% on Caltech face/motorbike task and 98.4% on the MNIST dataset. On the ETH-80 datasets, which consists of images of various objects taken from different viewpoints, an accuracy of 82.8% was achieved. Diehl and Cook Diehl et al. (2015) presented an SNN with two-layers for digit recognition. In the first layer, the intensity of each pixel of the input image is converted to a Poisson-spike whose firing rate is proportional to the intensity. The second layer comprised excitatory and as many inhibitory neurons, with excitatory neurons being connected to inhibitory neurons in a one-to-one fashion, whereas each inhibitory neuron is connected to all the excitatory neurons except the one it received an input from. This arrangement provides lateral inhibition with competition among excitatory neurons. The changes in weights are given by different STDP-based learning rules that make use of exponential weight dependence. The demonstrated an accuracy of 95%, achieved on the MNIST benchmark. Tavanaei et al. Tavanaei and Maida (2017) trained a SNN consisting of Izhikevich neurons (see section 3.4.2) for isolated spoken digit reconstruction. The training is performed using the STDP learning rule which is a

mixture of Hebbian and anti-Hebbian STDP in a supervised fashion, using a teacher signal at the output that determines the type of STDP to be used based on whether the desired output neurons spike (Hebbian) or not (anti-Hebbian). When tested on the Aurora dataset Hirsch and Pearce (2000), the overall classification accuracy was 90.8% without noise and 70.2% under 10dB noise.

3.7 NeuCube

3.7.1 From evolving connectionist systems to dynamic evolving SNNs

Evolving connectionist systems (ECOS) are generally modular systems that evolve both their structure and functionality from incoming information/data, in a way that is continuous, self-organised, online, adaptive and interactive N. Kasabov (1998); N. K. Kasabov (2007). Translating the principles of ECOS to SNNs, neurons are created (evolved) incrementally clustering the input data in either supervised or unsupervised way. This paradigm results in what we now call evolving spiking neural networks (eSNN) N. K. Kasabov (2007); Wysoski, Benuskova and Kasabov (2010) that can learn patterns within the data by creating (evolving) and merging (connecting) spiking neurons. The dynamic eSNN (deSNN), introduced in N. Kasabov, Dhoble, Nuntalid and Indiveri (2013), combines rank-order (see section 3.5.2) and temporal (e.g. STDP, see section 3.6.2) learning rules. All ECOS-type systems, from simple ones to eSNN and deSNN, are generally guided by the same main principles (i.e. they have evolving structures; they learn and partition the problem space locally, allowing for a faster adaptation; they learn in a continuous, on-line, incremental way N. Kasabov, Dhoble et al. (2013)). It has been argued that SNNs are in general the most convenient way towards the creation of an integrative computational framework for processing various spatio- and spectro-temporal brain data (STBD) N. Kasabov (2014), such as EEG, fMRI, DTI, MEG, and NIRS. This is mainly because SNNs in their implementation follow the same computational principle

that generates STBD: spiking information processing, as described in the previous sections of the paper. In the following subsection, we describe the latest version of an implemented SNN architecture, called NeuCube, first introduced in N. Kasabov (2014), with its specific input encoding, STDP learning and deSNN-based output representation. NeuCube is mainly designed for the creation of models that map, learn and help in the understanding of STBD. However, we will discuss some current problems in the area of brain-computer interface (BCI) and affective computing that could be addressed through an SNN-based architecture like NeuCube.

3.7.2 SNN implementation – NeuCube framework

General principles of the NeuCube architecture were first presented in N. Kasabov (2012), followed by a detailed description of the entire architecture in N. Kasabov (2014). In this paper we focus on the latest version of the SNN-based part of the NeuCube architecture, depicted in Figure 3.8. It consists of the following modules:

- Input module: implements input data encoding
- Representation module: 3D SNN reservoir module, i.e. the SNN cube (SNNc) module
- Output module: implements the output function, i.e. classification

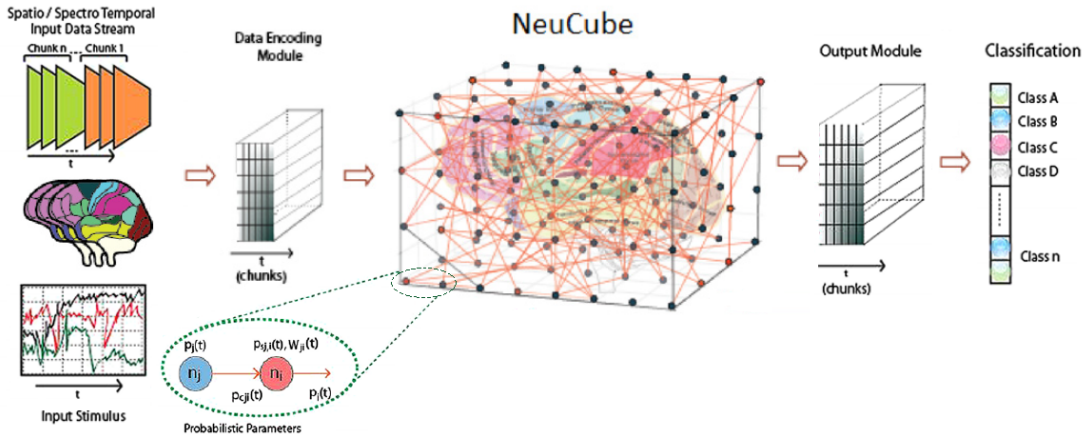


Figure 3.8: A schematic representation of the SNN-based NeuCube architecture, consisting of: input data encoding module; 3D SNNc module; output function module (e.g. for classification or prediction). The gene regulatory networks (GRN) module is optional and is left out for the purposes of this paper. Adapted from N. Kasabov (2014).

The gene regulatory network (GRN) module, parameter optimisation module, and visualisation and knowledge extraction module are all optional and are left out for the purposes of this discussion. The MATLAB-based implementation of the architecture is shown in Figure 3.9.

Input module: encoding

The goal of spike encoding (background described in section 3.5) is the transformation of continuous input timeseries into sparse spike trains of exciting (+1) and inhibiting (−1) spikes. The thresholding representation (TR) algorithm is the most commonly used spike encoding algorithm. It is also called the Address Event Representation (AER) method (as in Lichtsteiner and Delbruck (2005)). The method is based on thresholding the rate of change (x' in equation (3.14)) of the same input variable over time and is suitable when the input data is a stream (which is mostly the case).

The NeuCube implementation of the TR spike encoding algorithm is based on the variable threshold value that is calculated for each of the input data channels. The point of this variable threshold array is for it to be suitable concerning the specific signal dynamics, that could vary between different applications or even different input channels of the same application. For each of the input channels, the variable threshold is calculated based on one scalar input parameter (α_{TR}) in the following way:

$$VT(k) = \frac{1}{N}\sum_{i=1}^{N}(\mu + \sigma \cdot \alpha_{TR}) \tag{3.11}$$

$$\mu = \frac{1}{T}\sum_{j=1}^{L-1} x' \tag{3.12}$$

$$\sigma = \sqrt{\frac{\sum_{j=1}^{T-1}(x'-\mu)^2}{T-2}} \tag{3.13}$$

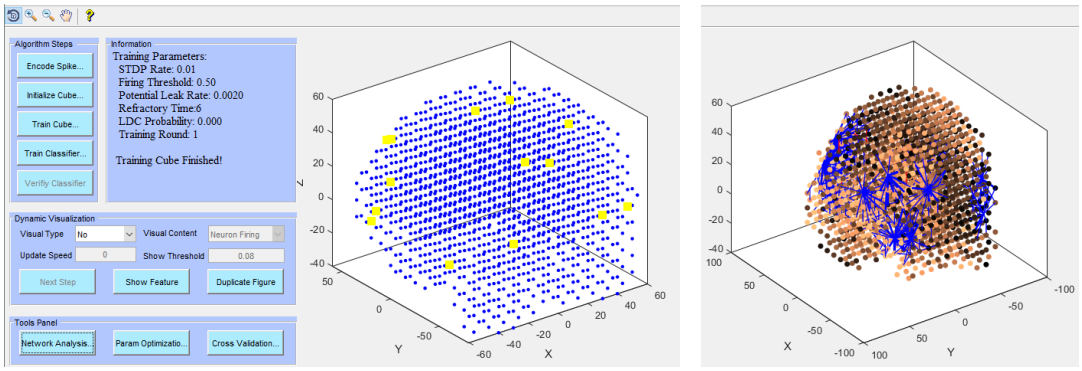
$$x' = \left|X(j+1,k,i) - X(j,k,i)\right| \tag{3.14}$$

where N is the number of samples, T is the signal length (number of time points per data sample), and k goes from 1 to the number of channels N_{input}. X is a $T \times N_{\text{input}} \times N$ data matrix, α_{TR} is the spike threshold parameter, and VT is the resulting variable threshold array. Equation (5.1) represents the threshold value for the k-th input channel.

Once we have determined the threshold values for each of the channels (according to equation (5.1)), every signal in the dataset is transformed in the following way:

- Exciting spike train is a sparse signal of the same length as the input signal, with a value of 1 at each time step where the positive signal difference (rate of change) exceeds the variable threshold.
- Inhibiting spike train is a sparse signal of the same length as the input signal, with a value of 1 at each time step where the negative signal difference (rate of change) exceeds the variable threshold.
- The two spike trains are subtracted, to get a spike train of values 1, 0 and -1. These sparse representations are then used as inputs to the spatially located neurons from the SNNc module (yellow neurons in Figure 3.9a).

Different spike encoding algorithms have different characteristics when representing the input data. Therefore, besides the described TR spike encoding algorithm, one can choose to use the BSA Nuntalid, Dhoble and Kasabov (2011), MW algorithm, or SF algorithm, as these are all implemented in the current version of the NeuCube N. Kasabov, Scott et al. (2016), but detailed explanations are here omitted. Our general recommendation when choosing a proper spike encoding algorithm would be to figure out what information the spike trains shall carry for the original signals. After that, the underlying patterns in the resulting spike trains will have higher interpretability and will hopefully yield a more successful SNNc representation. On the other hand, poor/inadequate spike encoding algorithm choice will essentially



(a) (b)

Figure 3.9: Current version of the MATLAB-based software implementation of the NeuCube architecture: an exemplary classification task with 3-class EEG data. (a) shows the brain/cube (SNNc) neuron coordinates, with the information regarding the unsupervised STDP training phase. (b) shows the connectivity of the SNNc after training (positive connections are represented in blue and negative in red; a brighter neuron has more connections)

mean loss of useful information and can introduce information noise. Hence, spike encoding, being the first link in the processing chain, is also a highly important one. Algorithm 3.2 (appendix) sums up equations 5.1-3.13, with the rest of the spike encoding procedure in a compact algorithmic form.

3D SNNc initialization

The 3D SNN reservoir module, also called the SNN cube (SNNc), is essentially a group of spatially located spiking neurons, usually with known coordinates of the input neurons. The 3D structure of the neurons and their connections require a visualisation that goes beyond a traditional 2D connectivity/weight matrix. Figure 3.9 shows a 3D SNNc of $N_{\text{cube}} = 1471$ brain-mapped spiking neurons, whose coordinates are based on the Talairach Atlas, a human brain template Talairach and Tournoux (1988). SNNc structure (spiking neuron coordinates) can be defined automatically (by specifying how many equally-spaced neurons shall be created for x, y, and z coordinates; resulting in a cuboid shape), or by loading the coordinates from a file. In our case (coordinates based on the Talairach Atlas), the resulting shape was brain-like, but generally, the resulting shape of the SNNc can be arbitrary.

In principle, either Talairach or MNI Evans et al. (1993) brain templates can be used to initialize a brain-like SNNc structure before variable mapping and learning. In some cases, the first can be suitable for cortical data, such as EEG, while the second could be better used for whole brain data such as fMRI. Both methods are used and compared on fMRI data in N. K. Kasabov, Doborjeh and Doborjeh (2016) and N. Kasabov, Zhou, Doborjeh, Doborjeh and Yang (2016).

The spike trains, obtained after encoding data using the TR algorithm (section 3.7.2), are entered into the SNNc via corresponding brain-mapped input neurons. The number of input neurons N_{input} is usually defined by the number of channels arising from the loaded dataset. Coordinates of the input neurons (input mapping locations) are defined either manually, automatically (by graph matching), or from a file. These should correspond to the locations of the origins where the data channels were collected (if such locations exist). In Figure 3.9a, the yellow input neuron coordinates naturally correspond to specific EEG electrode locations. These input neuron coordinates need to be a subset of the SNNc coordinates (usually $N_{\text{input}} \ll N_{\text{cube}}$).

L2 norm is calculated to get distances between pairs of neurons resulting in a $N_{\text{cube}} \times N_{cube}$ matrix of distances L_{dist}. The connections between the neurons in the SNNc are initialised using the small-world connectivity (SWC) approach, where a radius is defined as a parameter for connecting neurons within this radius (SWR). This results in an SNNc of sparsely connected neurons. Another parameter, LDC (long distance connectivity), can be used to initialise connections beyond the SWR. Generally, the LDC probability parameter would represent the probability of establishing a connection between two neurons with $L_{\text{dist}(ij)} > SWR$. Initially, all connections $C_{(ij)}$ between all neurons in the cube are set to 1. The connection flag between two neurons is set to zero (disconnected) if $L_{\text{dist}(ij)} > SWR$. A connection between a neuron i and a neuron j means that i is a pre-synaptic neuron in that connection, and j is post-synaptic. In a situation where a connection is 'bidirectional',

we make a random choice leaving only one of the two. After connection initialization, 80% of the weights (matrix W) are expected to be positive and 20% are expected to be negative:

$$W_{(ij)} = \text{sgn}\Big(\text{rand}(1) - 0.2\Big) \cdot \text{rand}(1) \cdot \frac{1}{L_{\text{dist}(ij)}} \tag{3.15}$$

rand(1) generates pseudorandom values drawn from the standard uniform distribution on the open interval $(0, 1)$. It is also important to note that all input neurons can only be pre-synaptic neurons. Algorithm 3.3 (appendix) sums up the SNNc initialization procedure described in this subsection.

3D SNNc STDP-based training

After the input data encoding (section 3.7.2), and connection and weight initialization (section 3.7.2), the SNNc is trained in an unsupervised manner using the STDP learning rule (introduced in section 3.6.2), based on the training data. This step was described in a compact algorithmic form in A. Abbott, Sengupta and Kasabov (2016). Here we provide a more detailed, precise description of the NeuCube's implementation of STDP learning. The algorithm relies on the following parameters:

- D (potential leak rate): the rate of passive potential degradation through inactivity
- R (refractory time): determining a period of resting between spikes
- η (STDP rate): learning rate, used for weight updating
- β (firing threshold): potential threshold for generating a spike
- N_{iter}: number of training iterations (passes through the training data)

A parameter that determines the LDC probability (explained in section 3.7.2) is here omitted, as it is not crucial for the understanding of NeuCube's STDP implementation.

At each timestamp (considering there are T timestamps for each datasample), N_{input} spike states are fed to the corresponding input neurons of the SNNc, and potential propagations are calculated. Let us consider the (i, j) neuron pair in which i is the pre-synaptic and j is the post-synaptic neuron. If a pre-synaptic neuron i fires, and neuron j is not in refractory time:

$$P_j(t) = P_j(t-1) + W_{(i,j)} \tag{3.16}$$

If at any time t, any neuron k exceeds the firing threshold potential β, it fires and its potential $P_k(t)$ is reset to 0. Its refractory counter R_k is set to R (refractory time). If at any time t, any neuron k did not exceed the firing threshold potential β, its potential is reduced: $P_k(t) = P_k(t-1) - D$, and refractory counter updates: $R_k = R_k - 1$.

Following the general STDP rule (section 3.6.2), if a neuron i as pre-synaptic fires at time t and post-synaptic j has fired last at t_j^f, the connection weight is updated:

$$W_{(i,j)} = W_{(i,j)} - \eta/(t - t_j^f + 1) \tag{3.17}$$

On the contrary, if a post-synaptic neuron j fires at time t and a pre-synaptic neuron i has fired last at t_i^f:

$$W_{(i,j)} = W_{(i,j)} + \eta/(t - t_i^f + 1) \tag{3.18}$$

At the beginning of each new training iteration n_{iter} (of N_{iter}) the learning rate η is adapted to $\frac{\eta}{\sqrt{n_{\text{iter}}}}$. It is important to note that all times considered in the learning algorithm are discrete (meaning t, t_j^f, t_i^f, R_k and R are all integers). Algorithm 3.4 (appendix) sums up NeuCube's unsupervised STDP-based learning procedure described in this subsection.

deSNN representation and supervised learning

After the unsupervised training, described in the previous section, the SNNc connectivity (Figure 3.9b) can be analysed and observed, hopefully yielding better interpretability of the data when compared to traditional data processing and learning methods. This functionality additionally allows to identify differences in the activity of various brain regions in case of STBD or to identify input channel interactions in cases of other types of input data. Furthermore, we end up with a trained network that will, once we feed it with a specific input spike sequence, produce an array of spatio-temporal neuron firing events that are a result of both the input spike sequence and specific neuron connections that have emerged from the training data. This array of spatio-temporal events forms a pattern that can be encoded into a representation that is generally suitable as input to any traditional supervised learning algorithm (i.e. SVM, ANN, kNN).

The same data that was used for unsupervised STDP-based training is propagated again through the trained SNNc and a supervised model is trained to classify the spatio-temporal spiking pattern left in the SNNc into predefined classes. The output module from Figure 3.8 represents the step where the SNNc spatio-temporal spiking pattern is transformed into a representation (or output neurons). The deSNN representation learning used in the NeuCube version presented in this paper is based on the RO learning rule (principles explained in subsection 3.5.2).

The algorithm parameters are:

- α (*mod*): important for weight update on first spike occurrence
- d (*drift*): used for weight update on subsequent spikes

Propagating a single training sample through the trained SNNc network results with a sparse $T \times N_{\text{total}}$ neuron activation matrix representing the trajectories of all neurons spiking states through the time T (which is, as already described, a dataset-specific datasample time length). We form a $1 \times N_{\text{total}}$ array of output weights

W_o, that will practically represent the output representation. Initially, all weights are zeros. These weights are updated by going through the $T \times N_{\text{total}}$ activation matrix for all neurons, at each timestamp. If a neuron i fires for the first time we set its weight to:

$$W_{o(i)} = \alpha^{\text{order}} \tag{3.19}$$

with order being the firing counter. We start at order $= 0$ and each time a neuron fires for the first time, the counter is incremented order $=$ order $+1$. Otherwise, if a neuron fires again, the weight is updated by the rule:

$$W_{o(i)} = W_{o(i)} + d. \tag{3.20}$$

and if it does not fire, the weight is updated by the rule:

$$W_{o(i)} = W_{o(i)} - d. \tag{3.21}$$

To practically interpret the resulting representation array: W_o is a relatively sparse representation where the neurons that fired among the first, and most often afterwards, contribute most to the final position of a data sample in the $1 \times N_{\text{total}}$ deSNN representation space. Therefore, the input neuron activations will contribute most significantly to the final representation for $\alpha < 1$.

The described algorithm provides a representation for each datasample in the dataset, including the validation/test data as well, although the SNNc connectivity is learned from just the training data in an unsupervised manner. With the $1 \times N_{\text{total}}$ representation, we can further address the problem via classical supervised learning, corresponding to the final module in the NeuCube schematic (Figure 3.8). The method used for this part of the algorithm could be chosen arbitrarily, from one of the many available and widely used supervised learning algorithms. The current MATLAB-based NeuCube employs a version of the kNN (k-nearest neighbours) algorithm, but a detailed explanation of the classification module is here avoided

as it exceeds the SNN nature of the paper. Algorithm 3.5 (appendix) sums up NeuCube's deSNN representation algorithm that is described in this section. A compact summary of the entire NeuCube-based supervised learning chain is bellow in algorithm 3.1, where we integrate all previous subsections (3.7.2, 3.7.2, 3.7.2, 3.7.2) and accompanying algorithms (3.2, 3.3, 3.4 and 3.5).

Algorithm 3.1 NeuCube-based supervised learning

Require: $X_{train} \in \mathbb{R}^{T \times N_{input}}, X_{test} \in \mathbb{R}^{T \times N_{input}}, Y_{train} \in classes, Y_{test} \in classes$
$\{hyperparameters := encoding\ pars.,\ SNNc\ struct.\ pars.,\ STDP\ pars.,\ \hat{Y_{test}} := f_{supervised}(X_{train}, Y_{train}, X_{test})\}$
Ensure: $\hat{Y_{test}} \in classes$
1: $N_{train} \leftarrow \#(X_{train})$ {number of samples in the train dataset}
2: $N_{test} \leftarrow \#(X_{test})$ {number of samples in the test dataset}
3: $S_{train} \leftarrow f_{encode}(X_{train}, encoding\ pars.)$ {see algorithm 3.2 and section 3.7.2}
4: $S_{test} \leftarrow f_{encode}(X_{test}, encoding\ pars.)$
5: $W \leftarrow f_{initialize}(SNNc\ struct.\ pars.)$ {see algorithm 3.3 and section 3.7.2}
6: $W \leftarrow f_{STDP}(W, S_{train},\ STDP\ pars.)$ {see algorithm 3.4 and section 3.7.2}
7: $\widetilde{W}_{train} \leftarrow f_{deSNN}(W, S_{train})$ {see algorithm 3.5 and section 3.7.2}
8: $\widetilde{W}_{test} \leftarrow f_{deSNN}(W, S_{test})$
9: $\hat{Y_{test}} \leftarrow f_{supervised}(\widetilde{W}_{train}, Y_{train}, \widetilde{W}_{test})$ {arbitrary supervised learning algorithm like kNN gives}
10: performance evaluation by comparing Y_{test} and $\hat{Y_{test}}$

3.8 Conclusion and future work

In this paper, we describe the development and discuss implementational aspects of spiking neural networks. The goal is to promote the use of SNN (as the third generation of ANN) and more specifically – the NeuCube architecture, as an off-the-mainstream approach in neural network processing. Comparative advantages of SNNs are clear, like computational speed, power consumption and biological plausibility. However, technical challenges continue to exist. As there is not yet a robust information theory supporting the design and implementation of SNNs N. Kasabov, Scott et al. (2016), choice of the network structure (including input neuron locations) and other hyperparameters for each specific application is still based on heuristic measures and expert knowledge.

Algorithm 3.2 NeuCube's TR spike encoding: $f_{encode} : \mathbb{R}^{T \times N_{input}} \rightarrow \{-1, 0, 1\}^{T \times N_{input}}$

Require: $X_{in} \in \mathbb{R}^{T \times N_{input}}, \{hyperparameters := \alpha_{TR}\}$
Ensure: $\widetilde{W}_{train} \in \{-1, 0, 1\}^{T \times N_{input}}$

$N \leftarrow \#(X_{in})$ {number of data samples in the dataset}
for $k = 1$ to N_{input} **do**
  $VT_k \leftarrow 0$
  for $i = 1$ to N **do**
    $x \leftarrow$ channel k of the i-th sample in X_{in}
    $x' \leftarrow |\delta x|$
    $\mu \leftarrow mean(x')$
    $\sigma \leftarrow st.dev.(x')$
    $VT_k \leftarrow VT_k + (\mu + \sigma \cdot \alpha_{TR})$
  end for
  $VT_k \leftarrow \frac{VT_k}{N}$
end for
for $k = 1$ to N_{input} **do**
  for $i = 1$ to N **do**
    $x \leftarrow$ channel k of the i-th sample in X_{in}
    $x' \leftarrow \delta x$
    $x_{out} \leftarrow 0^{T \times 1}$
    for $j = 2$ to T **do**
      if $x'_j > VT_k$ **then**
        $x_{out(j)} \leftarrow 1$
      else if $x'_j < -VT_k$ **then**
        $x_{out(j)} \leftarrow -1$
      end if
    end for
    store spike train x_{out} in X_{out} for channel k, sample i
  end for
end for

As argued in previous works N. Kasabov (2014); N. Kasabov, Scott et al. (2016), NeuCube's 3D SNN structure allows for the integrative modelling of various STBD, thus opening new opportunities for a better understanding and interpretation of STBD. The feasibility of using NeuCube in solving STBD-based tasks was already demonstrated in various domains, like EEG-based brain-machine interface for limb movement Taylor et al. (2014) or EEG-based classification of activities of daily living (ADL) in neurorehabilitation J. Hu, Hou, Chen, Kasabov and Scott (2014). The network structure is relatively intuitive when it comes to STBD-based problems (brain-like structure shown in Figure 3.9), but the real challenge is designing the SNN structure for non-STDB-based classification and pattern recognition tasks. Successful

examples of such efforts exist, like multi-variable-based personalised early stroke prediction Othman et al. (2014) or video-based age classification N. Kasabov, Scott et al. (2016), but still warrant for further exploration.

Finally, we summarise some open research problems in SNN and NeuCube applications that we find most promising, some of them already raised in N. K. Kasabov (2018b):

1. Integrating into one SNN model STBD multimodal data, such as EEG Z. G. Doborjeh, Kasabov, Doborjeh and Sumich (2018), audio-visual Paulun, Wendt and Kasabov (2018), fMRI and DTI Sengupta, McNabb, Kasabov and Russell (2018).

2. Integrating into one SNN model heterogeneous data, such as quantum-, molecular-, brain-, physiological-, environmental information N. K. Kasabov (2018b).

3. Defining optimal depth and length in both space and time of the knowledge representation extracted from a trained SNN N. K. Kasabov (2018b).

4. Using SNN for the development of new information theories, such as information compression Sengupta and Kasabov (2017).

5. Human-machine and machine-machine transfer learning based on a common structural template for STBD, such as the Talairach Atlas Talairach and Tournoux (1988) or MNI Evans et al. (1993).

6. Learning long spatio-temporal patterns in the context of associative memory.

7. Developing methods for self-optimisation of learning in SNN: *learning to learn*.

8. Building societies of SNN machines for distributed deep learning and transfer learning.

9. Towards a symbiosis between humans and SNN machines.

Some of the research problems from the list above have been addressed previously (i.e. 1, 2, 3 and 4), but nevertheless, remain open. On the other hand, some will most probably remain unaddressed in the near future (i.e. 7, 8 and 9). Spiking neural networks and the NeuCube framework are currently at the point where a vast expansion of work in points 1-5 is expected. Some potential applications that we find most promising are in the area of sleep stage classification where vast amounts of data-intensive laboratory-based polysomnography (PSG) recordings exist O'reilly, Gosselin, Carrier and Nielsen (2014), thus allowing for the possibility of having a pre-trained SNNc for transfer learning EEG-based BCI applications (relates to the research problem 5). Another highly challenging, but promising application is the area of affective computing where semi-STBD-based datasets are available Koelstra et al. (2012); Soleymani, Lichtenauer, Pun and Pantic (2012). The primary challenge here would be the integration of pure STBD (EEG), which reflects the central nervous system activity with the peripheral measures of autonomic activity like (but not limited to) electrocardiography (ECG) and electrodermal activity (EDA) into a single SNN-based structure (relates to research problems 1 and 2). The question is can these peripheral measures of autonomic nervous system (ANS) activity, be observed as activity within the limbic structures of the brain (i.e. amygdala, hippocampus, thalamus, hypothalamus), and as such be simultaneously integrated with STBD data like EEG and NIRS into a single specifically designed SNN structure? This would also open questions on how to efficiently encode the peripheral physiological data into spikes suitable for simultaneous processing with spike encoded STBD.

The version of NeuCube described in this paper is the so-called Module M1 (official NeuCube modules range from M1-M4 N. Kasabov, Scott et al. (2016)) for research and teaching purposes, and can be found free of charge at `http://www.kedri.aut.ac.nz/neucube`. For commercial use or access to the full set of modules, please contact the corresponding author directly or via this web page. The NeuCube is PCT patent protected.

3.9 Appendix: Algorithms

Algorithm 3.3 NeuCube's weight and connection initialization: $f_{initialize}$

Require: $X_{brain} \in \mathbb{R}^{1\times 3}, X_{input} \subset X_{brain}$
$\{hyperparameters := SWR, p \in [0,1]\}$
Ensure: $C : \{0,1\}^{N_{cube}\times N_{cube}}, W : \mathbb{R}^{N_{cube}\times N_{cube}}$

$N_{cube} \leftarrow \#(X_{brain})$ {number of defined neuron coordinates}
$L_{dist} : \mathbb{R}^{N_{cube}\times N_{cube}} \leftarrow$ distances between all pairs of neurons
$C_{ij} \leftarrow 1, \forall i, j$
$W_{ij} \leftarrow 0, \forall i, j$
for $i = 1$ to N_{cube} **do**
  for $j = 1$ to N_{cube} **do**
    if $L_{dist(ij)} > SWR$ **then**
      $C_{ij} \leftarrow 0$
    else
      $W_{ij} \leftarrow sgn(rand - p) \cdot rand \cdot L^{-1}_{dist(ij)}$ {$rand$ represents a randomly generated real number from the interval $[0,1]$}
      if $C_{ij} = 1 \wedge C_{ji} = 1$ **then**
        if $rand > 0.5$ **then**
          $C_{ij} \leftarrow 0$
          $W_{ij} \leftarrow 0$
        else
          $C_{ji} \leftarrow 0$
          $W_{ji} \leftarrow 0$
        end if
      end if
    end if
  end for
end for

Algorithm 3.4 NeuCube's unsupervised SNNc weight learning: f_{STDP}

Require: $C_{init} \in \{0,1\}^{N_{cube} \times N_{cube}}, W_{init} \in \mathbb{R}^{N_{cube} \times N_{cube}}, S_{in} \in \{-1,0,1\}^{T \times N_{input}}$
$\{hyperparameters := D, R, \eta, \beta, N_{iter}\}$
Ensure: $W_{out} : \mathbb{R}^{N_{cube} \times N_{cube}}$

$N \leftarrow \#(X_{in})$ {number of samples in the (training) dataset}
$\chi \leftarrow [1, 2, ..., N_{cube}]$ {all neuron indices}
$P_k, \forall k \in \chi \leftarrow 0$ {initialize neuron potentials}
$R_k, \forall k \in \chi \leftarrow 0$ {initialize neuron refractory time counters}
find $inputneuron$ indices $\iota \subset \chi$
for $n_{iter} = 1$ to N_{iter} **do**
  $\eta' \leftarrow \frac{\eta}{\sqrt{n_{iter}}}$
  for $i = 1$ to N **do**
    $s \leftarrow$ all N_{input} channel spikes of the i-th sample in S_{in} {inhibiting spikes (-1) are passed to virtual inhibitory neurons that accept only the inhibiting spikes}
    for $t = 1$ to T **do**
      find $firingneuron$ indices $\tau = \{firingneurons \text{ in } \iota\} \cup \{k \in \chi \setminus \iota, P_k > \beta\}$
      for all $j \in \tau$ **do**
        find post synaptic neuron indices γ
        for all $k \in \gamma$ and $R_k = 0$ **do**
          $P_k \leftarrow P_k + w_{jk}$ {postsynaptic neuron potential update}
        end for
      end for
      $P_k \leftarrow 0, \forall k \in \tau$ {firing neurons reset potential}
      $R_k \leftarrow R, \forall k \in \tau$ {firing neurons reset refractory counter}
      $P_k \leftarrow max(0, P_k - D), \forall k \in \chi \setminus \iota \setminus \tau$
      $R_k \leftarrow max(0, R_k - 1), \forall k \in \chi \setminus \iota \setminus \tau$
      for all $j \in \tau$ **do**
        find post synaptic neuron indices γ
        for all $k \in \gamma$ **do**
          $w_{jk} \leftarrow w_{jk} - \eta(t - t_k^f)$
        end for
        find pre synaptic neuron indices γ
        for all $k \in \gamma$ **do**
          $w_{jk} \leftarrow w_{jk} + \eta(t - t_k^f)$
        end for
      end for
    end for
  end for
end for

Algorithm 3.5 NeuCube's deSNN output representation: f_{deSNN}

Require: $W \in \mathbb{R}^{N_{cube} \times N_{cube}}, S_{in} \in \{-1, 0, 1\}^{T \times N_{input}}$
$\{hyperparameters := \alpha, d\}$
Ensure: $\widetilde{W} \in \mathbb{R}^{1 \times N_{cube}}$

$\widetilde{W} \leftarrow 0^{1 \times N_{cube}}$ {initialize representation to 0}
$F \leftarrow 0^{1 \times N_{cube}}$ {initialize the neuron firing flags}
$c \leftarrow 0$ {initialize the neuron firing order counter}
$S_{cube} \leftarrow$ propagate S_{in} through the SNN defined by W {S_{cube} is a sparse matrix of all neuron firings through the data sample length period T}
for $i = 1$ to T **do**
  for $j = 1$ to N_{cube} **do**
    if $S_{cube}(i, j) = 1$ **then**
      if $F(j) = 0$ **then**
        {neuron fires for the first time}
        $\widetilde{W}(j) \leftarrow \alpha^c$
        $F(j) \leftarrow 1$
      else
        $\widetilde{W}(j) \leftarrow \widetilde{W}(j) + d$
      end if
      $c \leftarrow c + 1$
    else
      $\widetilde{W}(j) \leftarrow \widetilde{W}(j) - d$
    end if
  end for
end for

4 FacialSense: Emotional Valence recognition using Brain-inspired Spiking Neural Network

4.1 Prelude to Chapter 4 Manuscript

The need for establishing more meaningful human-machine interactions have led to research on Affective computing. Affective computing systems concentrate on identifying human emotions and allowing machines to generate appropriate reactions or affects. Affective computing systems utilise various input streams such as facial expressions, voice inputs, body language, and physiological data (such as ECG, body temperature, etc.). Of these, it can be argued that facial expressions provide the rich data input for human-human interactions. Therefore, working on affective computing systems that can recognise human facial expressions is critical. This work presents an approach to facial affect recognition using neural networks. The proposed system uses third generation spiked neural networks for the same. Unlike artificial neural networks (ANNs) which came before them, spiked neural networks or SNNs are biologically compatible as the neurons work on spiked signals rather than a continuous one as in ANN. This is similar to how biological neurons operate, making these third-generation spiked neural networks more relevant to human-machine interactions. The dexterity of SNNs in Spatio-temporal patterns recognition is ideally suited for mimicking human brain patterns, making SNNs the most compatible neural network structures for affect recognition. Evolving spiked neural networks or eSNNs are employed in this study. A specific eSNN configuration

termed as NeuCube, which is specifically designed to handle brain data such as EEG and fMRI data effectively is utilised in this research.

The study uses the multimodal affect dataset MAHNOB-HCI which contains physiological, audio, and visual (facial and gaze) data. However, for the purpose of this study, the visual data alone is employed for facial affect recognition in this study. For the process, expression changes are used to affect recognition. Therefore, videos that exhibit no events or changes are excluded from the analysis. All in all, the study had a dataset of 527 sessions recorded in colour at 60fps from 27 subjects. The study then proceeds to describe the various stages involved in the facial emotion recognition pipeline. These are face detection and tracking, identifying and extracting the necessary features, using the extracted features to detect an event or change in affect state, and then training the network in event classification using NeuCube structure and the data available with the video input. Of these, face detection was carried out using the Matlab toolkit Computer Vision (CV). The process involves using point tracking to identify the polygon enclosing the face. Once this polygon is constructed, an existing DLIB library containing 68 facial landmarks is used to identify events or changes in facial affect state. These are followed by feature extraction, where measurable parameters such as distance between eyelids, lips, etc. are measured. Measuring these parameters allows for identifying changes in these parameters throughout the video, thus enabling the identification of specific events or changes in affect state. The final process involved is to train the NeuCube to recognise facial emotions. This process involves multiple steps.

The first step of the process is to encode the facial affect data available into a train of spikes so that it can serve as input to the eSNN that makes up the NeuCube. This is achieved using a gaussian receptive field covering the entire possible data interval. The time-based variations in a signal are coded into spikes using the gaussian field as a reference. The SNN network then has to be initialised by assigning weights to various neurons, thus building an SNN reservoir. In this study, an 11*11*6 array

of neurons was constructed for facial affect recognition. The input neurons are then identified, and the SNNr is then fed input data for training. The unsupervised training allows the NeuCube to identify patterns in the input data. During this stage, the weightage of the neurons was adjusted using a learning rule termed as time-dependent spike plasticity. This is followed by supervised training, where the structure is taught to classify the specific patterns into specific emotions. At this stage, for each input, an output neuron was created and linked to all the existing neuron. The final step is to feed the network with test data to assess whether it can identify the facial affects states involved. Here, the weights obtained for each neuron in the supervised learning stage is used to assess the test data. The study also presents the initial conditions used for default parameters related to the NeuCube. The results of the study are then discussed. It is seen that this approach results in an accuracy of 77%. This approach, using only facial expression data, therefore has accuracy comparable to more complex and detailed approaches that employ physiological data such as ECG, combined with facial affects. This leads the researcher to conclude that physiological data may not be necessary to assess facial affects, given that it is often compromised by artefacts and the inability of neural networks to fully exploit the data available. However, using NeuCubes could address this limitation, and combining facial affects with physiological data in an approach using NeuCubes could result in improved accuracy. The paper then concludes with discussing the assumptions and limitations of the present study and suggests that future studies in the area can incorporate e micro-expression detection, the ability to read non-frontal head poses, and illumination variations.

In summary, this paper starts off by introducing spiked neural networks or third-generation neural network. Their similarity to biological neural networks makes them ideal for applications in affective computing. A 3D spatiotemporal structure made of SNNs, termed as NeuCube is used in this study. This is the first work using SNN based NeuCubes to address the facial emotion recognition problem. The input

is in the form of visual data alone, and no physiological inputs are provided for the NeuCube. Even then, it is seen that the method performs at par with more sophisticated approaches using multimodal data. It is therefore safe to assume that a NeuCube -based approach with multimodal inputs will have accuracies well over the 77% reported by this study.

4.1.1 Contributions and Publications

Contributions

1. *This work is the first of its kind, employing NeuCubes for Video's facial affect recognition.*
2. *Encoding the facial Affect data into a train of spikes so that it can serve as input to the evolving Spiking Neural Networks using the Gaussian Receptive Field covering the entire possible data interval*
3. *Using NeuCubes, facial expression data alone has accuracy comparable to more complex and detailed approaches that employ a combination of physiological data and facial affects.*

Publications

1. Tan, C., Et al. (2020). *FacialSense: Emotional Valence recognition using Brain-inspired Spiking Neural Network*. Manuscript submitted for publication.

4.2 Introduction

Enabling smooth communication between human and computer systems is the central aim of affective computing and it requires computer systems to correctly interpret the emotional behavior and the affective state of the human. Facial expressions are one of the most common ways through which humans communicate their emotions and they constitute about 55 percent of the information communicated during face to face human interaction Mehrabian (1968). Recognition of facial emotion has applications in several fields such as medicine Edwards, Jackson and

Pattison (2002), driver fatigue monitoring, human-computer interaction, sociable robotics Fong, Nourbakhsh and Dautenhahn (2003) and security systems, to name a few.

Representation of emotions in one of the key component of any facial emotion recognition (FER) system and there exist several models of emotions. The Eckman emotion model proposed in Ekman (1999) has six basic human emotions, which include anger, disgust, fear, happiness, sadness and surprise. Such discrete emotions are easier to understand but are limited in describing certain emotions in different languages. In contrast, dimensional models of emotions represent emotion as a point in multidimensional space, which is described by axes that exhibit the largest variance of all possible emotions. Another popular dimensional model proposed in Plutchik (2001) is based on evolutionary principles and has eight basic bipolar emotions. The two-dimensional model of arousal and valence proposed in Russell (1980) is the most widely adopted emotion representation model Gunes, Schuller, Pantic and Cowie (2011). Arousal ranges intensity of emotion from calm to excited, and valence ranges from unpleasant to pleasant.

Emotion Recognition can be achieved by analyzing facial expression, body movements, voice behavior, gestures, and an array of physiological signals, such as heart rate, sweat, pupil diameter, brain signals, to mention a few. For reviews of physiological measurement based emotion recognition, see Shu et al. (2018) and Bota, Wang, Fred and Silva (2020).

The problem of recognizing emotions by utilizing facial expressions from videos and static images have been addressed by several studies Danelakis, Theoharis and Pratikakis (2015); Poria et al. (2015); Yeasin, Bullot and Sharma (2006). Advances in deep learning methodologies have created huge interest in application of such methods in FER Fan, Lu, Li and Liu (2016); Gudi, Tasli, Den Uyl and Maroulis (2015); Ionescu, Popescu and Grozea (2013); Kahou et al. (2013); Tang (2013), most of which are based on supervised learning. For an excellent overview of the

application of deep learning and as well as shallow learning approaches to FER , the reader is directed to S. Li and Deng (2018) and the references there in.

Spiking neural networks (SNNs) represent the third-generation of neural networks, modelling neurons and interactions between them in a biologically more realistic manner compared to second-generation neural networks based on ANNs. SNNs are an ideal choice to handle the emotion recognition task from video data, given their ability to handle spatio-temporal data effectively Dhoble, Nuntalid, Indiveri and Kasabov (2012) (see section 6.5 for details).

SNN have been proven to be successful in Emotion Recognition field, for instance, Diehl et al. presented a neuromorphic device that responds to language input by producing neuron spikes in proportion to the strength of the appropriate positive or negative emotional response Diehl et al. (2016). Mansouri-Benssassi et al. predicts emotional states using SNN based on facial expression with accuracy of 89% and speech data with 70% of accuracy, Mansouri-Benssassi and Ye (2021). Al-Nafjan developed a NeuCube-based SSN tested in small version of DEAP dataset to detect valence emotion level using only 60 EEG samples with 84.62% accuracy Al-Nafjan, Alharthi and Kurdi (2020).

In this work, we propose to build FER system using SNNs. To this end, we use the NeuCube framework N. Kasabov (2014), which is a type of evolving SNN (eSNN) and develop an encoding method to map the continuous facial feature values to spikes based on population coding. We use the data from Mahnob-HCI dataset to test the NeuCube framework for the classification of binary valence in response to video stimuli. Although the authors have previously applied SNN on Emotion Recognition testing in MAHNOB-HCI Tan, Ceballos, Kasabov and Puthanmadam Subramaniyam (2020); Tan, Šarlija and Kasabov (2021), the work here presented using just Facial expression information seems to outstand as a good choice based in our accuracy and feasibility of facial measurements in contrast with physiological signal measurements.

The structure of the paper is organized as follows. In section 6.5 we provide some background on SNN and the NeuCube framework. Section 4.4 details the methodology used in our work and section 6.7 presents the results. In section 6.8 we discuss our results and in section 6.9 direction for future work and conclude the paper.

4.3 Spiking neural networks

Biological neurons communicate with each other via discrete events known as action potentials or spikes, following an all-or-none principle, where a neuron fires an action potential if the stimulus crosses a certain threshold, else it remains silent. Given this style of information processing, the biological neurons (in humans and other animals) still outperform the existing artificial neural networks in terms of both energy and efficiency LeCun, Bengio and Hinton (2015); W. Wang et al. (2018). Parhi et al. present an overview of the brain-inspired computing models starting with the development of the perceptron and followed by convolutional neural networks (CNNs) and recurrent neural networks (RNNs), Parhi and Unnikrishnan (2020). They also briefly review other neural network models such as Hopfield neural networks, Boltzmann machines, spiking neural networks (SNNs) and hyperdimensional computing. Compared to the traditional ANNs, spiking neural networks (SNNs) employ a more biologically plausible models of neurons Taherkhani et al. (2019), thus bridging further the gap between neuroscience and artificial learning algorithms. SNNs have been shown to be computationally more efficient than ANNs both theoretically Maass (1997a); Maass and Markram (2004) and in several real-world applications Bohte, Kok and La Poutre (2002). SNNs have been used in several real-world learning tasks such as unsupervised classification of non-globular clusters Bohte, La Poutré and Kok (2002), image segmentation and edge detection Meftah, Lezoray and Benyettou (2010), and epileptic seizure detection with electroencephalogram (EEG) Ghosh-Dastidar and Adeli (2007). Furthermore, Bohte and

colleagues devised a supervised learning rule for the SNNs and demonstrated its application in the XOR classification problem and several other benchmark datasets Bohte, Kok and La Poutre (2002).

It has been shown that SNNs can be particularly exploited to solve online spatio-temporal pattern recognition in an efficient manner due to their ability to handle information in time and space in an adaptive and self-organized fashion Dhoble et al. (2012). The eSNNs, first proposed in N. K. Kasabov (2007), can handle spatio-temporal data by increasing the number of spiking neurons in time to learn temporal patterns from data Wysoski et al. (2010). In addition to the open evolving structure of eSNNs that facilitate addition of new variables and neuronal connections, eSNN have the advantage of fast learning from large amounts of data and can interact with other systems actively. eSNNs also allow the integration of various learning rules such as supervised learning, unsupervised learning, fuzzy rule insertion and extraction, to mention a few and are self-evaluating in terms of system performance. These aforementioned properties constitute the evolving connectionist systems (ECOS) principles on which the eSNN is based N. K. Kasabov (2018b). An extension of eSNN known as dynamic eSNN (deSNN) was introduced in N. Kasabov, Dhoble et al. (2013) that combines rank-order and temporal learning rules (for example spike-timing dependent plasticity). Another variant of eSNN system known as NeuCube N. Kasabov (2012, 2014); N. Kasabov and Capecci (2015) initially proposed to handle problems of spatio-temporal pattern recognition in brain data such as EEG, functional magnetic resonance imaging (fMRI) to cite a few, has been further developed to handle various other types of spatio-temporal data such as audio-visual data, climate data, seismic data and ecological data N. Kasabov, Scott et al. (2016).

4.3.1 MAHNOB database

The MAHNOB-HCI dataset is a multi-modal database for affect recognition and implicit tagging Soleymani, Lichtenauer, Pun and Pantic (2011). In this database,

27 subjects (16 females and 11 males) aged between 19 and 40 years old were monitored while watching 20 stimulus clips (34.9 to 117 seconds long) extracted from Hollywood movies and video websites, such as www.youtube.com and blip.tv. Data such as EEG, electrocardiography (ECG), face video, audio, gaze, and peripheral physiological signals (respiration amplitude, skin temperature, galvanic skin response (GSR)) were acquired while the participants were watching the clips. The physiological signals were acquired with a sampling rate of 1024 Hz (later downsampled to 256 Hz), while six different views of participant's facial expressions were recorded simultaneously by six video cameras at 60 frames per second (fps). In this work, the video taken by only the colour camera above the screen was used. After watching each stimulus, the participants used a keyboard interface for answering five questions related to emotional label, arousal, valence, dominance and predictability. Participants answered each question using nine numerical keys, selecting nine emotional labels for the first question and nine possible levels for the last question. In this work, only the binary valence scale was used where levels 1 to 5 were considered as low valence (unpleasant) and levels 6 to 9 as high valence (pleasant). The database is available online http://www.ibug.doc.ic.ac.uk/resources/mahnob-hci-tagging-database/.

Since the proposed method uses event detection based on expression changes, the videos with no event detected are rejected for analysis. In total, 55 videos were rejected for this reason. Also, for some subjects, less than 20 sessions of data were available, with the least number of sessions being 14. Since all the facial expressions are spontaneous, frequent expression changes and subject dependent variations are obvious. However, for some sessions in the database, several channels of the data were missing during the original data collection. Overall, 527 sessions from 27 subjects were actually used for this study.

4.4 Methodology

The facial emotion recognition (valence) pipeline starts with face detection and tracking in video (see Fig. 4.6a), followed by face landmark detection, features extraction, event detection, training and event classification using NeuCube and video classification.

4.4.1 Face detection and tracking

The first step for analysing face emotion recognition in video is face detection and tracking in frames. Computer Vision (CV) Matlab Toolbox was used for this task. The output of this step is the corner coordinates for the polygon enclosing the face for each frame in the video.

The face detection carried out in this work included the following steps,

1. The face in the first frame was detected using the *vision.CascadeObjectDetector* object in the CV toolbox. This function uses the Viola-Jones algorithm Viola, Jones et al. (2001) to detect people's faces, noses, eyes, mouth, or upper body. It outputs the region of interest (ROI) for the face as a polygon, enclosing the face. Specifically, the algorithm uses the histogram-of-oriented gradients (HOG), Local Binary Patterns (LBP), Haar-like features and a cascade of classifiers trained using boosting.

2. The corner features in the first frame ROI were detected using the *detectMinEigenFeatures* function in CV toolbox, which uses the minimum eigenvalue algorithm Shi et al. (1994).

3. For tracking of feature points in the remaining frames, we used Kanade-Lucas Tomasi (KLT) algorithm Lucas, Kanade et al. (1981); Shi et al. (1994).

4. Finally, in order to estimate the motion of the face, we used *estimateGeometricTransform* function in the CV toolbox to apply the same transformation to the ROI detected in the previous frame to obtain the ROI in the current frame.

Figure 6.1 shows the output of the face detection step. We found that point tracking in frames to detect face is computationally more efficient than face detection in each frame. Furthermore, point tracking can manage problems that can emerge in face detection such as making gestures with hand that may occlude parts of the face.

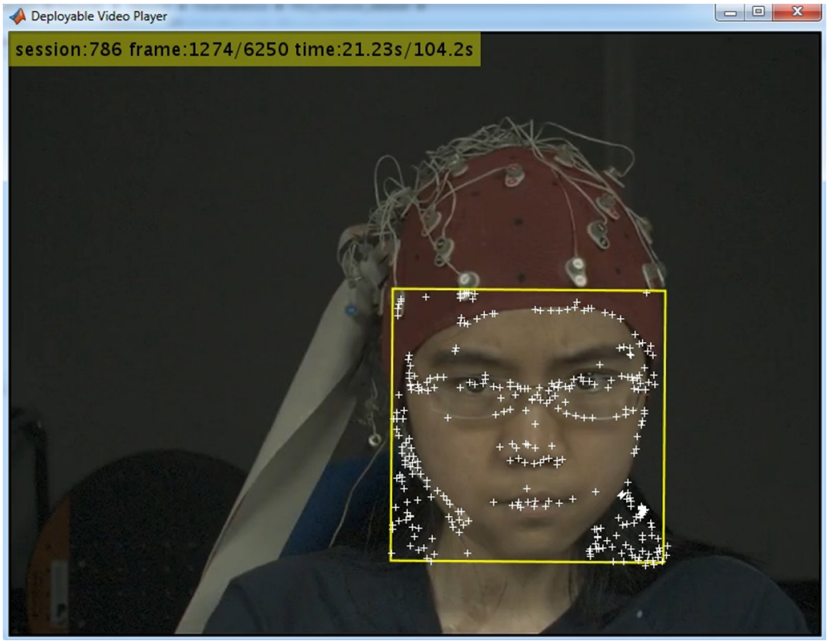
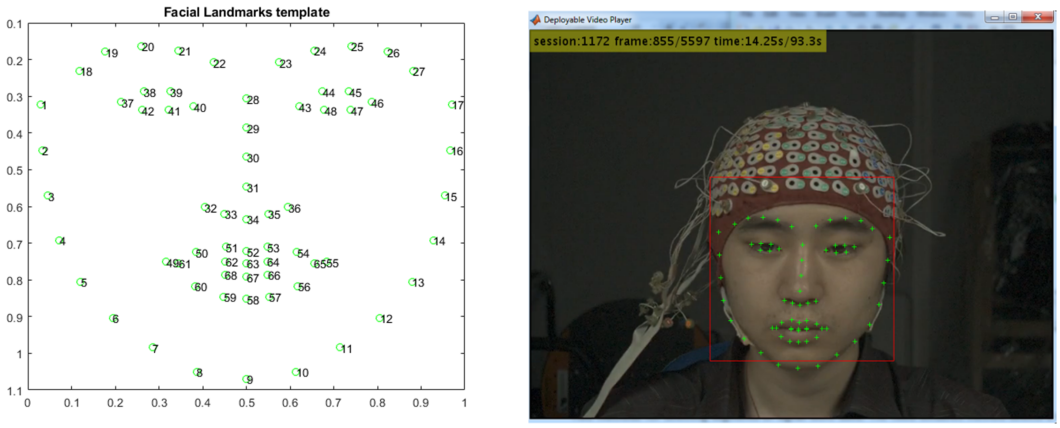


Figure 4.1: Example of face detection in Mahnob- HCI dataset showing the feature points tracked along the video

4.4.2 Face landmarks detection

Using the detected ROIs (See section 6.6.2), a trained model (DLIB) for 68 facial landmarks detection was used for each frame in the video Kazemi and Sullivan (2014). DLIB library can be obtained from `http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2`. The processing time for this task was around 100 seconds per video (i.e., approximately 30 minutes per subject).

Figure 6.2 shows the model template (a) and one example video frame with detected facial landmarks (b) adjusted to relevant facial structures (mouth, eyebrows, eyes, nose and face borders).



(a) Model template

(b) Detected facial landmarks in one example video frame

Figure 4.2: Facial landmarks detection.

4.4.3 Face features extraction

We extracted the following features from facial landmarks (see figure 6.3),

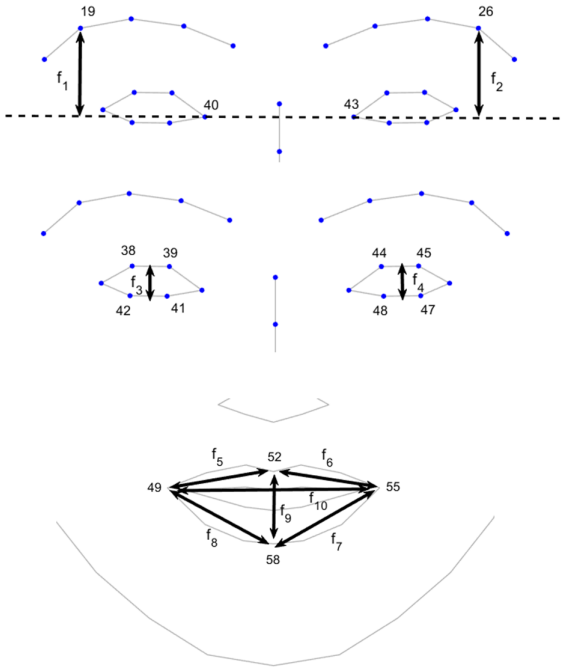


Figure 4.3: Facial features

1. Vertical distance between the horizontal line connecting the inner corners of the eyes and outer eyebrow (f1, f2).

2. Vertical distances between the upper eyelids and the lower eyelids (f3, f4).

3. Distances between the upper lip and mouth corners (f5, f6).

4. Distances between the lower lip and mouth corners (f7, f8).

5. Vertical distance between the upper and the lower lip (f9) and distance between the mouth corners (f10)

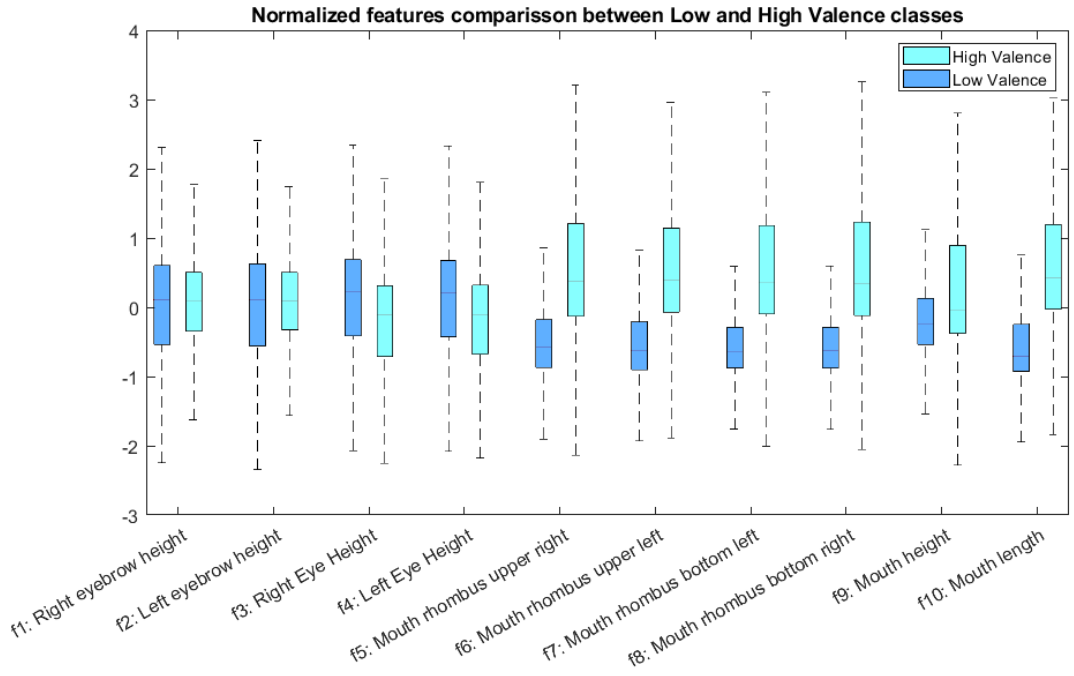


Figure 4.4: Boxplot for features in MAHNOB-HCI dataset for valence emotional dimension

We assume that participants hold a neutral face while first 2 seconds after starting the stimulus. As we want to detect changes in facial features, therefore the mean features in first 2 s are subtracted from facial features for each response video. Figure 6.4 shows the distribution of normalized features. It can be noted that mouth-related features such as mouth length have better discriminative power between low and high valence. Outliers are omitted for visualisation purposes.

4.4.4 Event detection

The event detection method is based on Otsu's algorithm Otsu (1979), a very popular method used for grayscale image segmentation. A power signal for the facial features is computed using the sum of the square of features. Otsu's algorithm is then used to select a histogram threshold by minimization of intra-class variance. The *graythresh* function from Image Processing toolbox in Matlab was used for this purpose. This function returns the threshold and an effective metric (EM) which is a value in the range $[0, 1]$ that indicates the effectiveness of thresholding the signal.

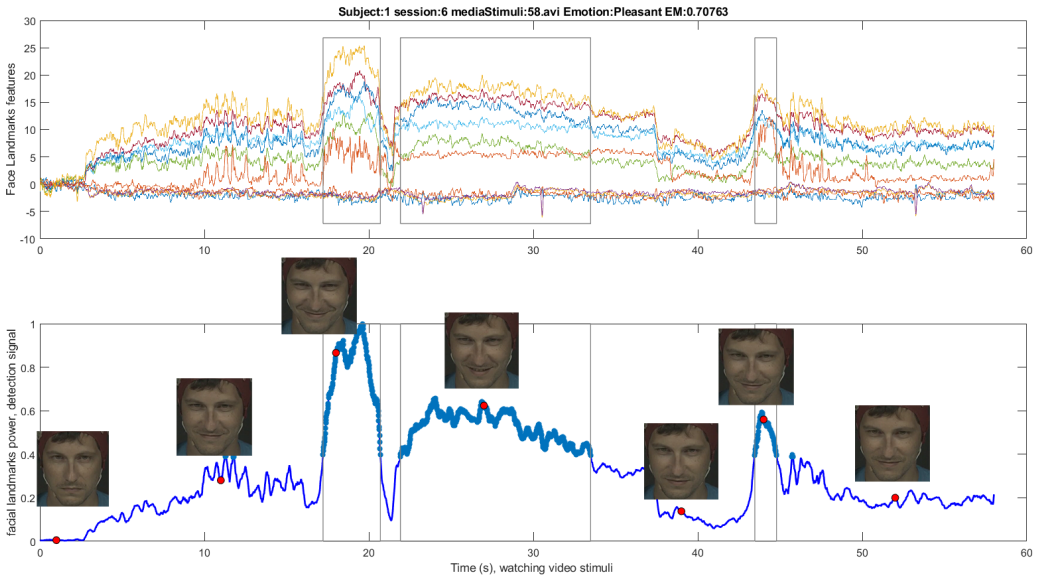


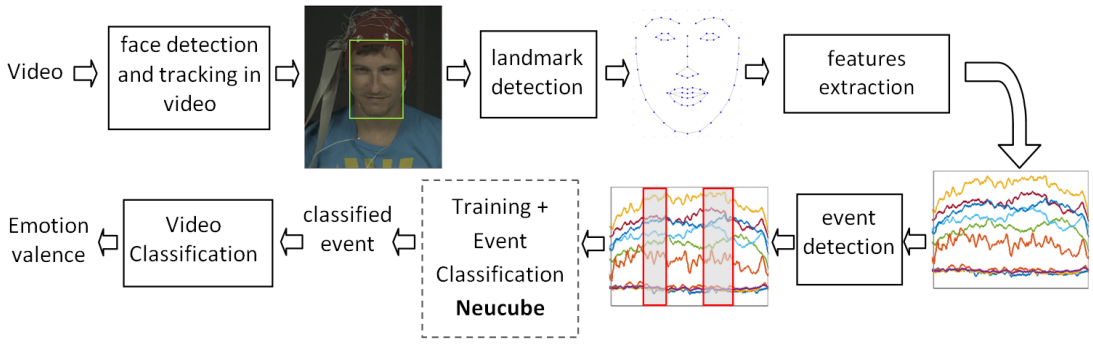
Figure 4.5: Event detection based on facial features power thresholding. Upper: features landmarks. Bottom: Facial features power, signal above threshold in bold, rectangles for detected events, red circles mark some points where the face is shown.

As we wanted to capture those video segments where facial expression occurs for classification purposes, the features power signal was mapped to 0-255 values, which was used as the input to *graythresh* function. Then we detected signal segments (events) where a signal was above the threshold for at least one second. This procedure resulted in events with different widths, but all detected events (DE) are resampled to a sample length of 64. The unique parameter used in the event detection is the minimum dwell time (DT) above the threshold ($DT = 1$ second),

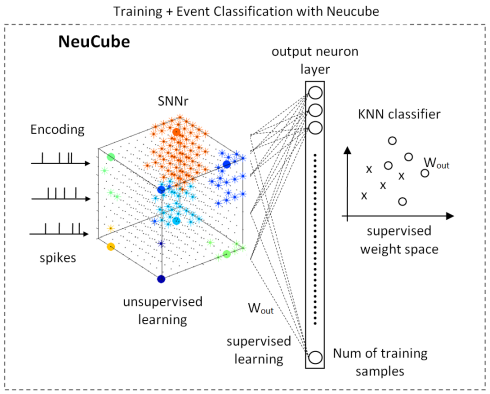
but EM could be used for selecting the more relevant events. Figure 6.5 shows an example of features facial power signal for a video with valence labelled as pleasant with three events detected (rectangles). It can be noted that faces inside event windows have more expressive smiles than those outside them.

4.4.5 NeuCube SNN for facial emotion recognition

We used an eSNN architecture called NeuCube proposed in N. Kasabov (2014) to build a system for emotion valence classification. A general scheme of our approach based on NeuCube is presented in Figure 4.6b. NeuCube structure includes Encoding, SNN reservoir (SNNr), output neuron layer and KNN classifier. Training and classifying spatio-temporal data using NeuCube have the following stages:



(a)



(b)

Figure 4.6: a) Proposed method for emotion valence classification using NeuCube. b) NeuCube architecture.

- **Encoding:** Encode the spatio-temporal data (features) into trains of spikes.

- **SNNr:** Construct a recurrent 3D SNNr and initialize the connection weights among neurons.
- **Input neurons location:** Locate the input neurons in the SNNr keeping related inputs near in space.
- **Unsupervised learning:** Feed the SNNr with training data to learn in an unsupervised mode the spatio-temporal patterns in the data.
- **Supervised learning:** Construct an eSNN classifier to learn to classify different dynamic pattern in SNNr activities.
- **Classification:** Feed the SNNr with testing data for classification purposes.

We briefly explain each stage in the following sections.

Encoding

The coding method we used was inspired by Gaussian Receptive Field population-based sparse coding proposed in Bohte, Kok and La Poutre (2002); Bohte, La Poutré and Kok (2002). This method codes each continuous value from a time-based feature to spikes emitted at different times by a neuron population. The whole feature range is covered for the neurons and the time for generating the spikes depends on the distance from the current value to the centre of a Gaussian receptive field covering each value interval. We used a population of five neurons per feature, in which only a neuron from the group spikes at the current time step. Fig. 6.7 shows an example of coding the mouth length feature. Note that the dimension of feature is 64 and the temporal dimension of each spikes train is 129 because zeros are inserted between the spikes.

It can be noted from the distribution of mouth length feature (Figure 6.7, left plot; blue: low valence, red: high valence, black: low and high valence), that there are two peaks in the distribution indicating the separation between the two class. In the middle plot (Figure 6.7), the time course of mouth length feature in a low

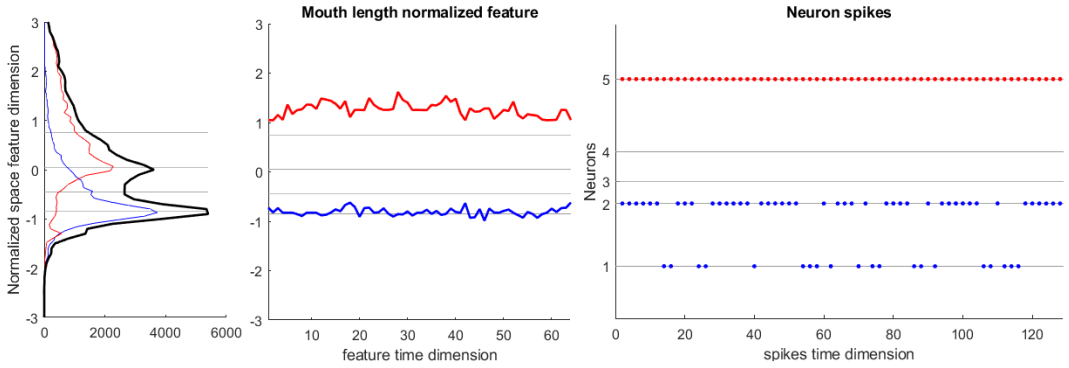


Figure 4.7: Encoding Continuous feature values to five neurons spiking

valence event (blue) and one high valence event (red) for the subject 1 are shown. In the right plot (Figure 6.7) spikes generated for these two events are shown (low valence in blue, high valence in red). Levels that define the receptive fields or range for exciting each neuron are defined using the feature distribution in the data from all detected events for all analyzed subjects. Levels for each five neuron population are obtained automatically by analyzing the histogram in such a way that the five ranges have the same count of value occurrences. Levels are shown as gray lines (Left and middle plots in Figure 6.7). Note that each feature value in time produces a spike in only one neuron from the population. As we have ten facial features, fifty input neuron are allocated in NeuCube network.

Construction of SNNr

When brain imaging data such as EEG is used, the SNNr can be built with a shape resembling the human brain N. Kasabov (2014) and the input neurons can be located based on the anatomical location of the EEG electrodes. However, in this study, as we are building a general classifier of facial features, we chose to build an $11 \times 11 \times 6$ array of neurons (equally spaced in x and y axes) as shown in Figure 6.8.

The SNNr was made with leaky integrate and fire model (LIFM) spiking neurons with recurrent connections. In this neuron model, the post-synaptic potential (PSP) increases or decreases with every input spike from pre-synaptic neurons. The effect of each spike is modulated by the corresponding synaptic connection weight. If

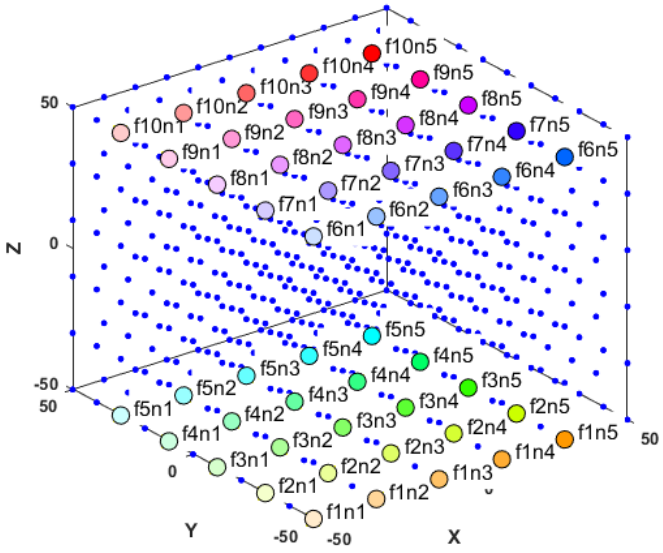


Figure 4.8: Input neurons location for facial features classification. Neurons coding the same feature are shown in degraded colour. n1 means for the neuron coding the lowest values and n5 the highest ones

PSP reaches a specific threshold (0.5 in this work), the neuron emits an output spike toward its connected neighbours and the PSP resets to a reference value. The PSP can leak between spikes with a predefined time constant τ, if we are using an exponential model or a constant leak time. The latter is used in this work and is set to 0.002. After a neuron outputs a spike, the absolute refractory time (equal to 1 in this work) is simulated by disabling it to increase the PSP until a certain unit time has passed. Figure 6.9 shows an example of LIFM neuron simulation with a refractory time equal to 3 seconds, potential leak rate equal to 0.02, a threshold of firing that is equal to 0.5 and synapses weights of 0.1, 0.1 and 0.35. It can be noted in figure 6.9 that the accumulation of spikes in time lead to an increase of PSP until a spike is generated and the effect of disregarding input spikes immediately after a spike is generated.

We set the initial connections (synapses strength) between neurons in SNNr using small-world connectivity Braitenberg and Schüz (2013); Bullmore and Sporns (2009). The connection probability was set such that neurons were more likely to be connected to neighboring neurons than to the distant ones. It has been shown that

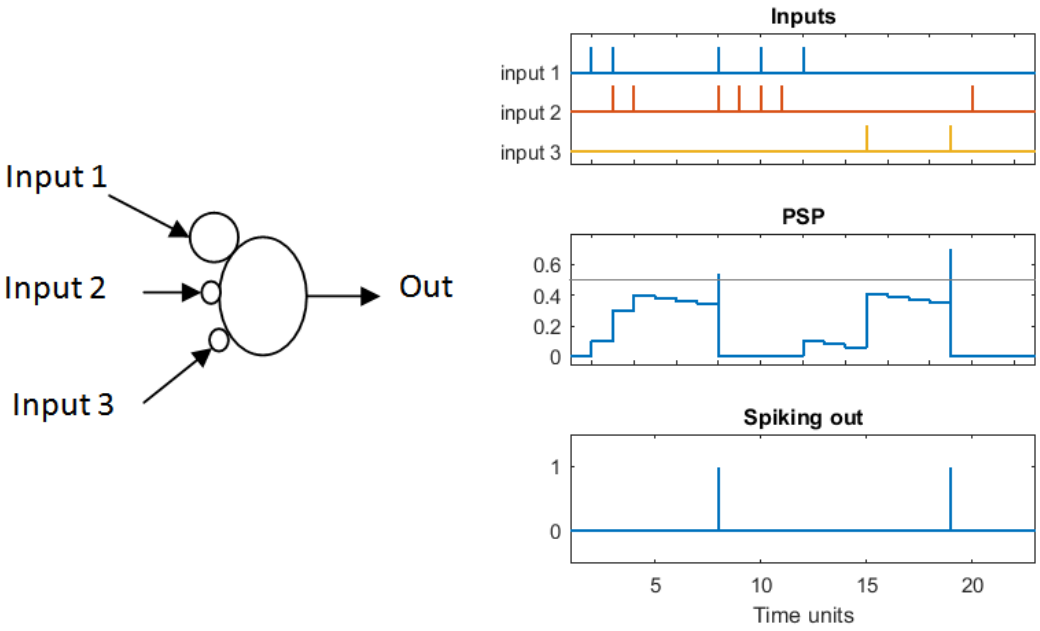


Figure 4.9: LIFM neuron model. Small circles at neuron inputs represent connection weights. Note that input 1 has a bigger weight and it produces a larger effect in PSP

such an approach brings some advantages with regard to learning speed, parallel processing, and also favours the linking of specialized processing cluster units Simard, Nadeau and Kröger (2005). Additionally, we defined a radius r to be the maximum distance of connections of one neuron to another in the reservoir ($r = 25$ in this study). The initial weights were assigned as the product of random values $[-1, +1]$ divided by Euclidean distance between pre-synaptic and post-synaptic neuron so that 80% of them were positive values (excitatory connection) while 20% of them were negative values (inhibitory connections). Neuron connections are unidirectional, and the direction of communication was selected randomly. Connections between input neurons and other neuron are always positive and with doubled weight in comparison with other random connections. These connections were modified in the unsupervised learning stage to adapt to spatio-temporal patterns in input data.

Input neurons location

Each of the five neuron population were spatially arranged in NeuCube structure in lines as illustrated in Figure 6.8. This arrangement enabled adjacent neurons to code

similar feature values favouring spatial neuron specialisation. Since facial features were defined as distances between facial landmarks points, the proposed mapping framework also maintained spatial differences in spatially separated neurons.

Unsupervised SNN training

We adjusted the connections between neurons using the training data and a learning rule-based on Hebbian plasticity called spike-time-dependent plasticity (STDP) S. Song et al. (2000). STDP learning modifies the neuronal connection weights taking into account the time difference between post- and pre-synaptic firing. A connection is strengthened, if postsynaptic firing occurs after presynaptic firing (long-term potentiation); otherwise, it is decreased (long-term depression). After STDP learning, the spatio-temporal pattern was saved in the value of connection weights in the SNNr. STDP learning rule is given as,

$$\Delta w = \text{sgn}(\Delta t)\frac{LR}{|\Delta t| + 1} \tag{4.1}$$

where LR is the STDP Learning Rate (0.001 in this work), sgn$(\cdot)$ is the function sign (-1 for negative values and 1 for positive), Δt is the difference between post- and pre-synaptic times ($\Delta t = t_{post} - t_{pre}$) and Δw is the change in the connection weight. The Hebbian relation Δw vs Δt is depicted in Figure 6.10 (solid line). Equation (4.1) is a computational efficient and simple approximation to the well know decaying exponential relation $\Delta w(\Delta t)$ found in literature such as in S. Song et al. (2000) and depicted in Figure 6.10 with dashed line for comparison purposes. Guyonneau, VanRullen and Thorpe (2005) have shown that neurons trained by STDP with repeated inputs in a SNN become selective to patterns with short response latency, i.e. neurons receiving similar spikes trains tend to focus on inputs that consistently fire early. Consequently, Neurons responds faster to learned patterns by using just first spikes.

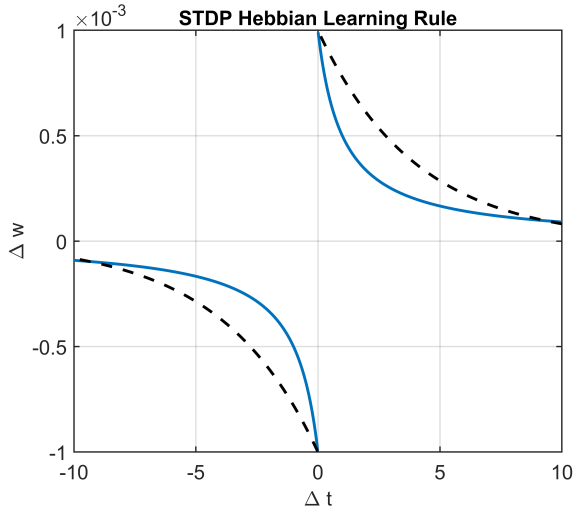


Figure 4.10: Hebbian Learning rule, connection (synaptic modification) vs difference between post- and presynaptic times. Simple approximation embedded in our SNN (solid line) and corresponding decaying exponential (dashed line).

Supervised output neurons training

The deSNN is applied for supervised learning N. Kasabov, Dhoble et al. (2013). For every single training sample, an output neuron was created and connected to all the neurons in the trained SNNr (see figure 6.6). Each output neuron was trained using the corresponding training sample by propagating the signal through the network once more. The neuron's connections weights $w_{i,j}$ between neurons i (in the reservoir) and j (output neuron) were initially established using rank order (RO) rule N. Kasabov, Dhoble et al. (2013). The RO method ranks the order in which the first spike arrives in the j neuron and the weights are given as,

$$w_{i,j}(0) = \alpha m^{\text{order}(i,j)} \tag{4.2}$$

where α is a learning parameter (in a partial case, equal to 1), m is a modulation factor that defines how important the order of the spike is (0.8 in this study), order(i,j) represents the order (the rank) of the first spike at synapse (i,j) ranked among all spikes arriving from all synapses to the neuron j. Furthermore, order$(i,j) = 0$ for the first spike to neuron j and increases according to the input spike order at other synapses.

Once a synaptic weight $w_{i,j}$ is initialized, based on the order of the first spike from i to j, the synapse becomes dynamic. It increases its value with a small positive value (drift $= 0.005$) at any time t a new spike arrives at this synapse and decreases its value if there is no spike at this time, as described in the following formula,

$$w_{i,j}(t) = \begin{cases} w_{i,j}(t-1) + \text{drift}, & \text{if} \quad S_{i,j}(t) = 1 \\ w_{i,j}(t-1) - \text{drift}, & \text{if} \quad S_{i,j}(t) = 0 \end{cases} \tag{4.3}$$

where $S_{i,j}(t)$ describes the existence of spike from neuron i entering to neuron j at time t. Every generated output neuron was trained to recognise and classify spatio-temporal patterns of weights adjusted by a corresponding labelled input training sample.

Classification

At the classification stage, the NeuCube was fed with validation data. For each sample data, synaptic weights for output neurons were calculated using the same supervised rules used in the supervised training procedure. The connection weights learned in this process were then classified using a K-nearest neighbour (KNN, with $K = 3$ neighbours) algorithm and the labels that were known for all the samples. To diminish the randomness due to NeuCube initialisation, we ran five times the whole NeuCube framework to classify each detected event in a Leave-One-Subject-Out (LOSO) mode and then applied majority vote for assigning the class to the current testing event sample. Since we used NeuCube framework for event classification, and eventually, there are multiple events per video, we used the majority vote of event classification for labelling the videos.

NeuCube parameters

NeuCube performance in analysing spatio-temporal data depends on several parameters. We chose a set of default parameter values equal to that used in the NeuCube development system publicly available online http://www.kedri.aut.ac.nz/

neucube, with the exception in refractory time. We used 1 time unit for this parameter in order to increase neuron activity. The NeuCube parameters used in this work are given in Table 6.1.

Table 4.1: NeuCube parameters

Small world radius (r)	25
STDP learning rate (LR)	0.001
Threshold of firing	0.5
Potential leak rate	0.002
Refractory time	1 second
mod	0.84
drift	0.005
K	3

4.5 Results

As a primer test, we explored the difference between mean face landmarks between two conditions (low and high valence, see Figure 6.11). It can be noted from Figure 6.11), the mouth related landmarks are varying in the most significant way. This visual inspection also agrees with box plot in Figure 6.4 and finding made in Soleymani et al. (2011) about the more relevant features in MAHNOB-HCI dataset.

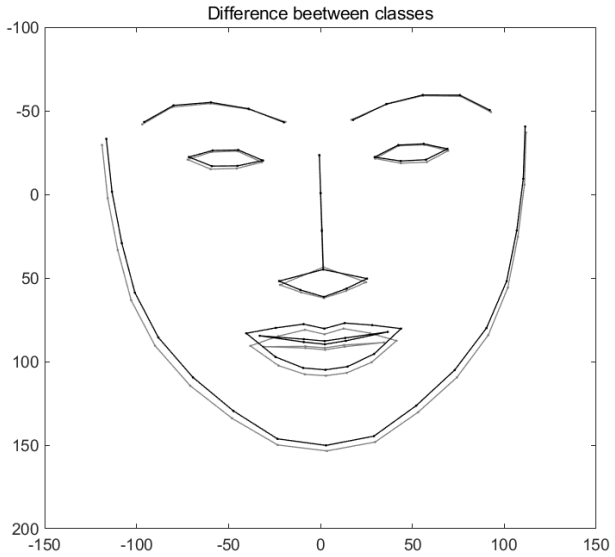


Figure 4.11: Difference between mean facial landmarks for both valence classes.

The effect of DT in event detection was analysed by studying the number of events detected with DT equal to 1 and 2 seconds. Table 6.2 shows how the

Table 4.2: Number of event videos with an event detected regarding dwell time

subject ID	No. of videos rated as low or high	No. of videos with $DT = 1s$	No. of videos with $DT = 2s$
1	15	15	14
2	16	14	11
3	12	12	10
4	14	12	12
5	16	14	13
6	17	17	14
7	16	15	15
8	14	13	11
9	10	9	9
10	13	7	5
11	14	14	12
13	14	11	7
14	14	12	9
16	11	10	8
17	16	8	6
18	12	5	1
19	13	12	7
20	15	15	14
21	15	12	11
22	15	13	10
23	16	14	13
24	13	11	10
25	18	17	15
27	16	14	13
28	15	12	9
29	16	15	12
30	14	12	11

number of events detected is reduced when DT is increased from 1 second to 2 seconds. For $DT = 1$ second, events can be detected in 85.89% of videos, while for $DT = 2$ seconds, events can be detected in 72.30% of videos. Because using $DT = 1$ second retains more events to analyse videos, this value was used for testing out classification method. NeuCube framework was fed with coded data under a LOSO

cross-validation scheme, i.e. all events from a specific subject was excluded from the training set. All the parameters were fixed with the values mentioned in the section 4.4. Table 4.3 shows the event valence classification accuracy (EVCA) and video valence classification accuracy (VVCA) results in MAHNOB-HCI dataset. Note that not all the videos were analysed because there were videos which did not have detected events, as mentioned previously.

Table 4.3: Event and movie valence classification in MAHNOB-HCI dataset using NeuCube

Subject ID	No. of DE	EVCA (%)	Videos analysed (with DE)	Videos with no DE	Videos correctly classified	VVCA (%)
1	52	86.54	15/15	0	12/15	80.00
2	55	87.27	14/16	2	13/14	92.86
3	48	91.67	12/12	0	10/12	83.33
4	39	89.74	12/14	2	10/12	83.33
5	58	87.93	14/16	2	11/14	78.57
6	39	76.92	17/17	0	14/17	82.35
7	64	90.63	15/16	1	14/15	93.33
8	42	90.48	13/14	1	11/13	84.62
9	32	87.50	9/10	1	8/9	88.89
10	23	30.43	7/13	6	3/7	42.86
11	50	70.00	14/14	0	8/14	57.14
13	39	82.05	11/14	3	9/11	81.82
14	31	90.32	12/14	2	10/12	83.33
16	36	86.11	10/11	1	9/10	90.00
17	16	62.50	8/16	8	5/8	62.50
18	9	55.56	5/12	7	3/5	60.00
19	32	84.38	12/13	1	10/12	83.33
20	64	53.13	15/15	0	11/15	73.33
21	42	95.24	12/15	3	11/12	91.67
22	41	82.93	13/15	2	9/13	69.23
23	56	82.14	14/16	2	12/14	85.71
24	42	80.95	11/13	2	8/11	72.73
25	66	54.55	17/18	1	10/17	58.82
27	83	72.29	14/16	2	9/14	64.29
28	29	48.28	12/15	3	8/12	66.67
29	44	50.00	15/16	1	9/15	60.00
30	78	84.62	12/14	2	11/12	91.67

In total, 1210 events were detected (640: 52.89% low valence, 570: 47.1% high valence) in 390 videos (207: 53.07% low valence, 183: 46.92% high valence). From these 390 videos, 55 (14.1%) were disregarded because they do not have relevant facial expression events. Event and movie binary classification accuracy equal to 77.52% and 77.01% were respectively achieved in 1210 events and 335 videos.

Some additional tests were performed to explore the effect of some parameters. For example, the NeuCube resolution was increased by building a Neuron reservoir with $11 \times 11 \times 11$ neurons and we obtained a slightly lower accuracy of 76.42% in movie classification.

In order to test that effectively the SNN is taking advantage from the time course of features, we flattened the signal features signals in each detected event (repeating the mean 32 times) and we obtained a 73.64% in event classification accuracy and 71.34% in video classification accuracy. These values, which is lower than the benchmark approach, suggest that NeuCube framework is using the time variation of features for reaching 77.01% in movie accuracy classification. We also, set interpolation length for features to 16 instead 64 and this reduced the movie classification accuracy to 74.33%.

Since it is obvious that mouth-related features are the more significant discriminant (as depicted in the boxplot in Figure 6.4), we tested the NeuCube framework using only the mouth related features (f5 to f10) and an accuracy of 71.07% in movie classification was achieved. This result suggests that all other features are also important for achieving high accuracy.

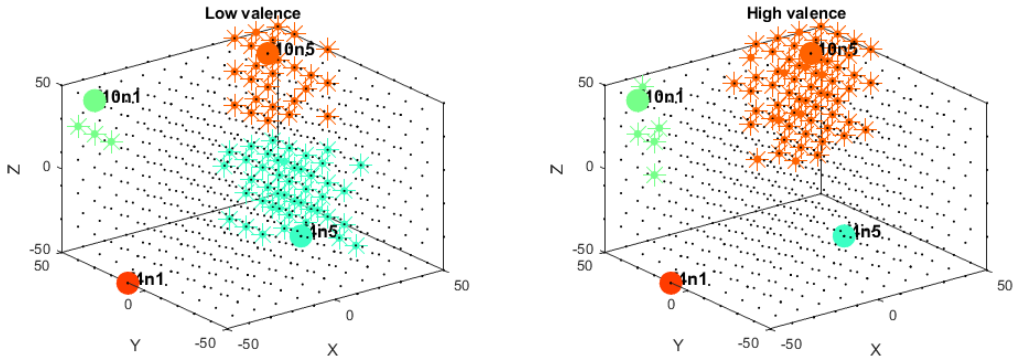
The last test was to increase the DT to 2 sec, and an accuracy of 76.24% was achieved. Table 4.4 shows how classification accuracy is affected as certain NeuCube parameters are changed.

Table 4.4: Testing some parameter variation effect in movie classification accuracies

Parameters	Video accuracy (%)
None	77.01
NeuCube resolution: $11 \times 11 \times 11$	76.42
Feature signal to 32 constant value (mean)	71.34
Interpolation of features to 16 points instead of 64	74.33
Only mouth features	71.07
DT 2 seconds	76.24

4.5.1 Clustering Spike Communication

NeuCube framework has an option to analyse clusters of neuron-surrounding input neurons using the spiked amount communicated between a pair of neurons. Figure 4.12 shows an example using this tool when the neuron reservoir is trained separately with one class (low valence) and the other one (high valence). For visualisation purposes, only input neurons f10n1, f10n5, f4n1 and f4n5 are shown. These neurons correspond with the neurons that code for low and high values of mouth length and low and high values of left eye height features respectively. Figure 4.12 shows that there is more activity related with f10n5 for high valence data than that in low valence data (orange stars) and there is more activity related with f4n5 for low valence data than that in high valence data. These results again agree with features distribution shown in Figure 6.4.



(a) Low valence **(b)** High valence

Figure 4.12: Neuron activity pattern example when NeuCube is trained using each separate data (low and high valence).

4.6 Discussion

In this work we developed an approach based on NeuCube N. Kasabov (2014), which is an eSNN framework, to classify emotional valence. We used a population coding scheme, based on Gaussian Receptive Fields to encode input data into spikes, that SNNs can handle. When tested on the benchmark dataset, the MAHNOB-HCI, our approach resulted in a accuracy about 77 % for both event and movie classification.

4.6.1 Related work

The MAHNOB-HCI dataset has been used in several studies due to its difficulty for the classification of spontaneous emotional responses from subjects. A convolution deep belief network (CDBN) was proposed in Ranganathan, Chakraborty and Panchanathan (2016) to learn emotional features from multimodal datasets and the authors reported a classification accuracy of 58.5% with the MAHNOB-HCI dataset (see accuracies comparison Table 4.5). Since MAHNOB-HCI dataset also contains physiological signals such as EEG, ECG and GSR, several studies have used either just the physiological signal(s) or combined information from facial expressions and physiological data. Using group nonnegative matrix factorization technique, Hajlaoui and colleagues obtained classification F1-score of 0.69 and 0.59 for valence and arousal respectively, using information only from the EEG sensors Hajlaoui, Chetouani and Essid (2018). Koelstra and Patras combined EEG and facial expressions for the first time, to perform affect recognition and implicit tagging Koelstra and Patras (2013). They used power spectral density (PSD) features for EEG, where as for facial expression they used facial Action Unit (AU) detection method, originally proposed in Koelstra, Pantic and Patras (2010) which uses Free-form Deformations (FFDs) and Motion History Images to detect AUs and their temporal models. For facial recognition, they trained the system using the MMI dataset Valstar and Pantic (2010) and obtained 64.5% of binary valence classification using only facial features and 74% by combining facial and EEG features. Although single-channel features

such as PSD estimates are commonly used for emotion recognition, multivariate measures based on functional brain networks can also be used. This is motivated by the fact that interaction between the brain regions play a key role in emotional states and functional network measures such as phase locking value (PLV) reflect such interaction. Li and colleagues used Phase Locking Value (PLV) to estimate functional networks and used various network statistics in combination with single-channel features such as PSD and used an SVM classifier after performing feature selection using F-score P. Li et al. (2019). They reported a classification accuracy of 68% on the MAHNOB-HCI dataset. Boxuan and colleagues developed a temporal information preserving framework by splitting signals into multiple stages in each video. They achieved a valence (unpleasant, neutral, pleasant) classification accuracy of 54% using only facial expression and 69% when fusing with physiological signals Zhong et al. (2017). They used Affdex SDK software McDuff et al. (2016), trained in 10,000 manually labelled facial images, which classify emotion-based on HOG features and support vector machine (SVM) classifier. Huang and colleagues obtained 50.57% for valence classification using appearance descriptors based facial features (Local binary pattern from three orthogonal planes, LBP-TOP) and 66.28% using fusion it with global EEG features X. Huang et al. (2016). They used the LOSO cross-validation scheme in nine emotion categories.

Table 4.5: Comparison with related works on valence classification using Mahnob-HCI dataset

Works	Features	Method	Classes	Crossvalidation	Accuracy %
Koelstra and Patras (2013)	Facial + EEG	Free-form Deformation and Motion History Images	Binary valence	Trained with MMI dataset, and data from the same subject	74
Zhong et al. (2017)	Facial + Physiological	Temporal Information Preserving Framework, SVM	Valence (3 classes)	LOSO	69

Works	Features	Method	Classes	Crossvalidation	Accuracy %
X. Huang et al. (2016)	Facial + EEG	LBP-TOP, Transfer learning CNN, SVM	9 emotion categories.	LOSO	66.28
Ranganathan et al. (2016)	Facial + Body + Physiologic	Convolutional deep belief network (CDBN) and SVM	Not mentioned.	LOSO	58.5
Hajlaoui et al. (2018)	EEG	group nonnegative matrix factorization	Binary valence	LO-session-O Train with data from each subject	F1-score = 0.69
P. Li et al. (2019)	EEG	PLV, PSD, F-score, SVM	Valence (3 classes)	10-fold cross validation for each subject. Train with data from each subject	67
Torres-Valencia, Álvarez-López and Orozco-Gutiérrez (2016)	EEG + Peripheral	Discriminant-based algorithm, SVM	Binary valence.	80% train data-20% test data	66.09
J. Liu, Su and Liu (2018)	Facial + EEG	LSTM-RNN	Valence, 9 classes	24 subject training and 3 for testing	74.5
Y. Huang, Yang, Liu and Pan (2019)	Facial + EEG	Pretrained CNN	Binary valence	LOSO	75.21
Tan, Ceballos et al. (2020)	Facial + Peripheral	SNN, feature-level fusion	Binary valence	LOSO	73.15
Tan et al. (2021)	EEG + Facial	SNN	Binary valence	LOSO	72.12
W. Zhang, Yin, Sun, Tian and Wang (2020)	EEG	shared-subspace feature elimination	Binary valence	LOSO	65.37
Y. Zhu, Wang and Ji (2014)	EEG+video stimulus	Canonical Correlation Analysis + SVM	Binary valence	leave one video out	58.16
T. Song, Lu and Yan (2020)	Peripheral	CNN	3 classes	LOSO	59.88
Wiem and Lachiri (2017)	Peripheral	SVM	Binary valence	No mentioned	68.75

Works	Features	Method	Classes	Crossvalidation	Accuracy %
R. Li et al. (2021)	EEG+Facial	CNN+LSTM (long short-term memory network)	Binary valence	train with 20 subjects Evaluate in 3 specific	78.56
Ours	Facial	SNN	Binary valence	LOSO	77.01

For comparative purposes Table 4.5 shows some other methods reported in literature for Valence Emotion Recognition in MAHNOB-HCI dataset. ANNs approaches stand out with higher accuracies such as those in J. Liu et al. (2018) using LSTM-RNN, Y. Huang et al. (2019) with CNN, R. Li et al. (2021) with CNN + LSTM and our works based in SNN, Tan, Ceballos et al. (2020); Tan et al. (2021).

It is evident from the aforementioned studies that by fusing information from facial expression and physiological data such as EEG, a higher classification accuracy can be obtained. Given that the acquisition of physiological signals can be time-consuming and laborious compared to acquiring videos/images, there is a lack of availability of large-scale physiological datasets pertaining to such tasks Siddharth, Jung and Sejnowski (2019). Furthermore, physiological signals can be of varying nature (EEG vs ECG) and riddled with artifacts, deep learning methods have not been fully exploited to utilize information from physiological signals. Recent efforts have been pursued in the direction of addressing these issues (for example see Siddharth et al. (2019)).

Despite using only information from facial expression, we observe that the NeuCube framework could provide comparable classification accuracy to studies that also exploit information from EEG (except for subject 10, which resulted in poor classification accuracy, which needs to be investigated further). Thus, it seems logical to us that by fusing facial expression with physiological data, should lead to increased classification accuracy. A future direction of research will be to test if the NeuCube framework can outperform deep learning methods when both are supplied with multimodal data (facial expressions, EEG, ECG, GSR, etc.).

4.6.2 Limitations and future work

We recognize that our study has several limitations that need to be addressed in the future. We tested our approach only on one benchmark dataset. To test whether the algorithm is generalizable, testing on other dataset such as DEAP, AMIGOS, etc., should be carried out.

We assumed that the face captured during the first two seconds after the stimulus is neutral and consider it as the baseline. This could be problematic, especially if the participant is tired. Since we used Otsu's algorithm Otsu (1979) for event detection, we do not take into account the long-lasting facial expression. It could be interesting if long term facial variation inside the video could be considered as detected events. Also, it could be interesting to incorporate detecting facial micro-expressions in our framework but this is in general challenging due to limited availability of such data and as well as difficulties in analyzing minute changes in expression Y. Wang, See, Phan and Oh (2015). Few methods have been proposed to address the problem of detecting micro-expressions using spatio-temporal local texture descriptor X. Li, Pfister, Huang, Zhao and Pietikäinen (2013), Gabor filter with SVM classifier Q. Wu, Shen and Fu (2011) and LBP-TOP with nearest neighbor classifier Guo, Tian, Gao and Zhang (2014), which can be incorporated to add more information for the SNN framework. Another improvement could be to normalise expressions between subjects by using pose estimation Murphy-Chutorian and Trivedi (2008), correction of a 3D model Jourabloo and Liu (2015); X. Zhu, Lei, Liu, Shi and Li (2016). Further improvements could be made along the lines of detecting non-frontal head poses, identity bias and as well as illumination variation.

Although we studied the effect of varying certain NeuCube parameters, the performance of the proposed system may be affected by the choice of several other parameters. For instance, effect of varying other parameters such as radio, firing threshold, refractory time and NeuCube resolution resolution should be carefully

investigated. The NeuCube framework also provides parameter optimization tool, which could be utilized instead of setting the parameters in an ad hoc manner.

4.7 Conclusion

Deep learning methods utilizing ANNs have been leveraged to solve the problem of FER in affective computing. SNNs offer biologically more realistic models of neurons compared to ANNs.

In this work, we proposed a novel solution of using a variant of SNN known as NeuCube which is an eSNN, to solve the FER problem and exploited some of the ECOS principles on which eSNN is based on, such as efficient processing of spatio-temporal data and open evolving structure.

We tested our approach on the MAHNOB-HCI dataset. We also proposed methods for feature extraction using event detection based on Otsu's thresholding algorithm. We showed that our approach, utilizing only face expressions, gives results comparable to deep learning methods that utilize multimodal data including face expression and physiological signals.

5

NeuroSense: Short-Term Emotion Recognition and Understanding Based on Spiking Neural Network Modelling of Spatio-Temporal EEG Patterns

5.1 Prelude to Chapter 5 Manuscript

Human-machine interactions are now an integral part of daily life. These have, in turn, given rise to a need for making these interactions more humane and organic. This need has, in turn, lead to the field of affective computing. Affective computing systems primarily work by identifying human emotions and then facilitate an appropriate emotional response (or affect) from the machine. Multiple inputs can be used in affective computing ranging from physical attributes such as expressions, voice changes, and body language, and physiological data such as EEG readings and body temperature variations. This study uses EEG data in conjecture with facial expressions to enable Affect recognition from EEG data. However, before proceeding to the specifics of the study, the paper presents a concise overview of the research in this area to highlight the importance of ECH based affect recognition. The study employs a third-generation spiked neural network (SNN) for its purposes. Specifically, a dynamic, evolving spiked neural network (deSNN) is employed. Unlike the previous generation of neural networks, these spiked artificial neural networks are better suited to accommodate the non-linearities associated with biological systems and process information as spiked signals much like a biological neuron.

The study uses a three-dimensional SNN structure termed as NeuCube, which is essentially a collection of eSNNs configured in a 3d format for spatiotemporal reasoning. Such a system can encapsulate the structural information in terms of neural networks and connectivity, just like a biological brain. Unlike other SNN configurations, this 3D approach is well suited for Spatio-temporal brain data modelling. In this particular study, facial expressions from video data are correlated with variations in EEG to identify EEG variations specific to a particular emotional affect. Two databases, namely the DEAP and MAHNOB-HCI, are utilised for this purpose. Of these, DEAP has multimodal physiological data, including EEG, EEG, and EMG. Frontal videos were available for 22 of the 33 subjects recorded. Therefore, 22 subjects were compatible with the requirements of this study. MAHNOB-HCI, on the other hand, contains 27 subjects whose frontal videos and physiological data were available. Using a tracking algorithm on the videos, specific arousal and valance classifications were obtained. The samples for a specific participant was utilised if there were at least five samples obtained pertaining to that participant. This led to a final data set of 214 samples for arousal and valance classification each from the DEAP dataset and 313 arousal classifications and 191 valance classifications from the MAHNOB-HCI dataset. Once the data sample was determined, the next step was spike encoding where the EEG data was encoded for each subject.

A simple threshold-based representation is employed to generate positive and negative spikes to encode the EEG data of each subject. The necessity for optimising this process is also highlighted. Given that spike encoding is simultaneously noise reduction and information compression, it is essential that no useful information is lost in the process while ensuring a low level of noise. Using Bayesian spike encoding optimisation, the final encoded spike trains are obtained. It is then fed to a 3D SNNr or SNN reservoir. After initialising the connections between neurons to obtain a sparsely connected SNNr, the structure is trained in affect recognition using data available. A similar representation using deSNNs is also carried out. The trained 3D

structures are then used for affect recognition. It is seen that the simple spike-based system has accuracies in the range of less than 65% or DEAP dataset and near 75% for MAHNOV-HCI dataset. However, with the introduction of deSNN, the accuracy values improved across the data sets. In fact, the lowest accuracy value, in this case, is 67.7% for valance classification in DEAP dataset while the highest was 79.4% for MAHNOB-HCI arousal classification. In general, it seems that this method if more suited for arousal classification than valance classification across datasets. Also, the left side of the 3D SNN is found to be more connected than the right, probably because in most people, the left hemisphere is better developed. Besides, brain regions corresponding to the temporal lobe, occipital lobe, and amygdala have more connections.

To summarise, this paper presents a method for identifying and codifying the variations in EEG values when there are emotional changes in a person. This is done by identifying samples from two databases, DEAP and MAHNOB-HCI. These databases contain both frontal visual data and the corresponding EEG readings for subjects. This paper works on correlating specific emotional Affects that appear on the face to specific variations in EEG patterns. The patterns once identified are coded into positive and negative depending on when the changes cross an absolute threshold value. These coded spike trains are optimised and fed to a 3D SNN reservoir. A reservoir utilising deSNN was also tested. It is seen that in general deSNNs perform better than simple SNNrs. Additionally, the patterns established in the 3D structure shows that regions corresponding to the temporal lobe, amygdala, etc. which are expected to be more active during emotional changes do appear prominent in the model as well. In addition, there is more connectivity in the left side of the brain in these 3D structures. This agrees with the knowledge that most people have a better developed left hemisphere than a right hemisphere. When tested, the accuracy values were generally better for deSNN based structures. However, it is

essential that this experiment is carried out with a much larger data sample at a later stage to assess its utility better.

5.1.1 Contributions and Publications

Contributions

1. *For the first time introduces a brain-inspired SNN architecture to recognise, and most importantly, to explain for a better understanding, EEG data measuring two primary emotions: arousal and valence*
2. *EEG is processed directly without a need for extracting hand crafted features, and the method is faster than classical methods.*
3. *Ability to analyse and capture emotional information from the Spatio-temporal EEG pattern using 3D neural network structures*
4. *The subject-specific encoding included in this model makes the model accurate independent of subject-specific facial characteristics, making the method truly subject independent.*
5. *The first use of hyperparameter optimisation for spike encoding*

Publications

1. Tan, C., Sarlija, M., & Kasabov, N. (2021). NeuroSense: Short-Term Emotion Recognition and Understanding Based on Spiking Neural Network Modelling of Spatio-Temporal EEG Patterns. *Neurocomputing*, 434, 137-148. [23238]. https://doi.org/10.1016/j.neucom.2020.12.098

5.2 Introduction

The interaction between humans and computational devices is becoming more and more common with the advent of personal digital devices, wearable systems, and other technological interventions. The field of affective computing (AC) combines computer science and emotion research to enable computational systems to identify the emotional states of users (affect recognition) and to generate responses that humans are likely to perceive as emotional (affect generation). It has also been argued that over time, it may be possible for systems to actually ”feel” emotions

Picard (1995). The introductory work by Picard Picard (1995). was followed by much research at the intersection of diverse fields such as neuroscience, ethics, psychology, and engineering, among others. To this day, the work on AC has resulted in the improvement of systems that are capable of interpreting, identifying, and responding to the emotional states of users. For this purpose, affective computing makes use of various multimodal inputs such as facial images, i.e. facial expression, voice data, biometric data, e.g. physiological changes, and body language of the user. Computational models termed as "affect models" are then employed to make sense of these input parameters and identify the emotional state of the user Tao and Tan (2005). The significance of AC is on the rise, given the increased degree of human-computer interactions. According to a recent study[1], 95% of the American population now owns a cell phone of some kind among which 77% use a smartphone. To compare with, the percentage of smartphone users was 35% in 2011. Instead of the one-sided interactions that humans normally expect from machines, utilising AC can make these systems respond in more effective ways, making the whole technology experience more satisfactory to the user Brigham (2017). Devices often come with built-in sensors that collect user data. The key challenge, lying at the core of AC is recognising the emotional state of a person based on the available data.

Nowadays, AC is finding applications in various fields ranging from gaming Guthier et al. (2016), education and e-learning Duo and Song (2012); Yadegaridehkordi et al. (2019) to medicine Aung et al. (2015); Luneski et al. (2010), wearable devices Picard and Scheirer (2001), robotics Cid, Moreno, Bustos and Núnez (2014), etc. For example, research shows that effective computing systems may aid in the diagnosis of seemingly hidden and unobservable medical conditions such as depression and chronic pain Aung et al. (2015). Similarly, affective systems can provide more empathetic, personalised feedback to students, making online learning more efficient Grafsgaard et al. (2013). As briefly stated earlier, in order to

[1] https://www.pewresearch.org/internet/fact-sheet/mobile/

sense affective states, four types of inputs are typically used: 1) facial expression recognition Ko (2018); Koelstra et al. (2012); Soleymani et al. (2012), 2) voice recognition T.-Y. Huang, Li, Chang and Lee (2019); Mijić, Šarlija and Petrinović (2019); B. Schuller, Rigoll and Lang (2004), 3) gesture recognition Camurri, Lagerlöf and Volpe (2003); Piana et al. (2014) and 4) biometrics Greco, Valenza, Citi and Scilingo (2016); Kukolja, Popović, Horvat, Kovač and Ćosić (2014); X. Zhang et al. (2018). Various instruments are used to gather the aforementioned types of data, like cameras, microphones, sensors and biometric devices (e.g. heart rate, blood pressure, skin conductance or electroencephalography (EEG) measurements). The vast majority of AC research if focused on detailed analysis of the collected data, in order to determine the emotional state of the user, where various machine learning techniques have been employed T.-Y. Huang et al. (2019); Rani, Liu, Sarkar and Vanman (2006); Rudovic (2016). Affect models are usually trained on large datasets of relevant data so that they can then be employed for emotion recognition and affect generation.

Various improvements in model architectures, feature selection methodology Kukolja et al. (2014) and deep-learning-based data representations have led to increases in emotion classification accuracies in the last years. However, exploration of novel approaches and concepts in the analysis of human affect could add new information and thus complement traditionally used approaches and features. As stated above, EEG is one of the well-established modalities in affective research Alarcao and Fonseca (2017); Koelstra et al. (2012); Soleymani et al. (2012). On the other hand, spiking neural networks (SNNs) have recently proven to be successful in modelling, recognition and understanding of EEG Spatio-temporal data in a wide array of domains Tan, Šarlija and Kasabov (2020), as described in section 5.3. Based on the importance of EEG in AC Alarcao and Fonseca (2017), as well as numerous EEG-based applications using SNNs (described in section 2), in this paper, we adapt and apply the concept of evolving spiking neural networks (eSNN) N. K. Kasabov

(2018a) and propose an SNN classification framework for EEG-based short-term emotion recognition.

Main contributions of our paper are:

- EEG signal segmentation strategy based on detection of changes in facial landmarks for short-term emotion recognition (section 5.4.2).
- SNN-based framework for subject-independent short-term emotion recognition (section 5.5).
- Hyperparameter optimisation strategy for spike encoding and the dynamic evolving SNN (deSNN) data representation (subsections 5.5.1 and 5.5.2).
- Comparison of emotion classification accuracies obtained by simple EEG spike-based features vs complex SNN-based representation of EEG spiking patterns (section 5.6).
- Novel SNN-based insights related to the neural mechanisms involved in short-term emotional processing of affective videos.

5.3 Spiking neural networks

Human brains encode information via discrete events known as action potentials or spikes, following an all-or-none principle, where a neuron fires a spike once the accumulated potential reaches a certain threshold, else it remains silent. Due to this binary nature of information representation, the human brain still outperforms the traditional artificial neural networks (ANNs) in terms of both energy and efficiency LeCun et al. (2015); W. Wang et al. (2018). Compared to the traditional ANNs, SNNs utilise a more biologically plausible model of neurons Taherkhani et al. (2020), thus bridging further the gap between neuroscience and learning algorithms. SNNs have shown the ability to integrate information encoded in time, phase, frequency, as well as handle large volumes of data in an adaptive and self-organised manner

N. Kasabov, Dhoble et al. (2013), making them particularly suitable for solving online Spatio-temporal pattern recognition problems. SNNs have been shown to be computationally more efficient than ANNs, both theoretically Maass (1997b); Maass and Markram (2004) and in several real-world applications Bohte, Kok and La Poutre (2002). SNNs have over the past years proven to be successful in several real-world learning tasks such as unsupervised classification of non-globular clusters Bohte, La Poutré and Kok (2002), image segmentation and edge detection Meftah et al. (2010), as well as in various tasks related to modelling, recognition and understanding of EEG Spatio-temporal data N. Kasabov and Capecci (2015); N. Kasabov, Hu, Chen, Scott and Turkova (2013); Kumarasinghe, Kasabov and Taylor (2020), such as Alzheimer's disease classification Capecci et al. (2016), epilepsy and epileptic seizure detection Ghosh-Dastidar and Adeli (2007), predicting human behaviour during decision making Z. G. Doborjeh, Doborjeh and Kasabov (2018), detection of limb movement execution and intention for brain-machine interface (BMI) applications Taylor et al. (2014), classification of activities of daily living J. Hu et al. (2014), modelling of peri-perceptual brain processes Z. G. Doborjeh, Kasabov et al. (2018), distinguishing brain states associated with depression and responsiveness to Mindfulness Training Z. Doborjeh et al. (2019), etc.

The evolving SNN, i.e. eSNN, is a class of SNN that utilises rank order learning S. Thorpe and Gautrais (1998) and was first proposed in N. K. Kasabov (2007). In addition to the open evolving structure which facilitates the addition of new variables and neuronal connections Wysoski et al. (2010), eSNNs have the advantage of fast learning from large amounts of data and can interact with other systems actively. eSNNs also allow the integration of various learning rules such as supervised learning, unsupervised learning, fuzzy rule insertion and extraction, to mention a few, and are self-evaluating in terms of system performance. These aforementioned properties constitute the evolving connectionist systems (ECOS) principles on which the eSNN is based N. K. Kasabov (2018b). In the rank-order learning scheme, the synaptic

weights are adjusted only once, making it not very efficient for Spatio-temporal data, where there may be a need to adjust synaptic weights based on the spikes arriving on a given synapse over time. To overcome this disadvantage, an extension of eSNN known as dynamic eSNN (deSNN) was introduced in N. Kasabov, Dhoble et al. (2013) that combines rank-order learning with temporal learning rules such as spike-timing-dependent plasticity (STDP), which allows dy- namic adjustment of the synaptic weights (more details in subsection 5.5.2). However, both eSNN and deSNN do not encapsulate the structural information of the brain in terms of neuronal locations and their connectivity, which may be crucial for modelling of Spatio-temporal brain data (STBD), such as EEG. The NeuCube architecture, first proposed in N. Kasabov (2012), aims at building an eSNN that incorporates structural as well as functional aspects of the brain along with utilising the unsupervised STDP learning algorithm. Below we give a brief introduction to the NeuCube architecture.

Traditional supervised learning methods such as support vector machines (SVM) or multilayer perceptron neural networks (MLP) typically deal with the spatial or temporal aspects of brain data, but cannot handle the dynamic interaction between these processes N. Kasabov, Hu et al. (2013). Furthermore, such models cannot incorporate any prior structural knowledge of the brain in an unsupervised manner, or handle multimodal brain data, e.g. EEG, functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI), positron emission tomography (PET). NeuCube is a specific implementation of an eSNN, initially proposed to handle pattern recognition problems related to STBD, but has further been developed and modified in order to handle various other types of Spatio-temporal data such as audiovisual data, climate data, seismic data and ecological data N. Kasabov, Scott et al. (2016); Tan, Šarlija and Kasabov (2020). The emotion recognition methodology proposed in this paper is based on the NeuCube framework, which has been adapted and further developed for this specific task, particularly in the directions of subject-specific spike encoding and hyperparameter optimisation (see section 5.5).

5.4 Data preparation

5.4.1 Datasets

Due to the high interest in EEG analysis for emotion recognition, several publicly available multimodal databases have been established to this day, like DEAP Koelstra et al. (2012), MAHNOB-HCI Soleymani et al. (2012), SEED Zheng and Lu (2015) or DREAMER Katsigiannis and Ramzan (2017), all of which include EEG. In this paper, we use DEAP and MAHNOB-HCI databases due to their similarity and availability of frontal face video data, which is needed for our event-detection-based signal segmentation.

DEAP is a widely used dataset for multimodal emotion analysis, consisting of 32 subjects Soleymani et al. (2012). 32-channel EEG and peripheral physiological signals, namely the galvanic skin response (GSR), electrooculogram (EOG), electromyogram (EMG), respiration, plethysmograph, electrocardiogram (ECG) and body temperature were recorded during subjects' exposure to 40 one- minute long affective music videos. After watching each video, each subject rated their emotional experience in five dimensions: valence, arousal, dominance, liking and familiarity. The rating values were on a continuous scale of 1-9, except for familiarity, which was rated on a discrete scale of 1-5. For 22 out of the 32 subjects, the frontal video was recorded, so we used data from those 22 subjects only.

MAHNOB-HCI is a multimodal dataset consisting of 27 subjects which participated in two experiments Soleymani et al. (2012). In the first experiment, similarly to DEAP, each participant watched 20 emotional videos which were between 34.9 and 117 seconds long. Recorded signals included 32-channel EEG, peripheral physiology (ECG, GSR, respiration, skin temperature), eye gaze data, audio data and video from 6 cameras recording facial expressions and head pose. After watching each video, the participants gave an emotional label/tag to the video, as well as rated their emotional experience in arousal, valence, dominance and predictability on a

1-9 scale. The second experiment was related to implicit tagging and was not used in this work.

5.4.2 EEG segmentation based on the analysis of facial landmarks

In this section, we describe the signal segmentation procedure, which was used to detect moments of participants' most intensive emotional engagement, based on detecting events in the time-varying facial landmarks. This step is important due to several reasons:

- The differences in utterance lengths vary between DEAP and MAHNOB-HCI, as well as within the MAHNOB-HCI database itself. With an event-based approach, we make sure all samples which are to be used for later emotion recognition are of the same length.
- The self-assessed emotional experience of the participant is not uniformly distributed across the entire duration of an emotional video but is more likely a result of one or more emotionally intensive events of shorter duration. We aim to capture the occurrences of such events by the analysis of facial landmarks.

The facial-video-based event detection algorithm comprises the following steps, which were taken in both datasets:

1. In each frame of each video, the participant's face was tracked. We have employed an implementation[2] of the Viola-Jones algorithm Viola et al. (2001) to detect participant's faces, noses, eyes, mouth, etc., in the first frame of the video. This step outputs a region of interest (ROI).
2. ROI from the previous step is then tracked from frame to frame based on detection and tracking of specific features by using a minimum eigenvalue feature detection algorithm developed by Shi and Tomasi Jianbo and Tomasi

[2]The *vision.CascadeObjectDetector* object from Matlab's Computer Vision Toolbox.

(1994) and a Kanade–Lucas–Tomasi (KLT) feature tracking algorithm Lucas et al. (1981).

3. Each detected and tracked ROI from the previous step was used as input to a facial landmarks detection algorithm[3] Kazemi and Sullivan (2014). Facial landmarks were 68 specific points on the face, such as mouth corners, eyebrow lines, eye lines, etc., as shown in Figure 5.1.

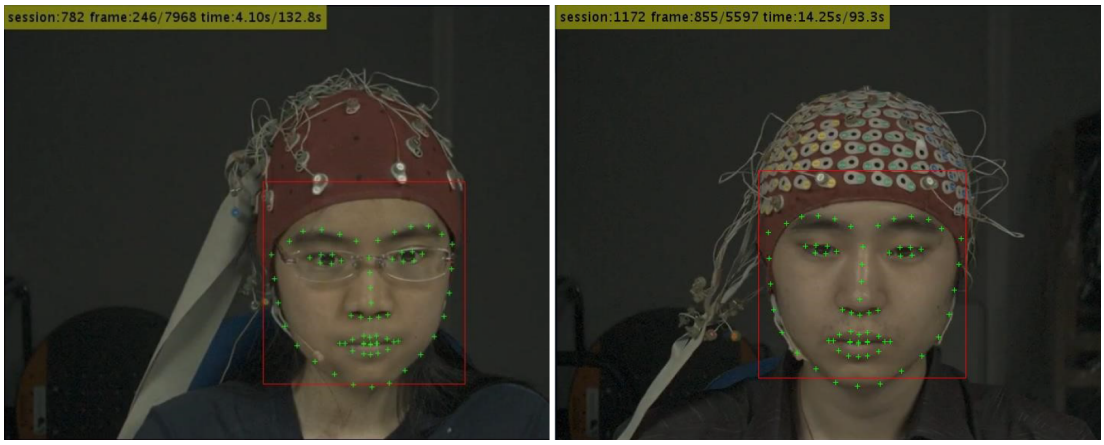


Figure 5.1: Two examples of facial landmarks detection in the MAHNOB-HCI dataset.

4. Based on the detected and tracked 68 facial landmarks, we compute an array of 20 specific geometrical features per frame, related to eyebrow, eye and lip positioning and shape, as described in Soleymani et al. (2012).

5. The energy of time-varying geometrical facial features (see Figure 5.2, top) is calculated disregarding the 10 features based on eye landmarks (features f5 to f14 in Soleymani et al. (2012)), as the eye-based features are very sensitive to blinking. The obtained energy signal is shown in Figure 5.2, bottom.

6. A simple liner-slope-based match filter for detecting increases in the facial features energy signal is applied, resulting in a detection signal (see Figure

[3]The pretrained DLIB model for facial landmarks detection: http://dlib.net/files/shape_predictor_6_face_landmarks.dat.bz2

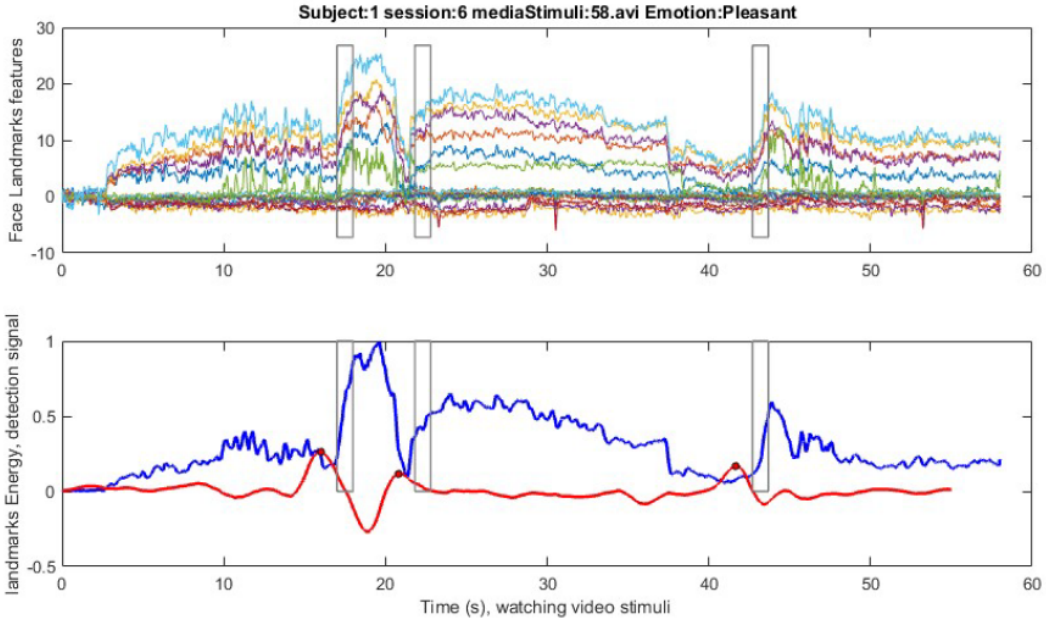


Figure 5.2: Detection of increases in facial activity. The top axes are showing the trajectories of facial features during a participant's exposure to a pleasant video. The bottom axes are showing the corresponding facial features energy signal (blue) and the detection signal (red). In this example, three events were detected.

5.2). A maximum in the detection signal marks the beginning of a one-second-long event. Figure 5.3 shows the face of the participant at the beginning and ending of the first event from Figure 5.2.

Figure 5.3: An example of a change in the participant's facial expression from the beginning to the ending of a detected emotional event.

7. EEG signal segmentation is finally performed based on the detected events from the previous step. EEG signals were preprocessed using the TEAP toolbox Soleymani, Villaro-Dixon, Pun and Chanel (2017), and all 32 available channels are used.

The described procedure resulted with a total of 224 samples for arousal classification and 224 samples for valence classification from the DEAP dataset, and a total of 162 samples for arousal classification and 208 samples for valence classification from the MAHNOB-HCI dataset. Data from participants with less than 5 samples per participant were excluded from the analysis, resulting with a total of 214 samples for arousal classification (125 labelled positively) and 214 samples for valence classification (112 labelled positively) from the DEAP dataset, and a total of 131 samples for arousal classification (47 labelled positively) and 191 samples for valence classification (94 labelled positively) from the MAHNOB-HCI dataset.

The number of extracted samples, as well as the class distributions, are not equal for valence and arousal due to our labelling strategy. For arousal, we used utterances which were labelled as either *calm* or *excited/activated*, thus excluding *medium arousal*. Accordingly, for valence, we also selected 2 classes (*pleasant* and *unpleasant*), thus excluding utterances labelled as *neutral valence*. This resulted with cases for which a participant rated his experience as, e.g. *pleasant* and *medium arousal*, in which case the data is processed for valence analysis and not processed for arousal analysis.

5.5 Emotion recognition methodology

In the following section, we describe our emotion recognition methodology based on SNN modelling of Spatio-temporal EEG spike patterns. Accord- ing to recent reviews of various studies on emotion recognition from EGG Alarcao and Fonseca (2017); Zheng, Zhu and Lu (2017), static features extracted on specific slices of

EEG data dominate the research landscape. These are most commonly related to spectral features, like changes in power over the theta, alpha, beta and gamma frequency bands Y. Huang et al. (2019); Koelstra et al. (2012); Lin et al. (2010) and spectral power asymmetry measures in pairs of symmetrical electrodes Lin et al. (2010); Soleymani et al. (2012). Besides the frequency domain features, which are predominant, various time domain and nonlinear features have been investigated as well Jenke, Peer and Buss (2014), such as Hjorth features Hjorth (1970), fractal dimensions, entropy features etc. Regardless of the feature extraction methodology, traditional approaches usually result in a static feature vector, thus usually well capturing the spatial aspect (emerging from the electrode positions), but neglecting the dynamical Spatio-temporal nature of EEG patterns. Time-frequency domain features are currently the only tool used to capture these dynamical changes in EEG Hadjidimitriou and Hadjileontiadis (2012). With our work, we aim to leverage the Spatio-temporal nature of EEG in emotion recognition by investigating the concept of spike encoding and SNN modelling of EEG. Therefore, we focus only on the comparison of the predictive power of simple spike-based features and complex SNN based Spatio-temporal spike-patterns in the task of EEG based emotion recognition. Accordingly, the possibility of enhancing the predictive power by the integration of EEG-spike-based features with traditional static features described earlier exceeds the scope of our work.

The proposed SNN-based framework for subject-independent emotion recognition includes the following processing steps (illustrated in Figure 5.4):

1. Subject-specific spike encoding of short-term EEG recordings obtained by the segmentation strategy described in section 5.4.2.

2. Unsupervised learning of a brain-like 3-D SNN reservoir (SNNr) module, based on STDP learning.

3. deSNN representation of Spatio-temporal spiking patterns obtained from the trained SNNr, given a specific input spike sequence. The obtained representation forms a static vector which is suitable as input to any traditional supervised learning algorithm (e.g. SVM, ANN, kNN, etc.).

4. Feature selection and supervised learning for emotion recognition from the representation obtained in the previous step.

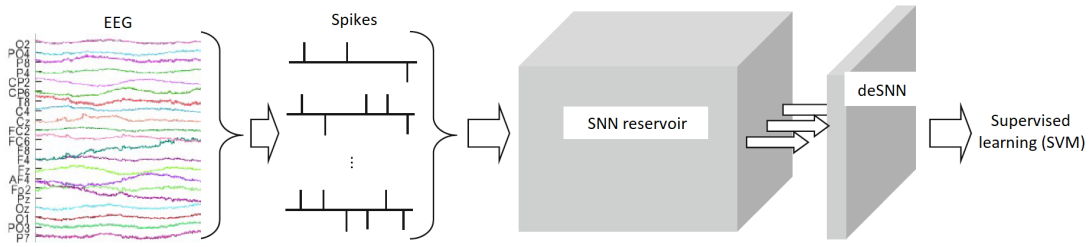


Figure 5.4: Illustration of the proposed SNN computational architecture for EEG-based emotion recognition.

5.5.1 Spike encoding

In an SNN-based architecture, information is processed in the form of binary spiking events. Accordingly, all continuous variables first need to be encoded into spike trains, shown as the first step in Figure 5.4. In this paper, by spike encoding, we namely focus on temporal spike encoding methods where spike timings usually mark changes in the signal value over time Petro et al. (2019). This approach is driven by a biologically plausible view that precise relative spike timing encodes information Bohte (2004). Most of the commonly used encoding algorithms N. Kasabov, Scott et al. (2016), such as threshold-based representation (TBR), step-forward (SF) encoding, moving-window (MW) encoding, or the Bens Spiker Algorithm (BSA) Nuntalid et al. (2011), rely on tracking the temporal changes in the signal, which then represent the exact timing of spikes. Such algorithms produce a bipolar spike sequence, where positive changes in signal value (increases) result with positive spikes, and negative changes in signal value (decreases) result with negative spikes.

In this paper, we transform the EEG data into spike trains using the relatively simple version of TBR, known as the address event representation (AER) method Delbruck (2007); Lichtsteiner and Delbruck (2005). The method is based on thresholding the rate of change of an input variable over time and is suitable when the input data is a stream, which is the case with EEG. The algorithm is based on the variable threshold value that is calculated for each of the 32 input data channels. The variable threshold array is calculated for each channel, as the signal dynamics and value ranges can vary between the input channels. For each of the input channels, the variable threshold is calculated based on one scalar input parameter (α_{TR}) in the following way:

$$VT(k) = \frac{1}{N}\sum_{i=1}^{N}\left(\mu + \sigma \cdot \alpha_{TR}\right) \tag{5.1}$$

where N is the number of samples, T is the signal length (number of time points per data sample), and k goes from 1 to the number of channels $N_{\text{input}} = 32$. α_{TR} is the spike threshold parameter, μ is the sample mean rate of change in the signal, σ is the sample standard deviation of the rate of change in the signal, and $\boldsymbol{VT}$ is the resulting variable threshold array. Equation (5.1) represents the threshold value for the k-th input channel. At each time point where the k-th input channel signal difference (rate of change) exceeds the corresponding variable threshold, a positive spike is generated. Accordingly, inhibiting, i.e. negative, spikes are generated when the rate of change exceeds the variable threshold in the negative, i.e. decreasing, direction. Algorithm 5.1 sums up the spike encoding procedure in a compact algorithmic form.

In affective applications, such as the subject-independent emotion recognition, the encoding algorithm should ideally provide similar spike trains for similar emotional states, regardless of the high inter-subject variability in the collected EEG signal properties. Therefore, the encoding procedure described in algorithm 5.1 is utilised in a per-subject fashion, with μ and σ being estimated separately for each subject.

Algorithm 5.1 Spike encoding: $f_{encode} : \mathbb{R}^{T \times N_{input}} \rightarrow \{-1, 0, 1\}^{T \times N_{input}}$

Require: $X_{in} \in \mathbb{R}^{T \times N_{input}}, \{hyperparameters := \alpha_{TR}\}$
Ensure: $X_{out} \in \{-1, 0, 1\}^{T \times N_{input}}$
$N \leftarrow \#(X_{in})$ {number of data samples in the dataset}
for $k = 1$ to N_{input} **do**
  $VT_k \leftarrow 0$
  for $i = 1$ to N **do**
    $x \leftarrow$ channel k of the i-th sample in X_{in}
    $x' \leftarrow |\delta x|$
    $\mu \leftarrow mean(x')$
    $\sigma \leftarrow st.dev.(x')$
    $VT_k \leftarrow VT_k + (\mu + \sigma \cdot \alpha_{TR})$
  end for
  $VT_k \leftarrow \frac{VT_k}{N}$
end for
for $k = 1$ to N_{input} **do**
  for $i = 1$ to N **do**
    $x \leftarrow$ channel k of the i-th sample in X_{in}
    $x' \leftarrow \delta x$
    $x_{out} \leftarrow 0^{T \times 1}$
    for $j = 2$ to T **do**
      if $x'_j > VT_k$ **then**
        $x_{out(j)} \leftarrow 1$
      else if $x'_j < -VT_k$ **then**
        $x_{out(j)} \leftarrow -1$
      end if
    end for
    store spike train x_{out} in X_{out} for channel k, sample i
  end for
end for

Optimisation of the spike encoding method

Spike encoding is the first link in the SNN processing chain. Choosing the appropriate spike encoding method, with the optimal hyperparameters, in our case α_{TR}, is extremely important, in order to retain the task-relevant information. The spike encoding can therefore be seen both as a primary information compression as well as a noise elimination step. Inadequate spike encoding algorithm can result with either loss of useful information on one end (high α_{TR}), or high levels of information noise on the other end (low α_{TR}), as shown in Figure 5.5. The problem of selection and optimisation of temporal spike encoding algorithms for SNNs has been most

recently tackled in Petro et al. (2019). An exhaustive approach should evaluate the effectiveness of the encoding within the context of the entire SNN processing and classification framework, i.e. the encoding algorithm is optimised according to the output of the whole SNN system, e.g. classification accuracy Sengupta and Kasabov (2017). The computational cost of such an approach would be extremely high, due to the need for simultaneous optimisation of all SNN system hyperparameters, thus making the approach infeasible. Another approach is to try and optimise the encoding step by itself, via application of the corresponding decoding algorithm and trying to minimise some error metric between the original and reconstructed signal, as in Petro et al. (2019). This approach neglects the classification task context, thus lacking final validation in terms of the obtained classification performance.

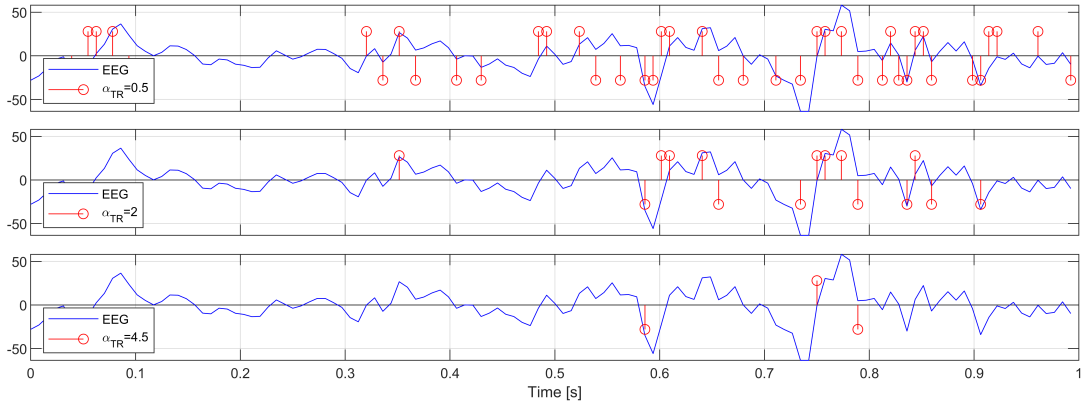


Figure 5.5: Spike trains resulting from different AER threshold parameter (α_{TR}) values. Lower α_{TR} yields a dense spike train with a tendency to encode changes which are most likely on the level of noise, while high α_{TR} value yields a sparse spike train with only major changes in signal value being encoded as spike events.

In this paper, we propose a compromise approach: we optimise the encoding method by itself, due to the already stated computational infeasibility of taking into account the rest of the SNN system, while we still take into account the specific classification problem, i.e. binary emotion classification in terms of valence and arousal. In order to avoid evaluating the performance of the entire SNN system, we calculate an array of simple spike-based features based on statistical descriptives and the widely used rate coding scheme Gautrais and Thorpe (1998). For each of the 32

EEG channels, encoded into spike trains as described in the previous subsection, we calculate the following 6 features:

- firing rate of positive spikes,
- firing rate of negative spikes,
- median timing difference between consecutive positive spikes,
- median timing difference between consecutive negative spikes,
- interquartile range of timing differences between consecutive positive spikes,
- interquartile range of timing differences between consecutive negative spikes.

We use a simple definition of firing rate as the temporal average, i.e. the number of spikes divided by the corresponding time interval duration (in our case 1 second for all data samples). Each sample is therefore described by $32 \cdot 6 = 192$ simple rate-coding-based and statistical features. Datasets are formed for an array of different α_{TR} values (ranging from 0.5 to 5, with a step of 0.5), standardised, and then tested for their task-relevant discriminative power. The used performance metric is the cross-validated classification error obtained with an optimised Inf-FS feature selection algorithm Roffo (2016); Roffo, Melzi and Cristani (2015) and an RBF-SVM classification algorithm. This metric employs a simple and generic feature selection and machine learning approach in order to estimate the maximum cross-validated nonlinear class separation that can be obtained by the calculated features. For each α_{TR} value, a non-convex Bayesian optimisation approach was employed in order to find the values of the remaining hyperparameter values:

- α_{FS}: Inf-FS parameter representing the trade-off between feature dispersion and feature correlation, which are the basis of the Inf-FS feature ranking algorithm (range set to $[0, 1]$),
- n_{FS}: number of selected features, based on the obtained feature ranking (range set to $[1, 50]$, integer grid),

- C: RBF-SVM parameter (range set to $[10^{-3}, 10^{3}]$, logarithmic grid),
- σ: RBF-SVM parameter (range set to $[10^{-3}, 10^{3}]$, logarithmic grid),

which maximize the leave-one-subject-out (LOSO) cross-validation accuracy. Matlab function *bayesopt* was used, with 500 objective evaluations and a 0.5 exploration ratio. Algorithm 5.2 sums up the spike encoding optimisation procedure in a compact algorithmic form.

Algorithm 5.2 Optimisation of the spike encoding threshold parameter α_{TR}

Require: $X_{in} \in \mathbb{R}^{T \times N_{input}}, y \in \mathbb{R}^{N_{samples} \times 1}$
Ensure: $\alpha_{TR,opt.}$
$\alpha_{TR,range} \leftarrow \{0.5k, k \in \{1, 2, ..., 10\}\}$ {search range for α_{TR}}
$N_{search} \leftarrow \#(\alpha_{TR,range})$ {from the above: $\#(\alpha_{TR,range}) = 10$}
$CV_{loss} \leftarrow 0^{N_{search} \times 1}$ {initialise objective array}
$\alpha_{FS,range} \leftarrow [0, 1]$ {search range for α_{FS}}
$n_{FS,range} \leftarrow [1, 50]$ {search range for n_{FS}, integer}
$C_{range} \leftarrow [10^{-3}, 10^{3}]$ {search range for C, logarithmic grid}
$\sigma_{range} \leftarrow [10^{-3}, 10^{3}]$ {search range for σ, logarithmic grid}
for $i = 1$ to N_{search} **do**
    $X \leftarrow f_{encode}(X_{in}, \alpha_{TR,range}(i))$ {spike encoding, see algorithm 5.1}
    calculate $X_{features} \in \mathbb{R}^{N_{samples} \times 192}$ based on X {features described in section 5.5.1}
    $CV_{loss}(i) \leftarrow Bayesopt(X_{features}, y, \alpha_{FS,range}, n_{FS,range}, C_{range}, \sigma_{range})$ {cross-validated classification error, other hyperparameters described in 5.5.1}
end for
$i_{min} \leftarrow$ index of smallest element in CV_{loss}
$\alpha_{TR,opt.} \leftarrow \alpha_{TR,range}(i_{min})$

Algorithm 5.2 has been applied to the tasks of arousal and valence classification, on both DEAP and MAHNOB-HCI, and the results of the spike encoding optimisation are shown in Figure 5.6. Threshold value $\alpha_{TR} = 1.5$ yields the most informative spikes for both arousal and valence classification. Such a result can be interpreted in the context of the behaviour of the encoding algorithm illustrated in Figure 5.5.

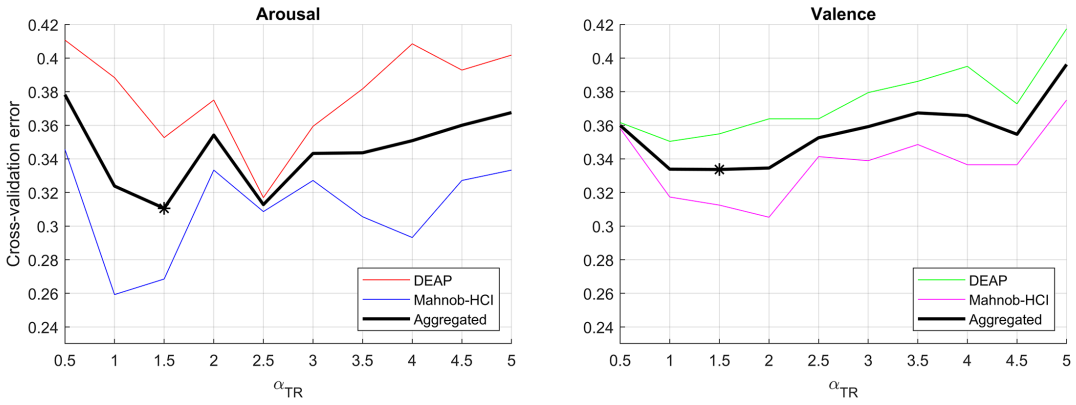


Figure 5.6: Class separation performance metrics for different α_{TR} values, using both datasets: arousal classification (left) and valence classification (right). Optimal α_{TR} value is indicated by an asterisk.

5.5.2 Spiking neural network processing

3D SNNr module

Once the EEG data has been encoded for each subject, the obtained sparse spike trains are used to train a 3D SNNr with $N_r = 1471$ leaky integrate and fire (LIF) model neurons. Each neuron has a predefined 3D spatial coordinate, according to the Talairach template coordinates from Talairach and Tournoux (1988), resulting with a brain-like shape. Accordingly, input spikes are passed over to the SNNr at the neuron locations corresponding to the mapping of 32 EEG channels, as shown in Figure 5.7.

The connections between the neurons in the SNNr are initialised using the small-world connectivity (SWC) approach N. Kasabov and Capecci (2015), where a radius is defined as a parameter for connecting neurons within this radius, i.e. small-world radius (SWR). This results in an SNNr of sparsely connected neurons. After initialisation, the connection weights $W_{(ij)}$ between the pairs of connected neurons (ij) are determined based on the following expression:

$$W_{(ij)} = \text{sgn}\Big(\text{rand}(1) - 0.2\Big) \cdot \text{rand}(1) \cdot \frac{1}{L_{\text{dist}(ij)}}, \tag{5.2}$$

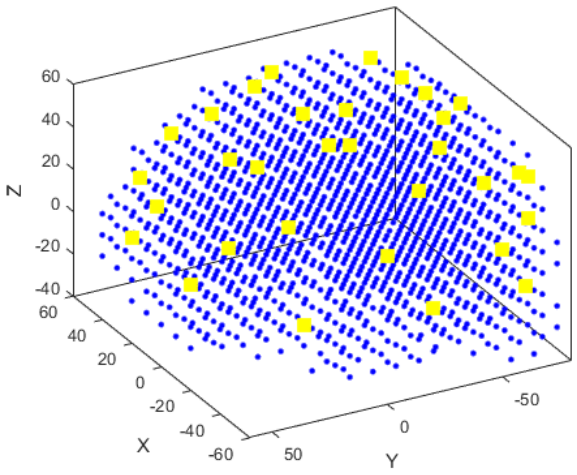


Figure 5.7: 3D SNNr structure, according to the Talairach template coordinates. Yellow neurons are considered input neurons, and correspond to the mapping of 32 EEG channels.

Where rand(1) generates a pseudorandom from a uniform distribution on the open interval $(0, 1)$, and $L_{\text{dist}(ij)}$ represents the distance between neurons i and j. The equation above results with an expectation of 80% positive weights and 20% negative weights (in matrix $\boldsymbol{W}$).

The STDP learning rule S. Song et al. (2000) is applied that allows the SNNr to adapt the connection weights based on the Spatio-temporal associations between the input-driven spikes. The used STDP algorithm relies on the following hyperparameters:

- D (potential leak rate): the rate of potential passive degradation through inactivity,
- R (refractory time): determining a period of resting between spikes,
- η (STDP rate): learning rate, used for weight updating,
- β (firing threshold): potential threshold for generating a spike,
- N_{iter}: number of training iterations,

and with the specific steps described in Algorithm 5.3.

Algorithm 5.3 Unsupervised SNNr weight learning: f_{STDP}

Require: $C_{init} \in \{0,1\}^{N_r \times N_r}, W_{init} \in \mathbb{R}^{N_r \times N_r}, S_{in} \in \{-1,0,1\}^{T \times N_{input} \times N_{samples}}$
$\{hyperparameters := D, R, \eta, \beta, N_{iter}\}$
Ensure: $W_{out} : \mathbb{R}^{N_r \times N_r}$

$\chi \leftarrow [1,2,...,N_r]$ {all neuron indices}
$P_k \leftarrow 0, \forall k \in \chi$ {initialize neuron potentials}
$R_k \leftarrow 0, \forall k \in \chi$ {initialize neuron refractory time counters}
find $inputneuron$ indices $\iota \subset \chi$
for $n_{iter} = 1$ to N_{iter} **do**
  $\eta' \leftarrow \frac{\eta}{\sqrt{n_{iter}}}$
  for $i = 1$ to $N_{samples}$ **do**
    $s \leftarrow T \times N_{input}$ spike matrix of the i-th sample in S_{in}
    for $t = 1$ to T **do**
      find $firingneuron$ indices $\tau = \{firingneurons \text{ in } \iota\} \cup \{k \in \chi \setminus \iota, P_k > \beta\}$
      for all $j \in \tau$ **do**
        find post synaptic neuron indices γ
        for all $k \in \gamma$ and $R_k = 0$ **do**
          $P_k \leftarrow P_k + w_{jk}$ {update potential}
        end for
      end for
      $P_k \leftarrow 0, \forall k \in \tau$ {reset potential}
      $R_k \leftarrow R, \forall k \in \tau$ {reset refractory counter}
      $P_k \leftarrow max(0, P_k - D), \forall k \in \chi \setminus \iota \setminus \tau$
      $R_k \leftarrow max(0, R_k - 1), \forall k \in \chi \setminus \iota \setminus \tau$
      for all $j \in \tau$ **do**
        find post synaptic neuron indices γ
        for all $k \in \gamma$ **do**
          $w_{jk} \leftarrow w_{jk} - \eta(t - t_k^f)$
        end for
        find pre synaptic neuron indices γ
        for all $k \in \gamma$ **do**
          $w_{jk} \leftarrow w_{jk} + \eta(t - t_k^f)$
        end for
      end for
    end for
  end for
end for

deSNN representation

As an output representation of the SNNr-based spike sequences we used the deSNN algorithm N. Kasabov, Dhoble et al. (2013). The method combines the simple rank-order (RO) learning rule S. Thorpe and Gautrais (1998) and the STDP-based

activation of the neurons in the previously trained SNNr (as described in Algorithm 5.3). In the previous section, we have described how the spike trains can be used to train the SNNr in an unsupervised manner. However, the SNNr now operates as an activation module, meaning that the new input spike trains are propagated through the trained SNNr, generating a complex Spatio-temporal neuron activation pattern. This pattern is used to generate (evolve) the output neurons, i.e. the deSNN representation. The algorithm hyperparameters are:

- α_{deSNN}: main deSNN hyperparameter that determines the output value based on the first spike occurrence of the corresponding SNNr neuron,
- d(drift): used for output update on the subsequent spikes of the corresponding SNNr neurons.

Algorithm 5.4 Output deSNN representation: f_{deSNN}

Require: $W \in \mathbb{R}^{N_r \times N_r}, s_{in} \in \{-1, 0, 1\}^{T \times N_{input}}$
$\{hyperparameters := \alpha, d\}$
Ensure: $W_{deSNN} \in \mathbb{R}^{1 \times N_r}$
$W_{deSNN} \leftarrow 0^{1 \times N_r}$ {initialize representation to 0}
$F \leftarrow 0^{1 \times N_r}$ {initialize the neuron firing flags}
$c \leftarrow 0$ {initialize the neuron firing order counter}
$S_{cube} \leftarrow$ propagate s_{in} through the SNN defined by W {S_{cube} is a sparse $N_r \times T$ matrix of all neuron firings for the entire sample duration length T}
for $i = 1$ to T **do**
  for $j = 1$ to N_r **do**
    if $S_{cube}(i, j) = 1$ **then**
      if $F(j) = 0$ **then**
        {neuron fires for the first time}
        $W_{deSNN}(j) \leftarrow \alpha^c$
        $F(j) \leftarrow 1$
      else
        $W_{deSNN}(j) \leftarrow W_{deSNN}(j) + d$
      end if
      $c \leftarrow c + 1$
    else
      $W_{deSNN}(j) \leftarrow W_{deSNN}(j) - d$
    end if
  end for
end for

From the deSNN procedure, described in Algorithm 5.4, it can be seen that the static output representation of the Spatio-temporal SNNr-based spiking patterns highly depends on the selection of the method's hyperparameters α_{deSNN} and d.

5.6 Experiments

In N. Kasabov and Capecci (2015) it has been suggested to repeat the processing steps described in sections 5.5.1, 5.5.2, and 5.5.2 for different hyperparameter values in order to optimise the final classification performance in the supervised learning step. Taking into account the entire proposed methodology, this includes the following hyperparameters: α_{TR} for spike encoding; SWR, D, R, η, β and N_{iter} for STDP-based unsupervised learning of the SNNr; α_{deSNN} and d for the deSNN representation; and finally α_{FS}, n_{FS}, C and σ for the supervised learning step, which includes feature selection and an RBF-SVM classifier. This makes a total of 13 hyperparameters that should be simultaneously optimised in an exhaustive procedure where each iteration includes data preprocessing, STDP-based learning as well as cross-validation of the obtained output representation. However, STDP-based training of SNNs is computationally very expensive and accounts for the most significant proportion of the total execution time.

In order to mitigate the effects of the computationally expensive STDP-based step on the duration of the hyperparameter optimisation procedure, we optimise the encoding hyperparameter α_{TR} via a compromise approach described in section 5.5.1, outside of the SNN context, while still taking into account the classification task performance. Parameter $\alpha_{TR} = 1.5$ is identified as optimal for both the arousal and valence classification tasks. For the STDP-based training of the SNNr module, we use a set of previously identified hyperparameters Capecci et al. (2016): $SWR = 2.5$, $D = 0.002$, $R = 6$, $\eta = .001$, and $\beta = 0.5$, in order to avoid training the SNNr at each step of the deSNN and supervised learning hyperparameter optimisation. N_{iter} was set to 5, due to the relatively small amount of available EEG data. This parameter

has to be taken with caution as high values may cause over-training of the SNNr N. Kasabov and Capecci (2015). The resulting brain-like SNNr connectivity can be visualised, analysed and interpreted for a better understanding of the EEG data and relative involvement of various brain regions. Figure 5.8 shows the emerging SNNr connectivity, obtained by using data with different emotional labels.

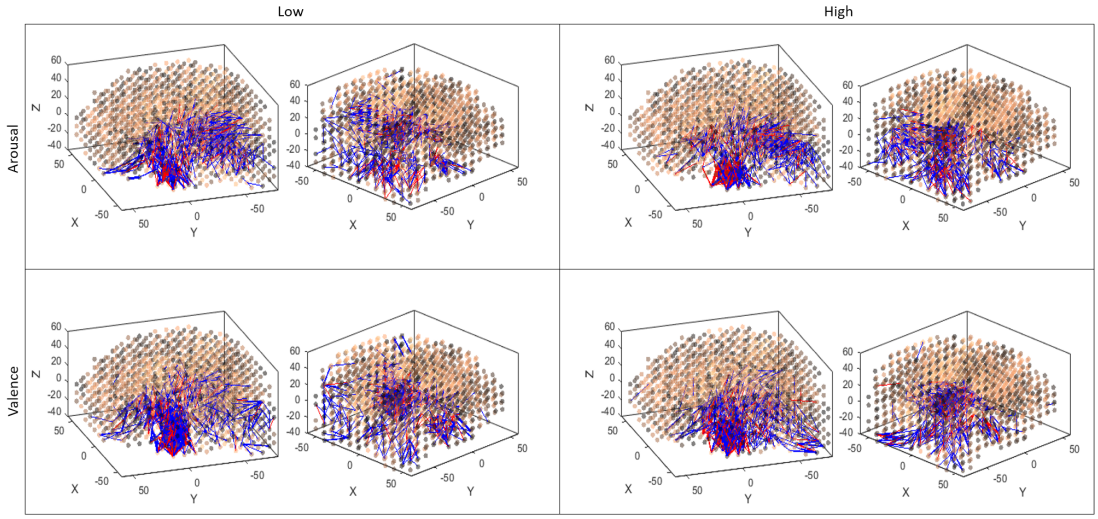


Figure 5.8: Connectivity of 4 different SNNr modules. Each SNNr is obtained by using the combined data from both DEAP and MAHNOB-HCI, labeled as either low or high in terms of either arousal or valence. For each SNNr two 3-D plots from different angles are given: the left plot shows frontal and left side of the brain, while the right plot shows the back (posterior) and right side of the brain. Neurons are plotted with a slight transparency in order to better highlight the 3D nature of the emerged connections. 500 strongest connections are displayed for each SNNr, with thicker lines denoting stronger connections. Brighter neurons are more active.

For each of the 4 available subtasks, i.e. arousal and valence classification with DEAP and MAHNOB-HCI, a Bayesian optimisation approach is utilised in order to identify the optimal set of the 6 remaining hyperparameters. α_{FS}, n_{FS}, C and σ are optimised using the same ranges as in section 5.5.1. The range for α_{deSNN} was set to $[10^{-2}, 2]$, and range for d was set to $[10^{-3}, 1]$. At each iteration both simple spike-based features from section 5.5.1, as well as the deSNN-based features, were fed to the feature selection and LOSO cross-validated RBF-SVM evaluation, in search of the maximum accuracy. The idea was to test the added value of deSNN-based

features in terms of classification accuracy improvement. Results are summed up in Table 5.1.

Table 5.1: Comparison of the obtained optimized LOSO cross-validation accuracies.

Feature set	Task	Dataset	Accuracy
Simple spike-bassed features	Arousal	DEAP	0.6384
		MAHNOB-HCI	0.7593
	Valence	DEAP	0.6473
		MAHNOB-HCI	**0.7212**
Simple spike-bassed features + deSNN	Arousal	DEAP	**0.7897**
		MAHNOB-HCI	**0.7939**
	Valence	DEAP	**0.6776**
		MAHNOB-HCI	0.7068

The upper half of Table 5.1 is based on the results obtained in section 5.5.1 (Figure 5.6), while the lower half sums up the results of deSNN optimisation procedure described in this section. The addition of optimised deSNN representation to the feature set significantly increased the maximal obtainable accuracy in arousal classification, for both datasets. In terms of valence classification, accuracy was slightly improved for DEAP, but not for MAHNOB- HCI.

5.7 Discussion

The obtained results demonstrate how complex SNN-based Spatio-temporal modelling of EEG spiking patterns can provide additional information value able for short-term emotion classification. Such SNN-based information was shown to be particularly useful for short-term arousal classification, and not as much for valence.

The main advantage of the described procedure is straight-forward spike-based processing of EEG, as opposed to the need for extraction of handcrafted features in case of the most classical methods, as well as the ability to capture information hidden within the complex Spatio-temporal EEG spiking patterns on relatively short time frames. Furthermore, the addition of subject-specific spike encoding, which

is crucial in order to successfully develop subject-independent SNN-based decision systems, as well as the proposed hyperparameter optimisation strategy is an added value to the original NeuCube framework. Hopefully, this will open up additional paths and opportunities for further utilisation of brain-like SNN models in various subject-independent classification tasks.

From the spike-only-based ”comparison approach” (Table 5.1) emerge both the strengths and limitations of this study. This approach demonstrates that the addition of deSNN features, which are a representation of spatio-temporal SNNr spiking patterns, to simple non-SNN-related spike features improves the classification performance. In contrast to traditional approaches, an SNN-based emotion classification framework presented in this paper can help in better understanding of short-term emotional processing. From Figure 5.8, it can be seen that the left side of the brain dominates in terms of connectivity strength, i.e. activation, in particular regions corresponding to the temporal lobe, amygdala and the occipital lobe. This result is not surprising, since both DEAP and MAHNOB-HCI induce emotions by means of emotional videos, which requires visual processing (occipital lobe) as well as the processing of language (temporal lobe). Additionally, it is well known that the amygdala, located in the medial temporal lobe, plays one of the key roles in general emotional processing Weymar and Schwabe (2016). The dominance of the left side most likely emerges from the dominance of the left temporal lobe in most people. For example, the lateral sulcus, area considered to be highly involved in language function, is longer on the left than on the right side of the brain Yeni-Komshian and Benson (1976). Additionally, the posterior views on the obtained connectivities from Figure 5.8 suggest that occipital lobe plays the most significant role in the processing of high arousal. This might indicate that the visual portion of the affective stimuli plays the most significant role in inducing high emotional arousal. Due to the indirect presence of the EEG data from all subjects in both the training and validation, as a result of a single unsupervised SNNr training for each

of the 4 subtasks, the reported accuracies likely overestimate the true predictive power of framework, and as such call for future work, with bigger amounts of data. Given that, as well as the hyperparameter optimisation strategy described in section 5.6, Table 5.1 should rather serve as a demonstration of ”added value” in comparison to the classical non-SNN- related spike-based features. Additionally, analysis of the deSNN features in the context of the traditional time- or frequency-domain EEG features would be useful but exceeds the scope of our work as such features do not match the idea of biologically inspired spike-based processing of EEG. However, the proposed approach should not be considered only in terms of the obtained classification performance, which is the case in most traditional ”black box” emotion recognition models. One of the main advantages of this approach is in the understanding of the induced and recognised emotions through a brain-inspired SNN model, as shown in Figure 5.8 and discussed above.

5.8 Conclusion

To conclude, this paper for the first time introduces a brain-inspired SNN architecture to recognise, and most importantly, to explain for a better understanding, EEG data measuring two basic dimensions of emotion: arousal and valence. The patterns obtained after deep learning in the SNN architecture are interpreted in terms of brain activities of subjects. The proposed method can manifest fast, incremental and transfer learning on new data related to emotions that make it suitable for further study and for real-time emotion recognition systems. Future work will involve improved SNN hyperparameter optimisation and better visualisation of the brain- structured SNN during and after learning for a better understanding of brain processes related to emotional processing in humans we believe that EEG technology will still remain a modality of choice in highly specialized emotion recognition tasks which are safety-critical or health-related, and as such require a comprehensive emotion recognition approach which would include the EEG as well. Examples of

such systems are EEG-based music therapy (Y. Liu, Sourina & Nguyen, 2011) or in safety-critical occupations were the interaction between human cognition and emotion is very important, like air traffic control (Borghini et al., 2017). In terms of temporal computation, spiking neural networks are inherently characterized by lower complexity compared to traditional neural networks, due to the sparse nature of SNNs and with computation at each time step being reduced just to the activated neurons and connections, rather than all connections at every timestep – which is the case with traditional deep neural networks.

6

FusionSense: Emotion Classification using Feature Fusion of Multimodal Data and Deep learning in a Brain-inspired Spiking Neural Network

6.1 Prelude to Chapter 6 Manuscript

Affective computing aims at making human-machine interactions more organic by allowing machines to recognise human emotions and respond accordingly. Over the years, affective computing has found applications in fields ranging from education to traffic safety. Neural networks which emulate the behaviours of biological neural networks have been critical in facilitating affective computing. The first generation of the same involved relatively simple weighted sum method-based neurons. Later, they were replaced by Sigmoid neurons which used smooth differentiable sigmoid function and allowed for stacking to generate deep neural networks. While these artificial neural networks (ANNs) were more successful than their first-generation counterparts, they could not emulate the nonlinearity associated with biological neurons that work on spikes or impulses. This then led to the formation of spiked neural networks or SNNs, which better mimic biological neurons. These have the capability to analyse spatial and temporal data such as those generated by EEGs, making them better suited for mimicking human behaviour. This specific paper deals with utilising a structure termed as NeuCube, which essentially is a

3-dimensional SNN structure. In this work, NeuCubes are used to build a Facial Emotion Recognition (FER) system.

Once the overall objective is identified, the paper proceeds to give a brief literature review on the existing modalities that serve as input for affective computing. These include individual aspects such as facial expressions, speech, body language, and physiological signals such as ECG, EEG, etc. In earlier literature, most neural networks considered any one of these Affects as input or carried out unimodal affect processing. However, a more recent trend is to consider multiple Affects simultaneously to give better, more accurate results. This multimodal affect recognition is also discussed in the paper, highlighting the two possible ways in which multimodal analysis can be carried out. The various Affects can be combiner at the feature level (feature level or new level fusion), and the combined single vector can then be used by the model. Alternatively, the model can analyse each input individually before combining to form a combined decision vector, and this approach is termed as Decision level fusion. The next part of the paper explores spiked neural networks in great detail and introduces evolving SNNs or eSNNs.. These fast learning eSNNs can further be improved by constituting them as dynamic evolving SNNs or deSNNs, which can further improve the way the neural network processes Spatio-temporal data. However, these neural systems are not sufficient to emulate a human brain, as locations of neurons and their connectivity are critical aspects of biological neural systems. Therefore, a 3D NeuCube of SNNs was proposed to mimic the human brain better. The paper also discusses various studies on NeuCubes before proceeding to the methods. The MANHOB-HIC database is employed for this study to detect facial affects. Facial features are then extracted, and physiological indicators such as Heart rate variability (HRV), respiration variability, respiration depth, skin temperature, GSR and pupil diameter are used as physiological features obtained from observing the video. The NeuCube then uses all this data (both facial affect data and physiological affect data). It should be noted that the NeuCube here is an 11*11*7 array of

neurons. The SNNr was made with leaky integrate and fire model (LIFM) spiking neurons with recurrent connections. In such a neuron configuration, the PSP of postsynaptic potential rises or falls with each unique input from the presynaptic neurons. The PSP can leak between spikes and has a time constant equal to 0.02 sec in this study for exponential decay.

These structures are first subjected to unsupervised training, where learning is centred on spike-time-dependent plasticity (STDP). The next step is supervised training where for each input and output neuron is generated and linked to all existing neurons in the system. The weightage for each connection is determined using rank-order method initially and are readjusted with each new generation of neurons. During the classification stage, the NeuCube was exposed to validation data, and with each iteration, the weight of each neural connection was optimised. The leave one subject out mode (LOSO) was employed for the same. The final stage was testing the NeuCube using test data. The network was then tested and exhibited an accuracy of 83.7% for classifying the training data, and nearly 81% for peripheral data. The accuracy on MANHOB-HIC dataset was 73.15% which is promising considering that this work pioneers multimodal analysis using SNNs. The study then moves on to discuss related work in the field and highlights the shortcomings of the work presented, such as not utilising EEG data or voice data at this stage of research.

In summary, the work presented introduces the possibility of using SNNs for multimodal emotion recognition. These third-generation neural networks are constituted as a three-dimensional array known as NeuCube for handling spatiotemporal data like a biological neural network. Using the MAHNOB-HCI dataset, the multimodal data consisting of facial expressions and physiological data (ECH, skin temperature, skin conductance, respiration signal, mouth length and pupil size) are used in tandem for identifying the emotional affect. This novel approach gives an accuracy of over 73% on the MAHNOB-HCI dataset, highlighting the potential for

application of SNNs in emotion recognition using multimodal data. It also raises the possibility that more modalities such as ECG data or voice data can further be incorporated into such systems, to raise accuracy.

6.1.1 Contributions and Publications

Contributions

1. *Application of evolving Spiking Neural Networks on Multimodal emotion recognition.*
2. *Employs a multimodal approach combining facial expression data and physiological data in tandem for identifying the emotional affect.*
3. *The novel approach gives an accuracy of over 73% on the MAHNOB-HCI dataset, highlighting the potential for application of SNNs in emotion recognition using multimodal data.*
4. *Addition of more modalities such as ECG data or voice data can further be incorporated into such systems, to raise accuracy.*

Publications

1. Tan, C., Ceballos, G., Kasabov, N., & Subramaniyam, N. (2020). FusionSense: Emotion Classification using Feature Fusion of Multimodal Data and Deep learning in a Brain-inspired Spiking Neural Network. *Sensors*, 20(18), [5328]. https://doi.org/10.3390/s20185328

6.2 Introduction

The central aim of affective computing is to enable seamless communication between humans and computers by developing systems that can detect and respond to the various affect states of the humans Calvo and D'Mello (2010). Affective computing is an interdisciplinary field of research that involves experts from computer science, psychology, social, and cognitive sciences. Affect recognition has important applications in several fields, such as medicine Edwards et al. (2002), driver fatigue monitoring, human-computer interaction, sociable robotics Fong et al. (2003), and security systems, to name a few.

Modelling affect can be classified into three categories: categorical, dimensional, and components. Categorical models classify emotions into a set of discrete classes, which are easy to describe and these include six basic emotions, such as happiness, sadness, fear, anger, disgust, and surprise. Owing to its simplicity, categorical models have been extensively utilized in affect research. In contrast, dimensional models represent emotion as a point in multidimensional space, where the dimensions include valence, activation, and control, allowing for the description of more complex and subtle emotions. However, such multidimensional space can pose a significant challenge to automatic emotion recognition system and, thus, researchers have mostly used the simplified two-dimensional model of arousal and valence proposed in Russell (1980), where arousal ranges the intensity of emotion from calm to excited, and valence ranges from unpleasant to pleasant Gunes et al. (2011). Finally, the component model of emotions arrange emotions in a hierarchical fashion, where complex emotions can be derived from the combination of a pair of basic emotions. The most popular component model proposed by Plutchik Plutchik (2001) is based on evolutionary principles and it has eight basic bipolar emotions.

Affect can be expressed via facial expression, body movements, voice behavior, gestures, and an array of physiological signals, such as heart rate, sweat, pupil diameter, brain signals, to mention a few. The problem of recognizing emotions by utilizing facial expressions from videos and static images have been addressed by several studies Danelakis et al. (2015); Poria et al. (2015); Yeasin et al. (2006). Advances in deep learning methodologies have created huge interest in application of such methods in facial emotion recognition (FER) Fan et al. (2016); Gudi et al. (2015); Ionescu et al. (2013); Kahou et al. (2013); Tang (2013), most of which are based on supervised learning. The methods do not allow for incremental, adaptive learning on new data, and they are not suitable for on-line applications. For an excellent overview of the application of deep learning and as well as shallow learning approaches to FER, the reader is directed to S. Li and Deng (2018) and the references

there in. Additionally, the reader can refer to this Chapter on Multimodal Affect Recognition in the Context of Human-Computer Interaction Schwenker et al. (2017).

Spiking neural networks (SNNs) represent the third-generation of neural networks, modelling neurons and interactions between them in a biologically more realistic manner as compared to second-generation neural networks based on ANNs. SNNs are an ideal choice to handle the emotion recognition task from video data, given their ability to effectively handle spatio-temporal data Dhoble et al. (2012) (see Section 6.5 for details).

In this work, we propose building an emotion recognition system for multimodal data. system using SNNs. To this end, we use the NeuCube framework N. Kasabov (2014), which is a type of evolving SNN (eSNN). In this paper, we develop an encoding method to map the continuous facial feature values to spikes based on population coding. We use the data from Mahnob-HCI dataset to test the NeuCube framework for the classification of binary valence in response to video stimuli.

The structure of the paper is organized, as follows. In Section 6.3, we provide some background literature on various data modalities used in affect detection, where, as in Section 6.4, we describe strategies for multimodal data fusion. In Section 6.5 we provide some background on SNN and the NeuCube framework. Section 6.6 details the methodology used in our work and Section 6.7 presents the results. In Section 6.8, we discuss our results and, in Section 6.9, the direction for future work is presented and it concludes the paper.

6.3 Signals for Affect Detection

6.3.1 Facial Expression

One of the immediate and natural ways for humans to communicate their emotions is through facial expressions, which constitute about 55% of the information communicated during face to face human interaction Mehrabian (1968). Thus,

affect research has primarily focused on detecting emotions from the face. Research on facial emotions have shown that the six basic emotions, such as fear, anger, sadness, enjoyment and disgust can be detected with facial expressions Ekman (1992a, 1992b) and dectecting an emotion is equivalent to detetecting the associated prototypic facial expression. Based on the Facial Action Coding System (FACS), which originally described 44 single action units (AU) including head and eye movements, with each action unit linked with an independent motion on the face and the correponding muscles, for example lip suck motion with the muscle orbicularis oris Ekman and Friesen (1976). Several deep learning techniques have been used to build automatic facial emotion recognition (FER) system, including deep boltzmann machine (DBM), deep belief networks (DBNs) El Kaliouby and Robinson (2005); P. Liu, Han, Meng and Tong (2014); Uddin et al. (2017), convolutional neural networks (CNNs) Breuer and Kimmel (2017); Gudi et al. (2015); H. Jung, Lee, Yim, Park and Kim (2015); Ng, Nguyen, Vonikakis and Winkler (2015); Zhao, Chu and Zhang (2016), auto-encoders Rifai, Bengio, Courville, Vincent and Mirza (2012); Sun, Zhao and Jin (2017); Zeng et al. (2018), and recurrent neural networks (RNNs), to mention a few.

6.3.2 Speech

Affective information from speech can contain lingusistic and paralinguistic features, which refer to what is said and how it is said, respectively. Although speech is a fast and efficient method of communication that can be exploited in affect research, detecting the emotional state of the speaker using speech signal is still a significant challenge. There is no clarity on which features of the speech signal are most powerful is distinguishing different emotions. It has also been shown that, as compared to facial expressions, the accuracy of affect detection from speech is lower El Ayadi, Kamel and Karray (2011). For instance, the basic emotions, such as sadness, anger, and fear, can be recognized using speech, where as disgust is hard to detect Calvo

and D'Mello (2010). Moreover, cultural differences among speakers has not been addressed thoroughly with most of the affect research involving speech focusing on monolingual emotion classification El Ayadi et al. (2011). The features that are typically extracted from speech signal include both global and local features, Local features refer to pitch and energy extracted from small segments, into which a speech signal is typically divided to make it stationary, whereas global features refer to statistics of all the local features extracted from a long signal. Studies have shown that global features have better classification accuracy than local features Picard, Vyzas and Healey (2001); Shami and Kamel (2005); Ververidis and Kotropoulos (2005). However, studies have shown that global features cannot distinguish between emotions that have similar arousal Nwe, Foo and De Silva (2003) and may prove to be sub-optimal when using classifiers, such as Hidden Markov Model (HMM) and Support Vector Machines (SVM), due to insufficient number of training vectors El Ayadi et al. (2011). Because the properties of the different speech sounds can be altered by different emotions, some studies have also explored the benefits of phoneme-level modeling for the classification of emotional states from speech rather than using the prosodic features, such as pitch and energy Lee et al. (2004). Their results showed that the using phoneme-class classifiers outperformed HMM classifiers just based on global features. Apart from using HMM or SVM classifiers, several deep learning techniques have been explored for emotion recognition from speech signals including DBM Albornoz, Milone and Rufiner (2011); Z.-w. Huang, Xue and Mao (2015), auto-encoders Cibau, Albornoz and Rufiner (2013); Deng, Zhang, Marchi and Schuller (2013), DBNs C. Huang, Gong, Fu and Feng (2014); Wen, Li, Huang, Li and Xun (2017), and CNNs Badshah, Ahmad, Rahim and Baik (2017); Z. Huang, Dong, Mao and Zhan (2014); Trigeorgis et al. (2016), to cite a few. Despite the aforementioned challenges, speech is still an important signal that can be used for affect detection, as it is non-intruive and has high temporal resolution.

6.3.3 Posture and Body Movements

In comparison to speech and facial expression, perceiving emotions through body movements and postures is a relatively less explored topic in affect research. In fact, 95% of the literature in research on human emotions focuses on facial expressions and less than 5% on speech and other physiological signals with the remaining little of body movements. Several studies in the past have shown that body movements and postures can contribute to the recognition of emotional states De Meijer (1989); Walk and Homan (1984), with perhaps the most influential work in this topic dating back to the second half of 19th century by Charles Darwin Darwin and Prodger (1998). Body postures may offer certain advantages in affect detection given the multiple degrees of freedom human body possesses, which can aid in communication of emotions and subsequently affect detection, even at long distances, at which facial emotions are unreliable Coulson (2004), indicating that postures contain information not present in facial expressions. Another advantage of posture-based affect system could be that, in comparison to facial expression, which may be intentionally controlled, postures and body movements are unintentional and, thus, less susceptible to social editing Calvo and D'Mello (2010). In a study on deception by Eckman and Friesen Ekman and Friesen (1969), it was shown that liars were less successful at deception through body movements as compared to more controlled channels of communication, such as facial emotions, which they referred to as nonverbal leakage. Gestures, which can be defined as collection of body movements or actions involving head, hands, and other parts of the body allow the communication of a range of thoughts and emotions. Some of the basic gestures have been shown to be similar across the cultures. Given the advantages of this non-verbal communication channel, relatively few studies have utilized deep or machine learning framework to recognize emotions using body movements, postures, and gestures Kosti, Alvarez, Recasens and Lapedriza (2017); Saha, Datta, Konar and Janarthanan (2014).

6.3.4 Physiological Signals

Physiological signals such as electroencephalography (EEG), electrocardiogram (ECG), electromyogram (EMG), skin conductance, also known as Galvanic skin response (GSR), skin temperature, as well as pupilary diameter can be used for affect detection, apart from the above mentioned non-physiological signals. Physiological signals for affect detection are typically acquired in a non-invasive manner using wearable sensors. Heart rate (HR) and hear rate variability (HRV) can be derived from ECG signals. Skin temperatue has been shown to be a effective indicator of the emotional state as shown in Barnea and Shusterman (1995) and it primarily reflects the activity of the autonomic nervous system (ANS). Another modality that captures the activity of ANS is the GSR or skin conductance, which can be obtained by measuring the electrical potential on the skin after passing a negligible amont of current. GSR is considered to be a reliable indicator of arousal Nakasone, Prendinger and Ishizuka (2005), as it captures the activity of the sweat glands on the skin.

In affect research, ECG signals are typically recorded by a pair of electrodes, which are a subset of lead I configuration comprising of 12 electrodes. Features such as HR and HRV can be further derived from ECG that can reflect the activity of the sympathetic and parasympathetic branch of ANS system. HR and HRV have both been used in several studies to asses the mental states of the subject Healey and Picard (2005); Hjortskov et al. (2004). An EMG signal is reflective of the strength of muscle movements and is typically recorded by a pair of electrodes placed on the body. Studies have shown that when the subject is under some emotional stress, the changes in the facial expression can be measured using EMG activity Ekman, Friesen and Ellsworth (2013); Scheirer, Fernandez and Picard (1999). Apart from using electrodes on the face, other studies have also looked into measuring the activity of jaws or shoulders in order to identify emotional states Healey (2009).

Breathing is another physiological process that is shown to be altered by basic emotions, such as happiness, sadness, and anxiety Homma and Masaoka (2008).

Researchers have observed rapid breathing during arousal state Nykliček, Thayer and Van Doornen (1997) and as well as changes in respiratory pattern of subjects looking at photographs that induce emotions Homma and Masaoka (2008). The respiratory rate is shown to be modulated by emotions, particularly anxiety affecting the expiration rate Homma and Masaoka (2008), where timing and volumetric aspects of breathing are altered by various physical and mental stress Grossman and Wientjes (2001).

Finally, EEG is probably the most widely used physiological signal to study emotion. EEG is a low cost technology a compared to other neuroimaging modalities and has very good temporal resolution. EEG electrodes record the activity of a large number of synchronous neurons as potential difference on the scalp. Several studies have utilized EEG for emotion recognition Chanel, Kronegg, Grandjean and Pun (2006); Horlings, Datcu and Rothkrantz (2008); Zheng, Zhu, Peng and Lu (2014) and classification of emotional states of arousal, valence and dominance. In addition to EEG, pupilary diameter size is also an indication of emotional state, with several studies reporting that the size of the pupil discriminates during and after different kinds of emotional stimuli Granholm and Steinhauer (2004); Partala, Jokiniemi and Surakka (2000).

Several deep learning methodologies have been utilized for emotion recognition using physiological signals Cho, Bianchi-Berthouze and Julier (2017); Jia, Li, Li and Zhang (2014); T.-P. Jung, Sejnowski et al. (2018); Q. Zhang, Chen, Zhan, Yang and Xia (2017). The reader is directed to Calvo and D'Mello (2010) for an exhaustive list of literature.

6.4 Multimodal Affect Recognition

Although a majority of the machine learning and deep learning framework for affect recognition uses data from one modality, i.e., video or audio or EEG, recently there has been considerable interest in fusing data from the above mentioned modalities.

Multi-sensor data fusion can be highly advantageous in terms of improving the reliability and accuracy of affect detection and, furthermore, multimodal systems have shown to outperform unimodal system as discussed in D'mello and Kory (2015). Multimodal fusion involves combining data from many different types of sensors and such fusion can be primarily performed at two distinct levels, known as feature-level fusion and decision-level fusion.

6.4.1 Feature-Level Fusion

In the feature-level fusion approach (also known as early fusion), features that are derived from different modalities are combined into a single feature vector, on which a classifier can then be trained. It is well known that humans use and integrate multiple sensory cues during face-to-face interaction to detect affective states and is the fundamental idea behind feature-level fusion Pantic and Rothkrantz (2003). The main advantage of feature-level fusion is that correlation between multimodal features at an early stage can lead to better performance, requiring only one learning phase on the feature vector. Several studies have utilized this approach for affect research Poria et al. (2015); C. Sarkar, Bhatia, Agarwal and Li (2014); S. Wang, Zhu, Wu and Ji (2014). However, feature-level fusion also has several challenges. Because features obtained from different modalities can have different time-scales, achieving time synchronization to bring the features in same format can be difficult and computationally expensive. Additionally, given the large feature set that one obtains with feature-level fusion, the classification accuracy can be severly affected if the training dataset is limited. Furthermore, learning cross-correlation between the heterogenous features can prove to be difficult Atrey, Hossain, El Saddik and Kankanhalli (2010).

6.4.2 Decision-Level Fusion

In the decision-level fusion approach (also known as late fusion), first the decisions based on features derived from each modality is obtained separately. A fused decision vector is then obtained using the local decisions, which can be used to obtain the final decision or classification Atrey et al. (2010). The fundamental advantage of decision-level fusion over feature-level fusion is that the decisions all have the same format and, hence, can be fused easily, thus avoiding synchronization issues. Furthermore, using decision-level fusion allows for the application of optimal classifier or method suited for each modality, thus providing more flexibility when compared to feature-level fusion Poria, Cambria, Bajpai and Hussain (2017b). Several studies have utilized decision-level fusion for affect research Alam and Riccardi (2014); Cai and Xia (2015); Yamasaki et al. (2015) and it has been noted that researchers prefer decision-level fusion over feature-level fusion Poria, Cambria et al. (2017b).

6.5 Spiking Neural Networks

Human brains encode information via discrete events that are known as action potentials or spikes, following an all-or-none principle, where a neuron fires an action potential if the stimulus crosses a certain threshold, else it remains silent. Due to this binary nature of information representation, the human brain still outperforms the existing artificial neural networks (ANNs) in terms of both energy and efficiency LeCun et al. (2015); W. Wang et al. (2018). When compared to the traditional ANNs, spiking neural networks (SNNs) utilize a more biologically realistic model of neurons Taherkhani et al. (2019), thus further bridging the gap between neuroscience and learning algorithms. SNNs have shown the ability to integrate information from different dimensions, such as time, phase, frequency, as well as handle large volumes of data in an adaptive and self-organized manner Dhoble et al. (2012); N. Kasabov, Dhoble et al. (2013), making them particularly suitable to solve online spatio-temporal pattern recognition. SNNs have been shown

to be computationally more efficient than ANNs both theoretically Maass (1997a); Maass and Markram (2004) and in several real-world applications Bohte, Kok and La Poutre (2002). SNNs have been used in several real-world learning tasks such as unsupervised classification of non-globular clusters Bohte, La Poutré and Kok (2002), image segmentation and edge detection Meftah et al. (2010), epileptic seizure detection with EEG Ghosh-Dastidar and Adeli (2007). Furthermore, Bohte and colleagues devised a supervised learning rule for the SNNs and demonstrated its application in the XOR classification problem and several other benchmark datasets Bohte, Kok and La Poutre (2002). The evolving SNN (eSNN) is a class of SNN that utilizes rank order learning S. Thorpe and Gautrais (1998) and was first proposed in N. K. Kasabov (2007). The eSNN handles spatio-temporal data by increasing the number of spiking neurons in time to learn temporal patterns from data Wysoski et al. (2010). In addition to the open evolving structure of eSNNs that facilitates the addition of new variables and neuronal connections, eSNN have the advantage of fast learning from large amounts of data and they can interact with other systems actively. eSNNs also allows for the integration of various learning rules, such as supervised learning, unsupervised learning, fuzzy rule insertion, and extraction, to mention a few and they are self-evaluating in terms of system performance. These aforementioned properties constitute the evolving connectionist systems (ECOS) principles on which the eSNN is based N. K. Kasabov (2018b).

Because, in the rank-order learning scheme, the synaptic weights are adjusted only once making it not very efficient for spatio-temporal data, where there may be a need to adjust synaptic weights that are based on the spikes arriving on the same synapse over time. To overcome this disadvantage, an extension of eSNN, known as dynamic eSNN (deSNN), was introduced in N. Kasabov, Dhoble et al. (2013), which combines rank-order learning with temporal learning rules, such as spike-timing dependent plasticity (STDP), which allows for dynamic adjustment the synaptic weights. However, both eSNN and deSNN do not encapsulate the structural

information of the brain in terms of neuronal locations and their connectivity, which may be crucial for modeling spatio-temporal data. The NeuCube architecture, first proposed in N. Kasabov (2012), aims at building a eSNN that incorporates structural as well as functional aspects of the brain along with utilizing STDP learning rules. The following section gives a brief introduction to the NeuCube architecture. The reader is directed to N. Kasabov (2012, 2014); N. Kasabov, Hu et al. (2013) for a more detailed introduction.

6.5.1 NeuCube

It is well known that the information in human brain is processed at different spatiotemporal levels, ranging from molecular information processing to higher order cogitive processes. The data can be acquired at different levels of these spatiotemporal processes and an efficient learning method should be able to handle complex spatio-temporal relationship from brain data at different levels. Some examples of spatio-temporal brain data (STBD) include EEG, functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI), and positron emission tomography (PET) to mention a few. Traditional methods such as support vector machines (SVM) or multilayer perceptron neural networks (MLP) typically deal with the spatial or temporal aspects of brain data and cannot handle the dynamic interaction between these processes N. Kasabov, Hu et al. (2013). Furthermore, they cannot incorporate any structural prior knowledge of the brain or handle multimodal brain data. NeuCube N. Kasabov (2012, 2014); N. Kasabov and Capecci (2015) and aslo N. Kasabov (2014); N. Kasabov and Capecci (2015); N. Kasabov, Scott et al. (2016); N. K. Kasabov (2018b) is a variant of eSNN, initially proposed to handle problems of spatio-temporal pattern recognition in brain data such as EEG, functional magnetic resonance imaging (fMRI) to cite a few, has been further

developed to handle various other types of spatio-temporal data, such as audio-visual data, climate data, seismic data, and ecological data N. Kasabov, Scott et al. (2016). The typical framework of the NeuCube system comprises of

1. An input encoding module, which converts the STBD into trains of spikes that captures temporal patterns present in the data. Various methods have been proposed to achieve this, including population coding Mastebroek, Vos and Vos (2001), address event representation S.-C. Liu and Delbruck (2010), and Bens Spike algorithm N. Kasabov, Dhoble et al. (2013).

2. A three-dimensional SNN reservoir (3D-SNNr), which takes the spike trains as input. The 3D-SNNr contains neurons that have pre-defined spatial co-ordinates and are modelled as leaky integrate and fire neurons. The initial structural connections between the neurons can be established in several ways, including small-world organization Bullmore and Sporns (2009) or based on the DTI data. Several studies utilizing EEG, fMRI, and MEG have demonstrated the presence of small-world connectivity in the brain Z. J. Chen, He, Rosa-Neto, Germann and Evans (2008); Stam (2004) and, thus, this is the preferred initial setup for the spatial structure of 3D-SNNr. Based on the temporal association between the input spikes, connections between the neurons is modified while using the spike timing dependent plasticity (STDP) rule. This is a deep unsupervised learning, as deep connectionist structures of many neurons are created as a results of the learning in space and time N. K. Kasabov (2018b).

3. A classification module, which takes the spiking patterns from 3D-SNNr as its input to perform classification.

4. An optional, Gene Regulatory Network (GRN) for controlling the learning parameter and optimization of 3D-SNNr, exploiting the fact that spiking activity is influenced by the gene and protein dynamics.

The details on the implementation of NeuCube for this study is further described in Section 6.6.6.

6.6 Methods

6.6.1 Mahnob Database

The MAHNOB-HCI dataset is a multi-modal database for affect recognition and implicit tagging Soleymani et al. (2011). In this database, 27 subjects (16 females and 11 males) aged between 19 and 40 years old were monitored while watching 20 stimulus clips (34.9 to 117 s long) that were extracted from Hollywood movies and video websites, such as `www.youtube.com` and `blip.tv`. The face video, audio and elicited physiological signals (EEG, ECG, respiration amplitude, skin temperature, GSR, and gaze data) were acquired while watching the clips. The ECG signal was obtained by subtracting a measurement from the upper left corner of chest, under the clavicle bone, from that one on left side of abdomen, below the last rib. The respiration signal was obtained by a belt placed in the subject's abdomen, skin temperature was acquired by a temperature sensor placed at the subject's little finger and GSR was obtained by passing a negligible current between the electrodes on the distal phalanges of the middle and index fingers of the subject. Gaze data were acquired with Tobii X1205 eye gaze tracker providing position of the projected eye gaze on the screen (at 60 Hz), the pupil diameter, the moments when the eyes were closed, and the instantaneous distance of the subject's eyes to the gaze tracker device.

Physiological signals, except the gaze data, were acquired at a sampling rate of 1024 Hz (down sampled to 256 Hz for further analysis), while six different views of subject's facial expressions were simultaneously recorded by six video cameras at 60 fps. In this work, the video taken only by the color camera above the screen were used. After watching each stimulus, the participants used a keyboard interface for answering five questions that were related to emotional label, arousal, valence,

dominance, and predictability. The participants answered each question using nine numerical keys, selecting nine emotional labels for the first question and nine possible levels for the last question. In this work, only the binary valence scale was used, where levels one to five were considered as low valence (unpleasant) and levels six to nine as high valence (pleasant). The database is available online http://www.ibug.doc.ic.ac.uk/resources/mahnob-hci-tagging-database/.

The multimodal emotion recognition (valence) pipeline starts with face detection in video, followed by face landmark detection, features extraction from face and peripheral signals, and ends with training and signals classification while using NeuCube.

6.6.2 Face Detection and Tracking

The first step for analyzing face emotion recognition in video is face detection and tracking in frames. Computer Vision (CV) Matlab Toolbox was used for this task. The output of this step is the corner coordinates for the polygon enclosing the face for each frame in the video.

The face detection that was carried out in this work included the following steps,

1. The face in the first frame was detected using the *vision.CascadeObjectDetector* object in the CV toolbox. This function uses the Viola-Jones algorithm Viola et al. (2001) to detect people's faces, noses, eyes, mouth, or upper body. It outputs the region of interest (ROI) for the face as a polygon, enclosing the face. Specifically, the algorithm uses the histogram-of-oriented gradients (HOG), Local Binary Patterns (LBP), Haar-like features, and a cascade of classifiers trained using boosting.

2. The corner features in the first frame ROI were detected using the *detectMinEigenFeatures* function in CV toolbox, which uses the minimum eigenvalue algorithm Shi et al. (1994).

3. For the tracking of feature points in the remaining frames, we used the Kanade–Lucas Tomasi (KLT) algorithm Lucas et al. (1981); Shi et al. (1994).

4. Finally, in order to estimate the motion of the face, we used *estimateGeometricTransform* function in the CV toolbox to apply the same transformation to the ROI that was detected in the previous frame to obtain the ROI in the next frame.

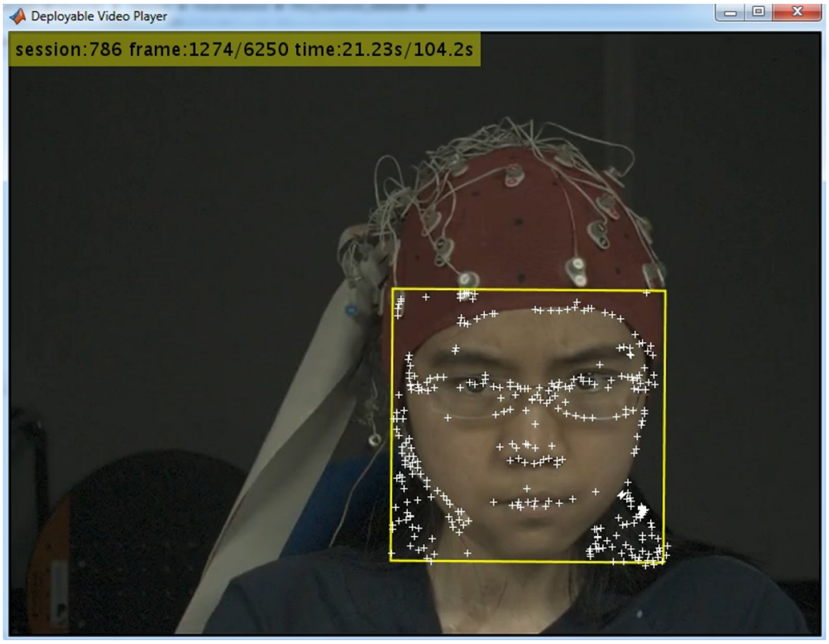


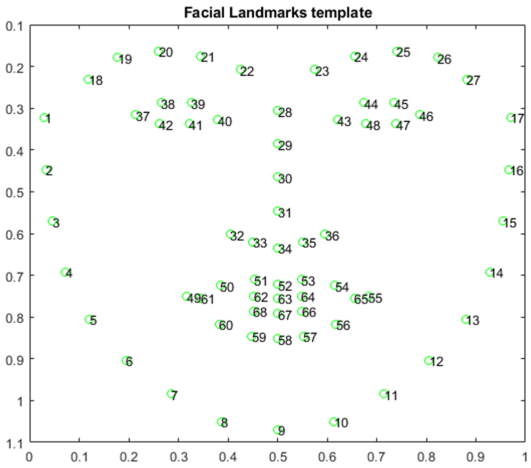
Figure 6.1: Example of face detection in Mahnob-HCI showing the feature points tracked along the video.

Figure 6.1 shows the output of the face detection step. We found that point tracking in frames to detect face is computationally more efficient than face detection in each frame. Furthermore, point tracking can manage problems that can emerge in face detection, such as making gestures with hand that may occlude parts of the face.

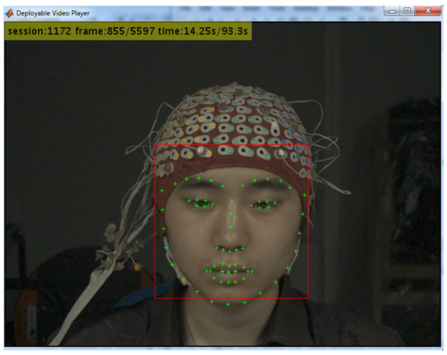
6.6.3 Face Landmarks Detection

Using the detected ROIs (See Section 6.6.2), a trained model (DLIB) for 68 facial landmarks detection was used for each frame in the video Kazemi and Sullivan

(2014). DLIB library can be obtained from http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2. The processing time for this task was around 100 s per video (i.e., approximately 30 min per subject). Figure 6.2 shows the model template (a) and one example video frame with detected facial landmarks (b) adjusted to relevant facial structures (mouth, eyebrows, eyes, nose, and face borders).



(a) Model template



(b) Detected facial landmarks in one example video frame

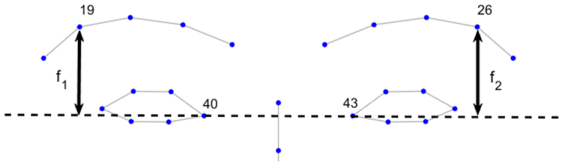
Figure 6.2: Facial landmarks detection.

6.6.4 Face Features Extraction

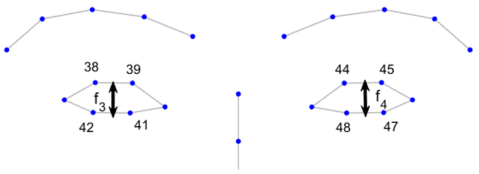
We extracted the following featured from facial landmarks (see Figure 6.3),

1. Vertical distance between the horizontal line connecting the inner corners of the eyes and outer eyebrow (f1, f2).
2. Vertical distances between the upper eyelids and the lower eyelids (f3, f4).
3. Distances between the upper lip and mouth corners (f5, f6).
4. Distances between the lower lip and mouth corners (f7, f8).
5. Vertical distance between the upper and the lower lip (f9) and distance between the mouth corners (f10)

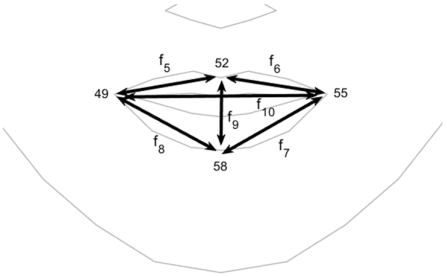
We assume that the participants hold a neutral face for the first two seconds after starting the stimulus. Because we want to detect changes in facial features, therefore the mean features in first 2 s are subtracted from facial features for each response video.



(a) Vertical distance between f1, f2



(b) Vertical distance between f3, f4



(c) Distance between f5, f6 and between f7, f8 and vertical distance between f9, f10

Figure 6.3: Facial features.

6.6.5 Physiological Features

Heart rate variability (HRV), respiration variability, respiration depth, skin temperature, GSR, and pupil diameter are used as physiological features in this study. The ECG signal is pre-processed by mean subtraction and band pass filtered with a low pass and high pass filter in cascade (Least Square FIR, 70 dB, 0.05–40 Hz, 1 dB ripple) for reducing high frequency noise as muscular activation and reducing shifting due to respiration. R waves are detected using Pan and Tompkins algorithm Pan and Tompkins (1985) for calculating the RR interval (for HRV) as a feature. The *findpeaks.m* function in Matlab (Signal Processing Toolbox) was applied to the respiration signal to detect valleys and peaks in signal and further obtain the respiration variability (time between cycles) and respiration depth (cycle amplitude). The raw Temperature (Celsius) and GSR measurements were also considered as feature signals. Additionally, from the gaze data, the mean pupil diameter (from

both eyes) was computed as an additional feature signal. Figure 6.4 shows examples of physiological features.

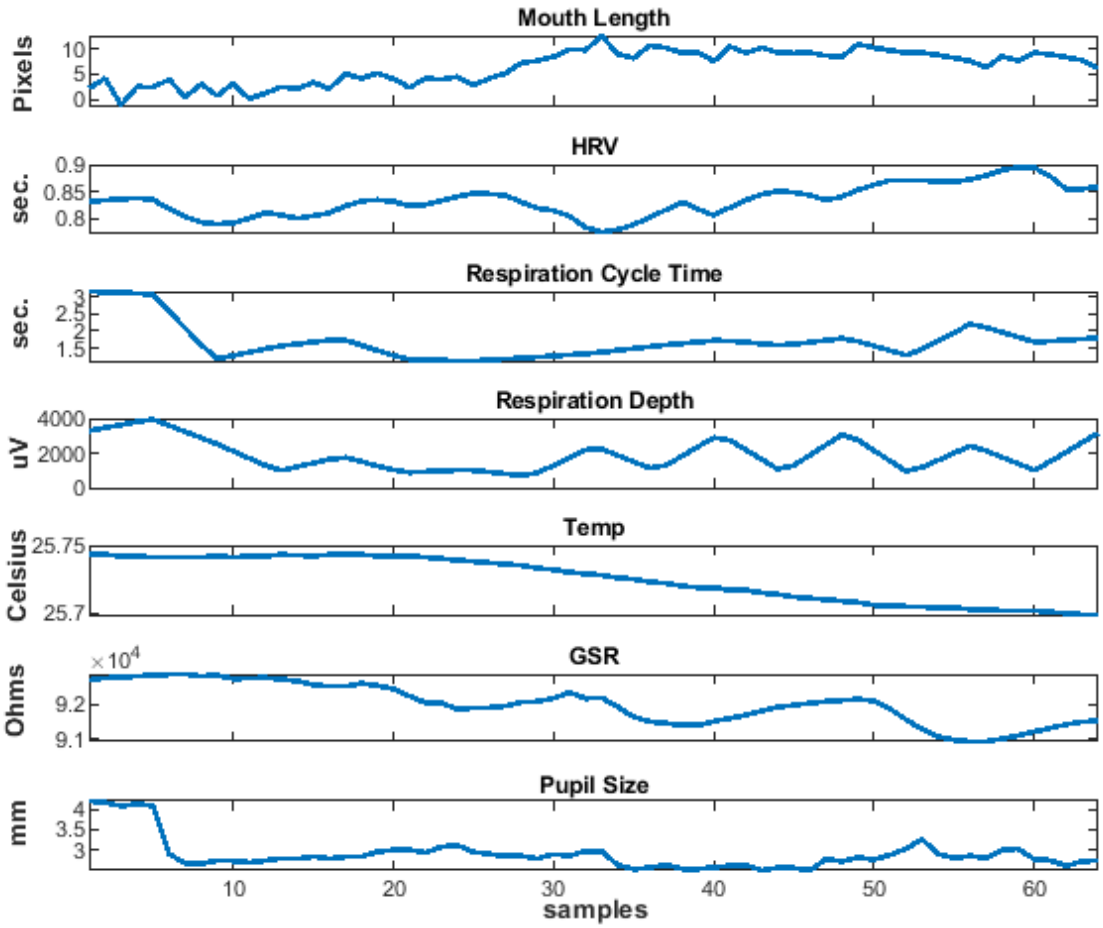


Figure 6.4: Elicited signal features in the last 30 seconds of video.

All of the facial and peripheral physiological features obtained in the analyzed window (last 30 s of video) were resampled to 64 samples. All of the features are calculated in whole video response too, and resampled to 64 points, in order to capture changes in physiological feature. The first sample is subtracted from features in windows for further analysis. We suppose that this first measurement in whole video means for resting or neutral state for physiological signals. Figure 6.5 shows the distribution of normalized features. It can be noted that mouth-related features and pupil size have better discriminative power between low and high valence. The outliers are omitted for visualization purposes.

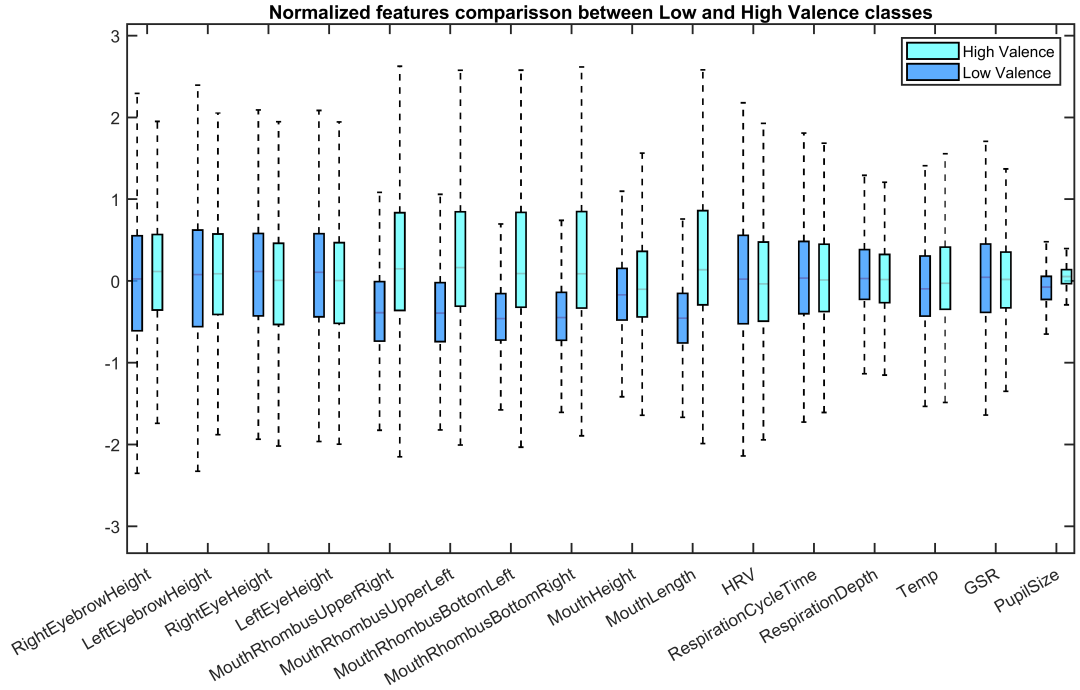


Figure 6.5: Boxplot for features in Mahnob-HCI dataset for valence emotional dimension.

6.6.6 NeuCube SNN for Facial Emotion Recognition

We used NeuCube proposed in N. Kasabov (2014) to build a system for emotion valence classification. A general scheme of our approach based on NeuCube is presented in Figure 6.6. As described in Section 6.5.1, the NeuCube structure includes Encoding, 3D-SNNr, output neuron layer, and KNN classifier. Training and classifying spatio-temporal data using NeuCube have the following stages:

- **Encoding:** encode the spatio-temporal data (features) into trains of spikes.
- **SNNr:** construct a recurrent 3D SNNr and initialize the connection weights among neurons.
- **Input neurons location:** locate the input neurons in the SNNr keeping related inputs near in space.
- **Unsupervised learning:** feed the SNNr with training data to learn in an unsupervised mode the spatio-temporal patterns in the data.

- **Supervised learning:** construct an eSNN classifier to learn to classify different dynamic pattern in SNNr activities.
- **Classification:** feed the SNNr with testing data for classification purposes.

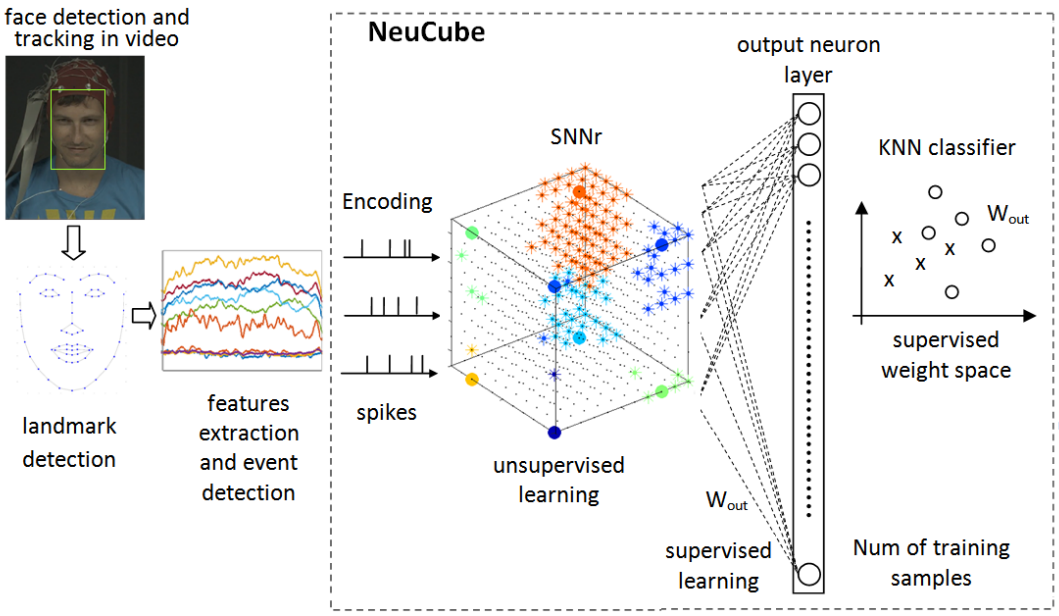
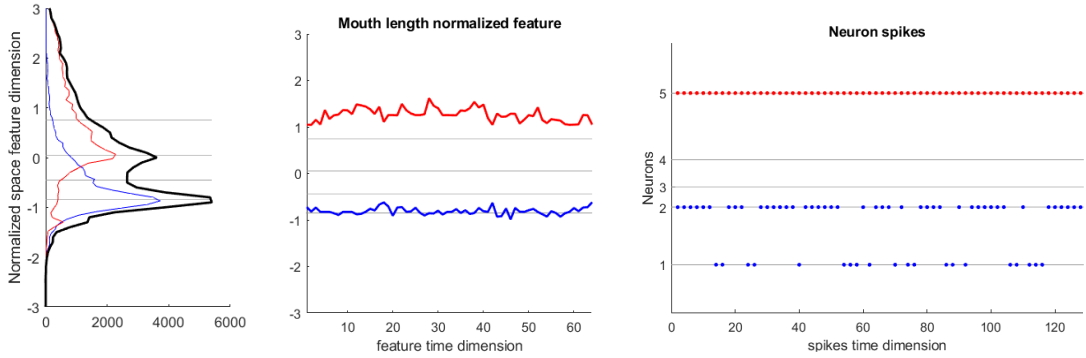


Figure 6.6: Proposed method for emotion valence classification using NeuCube.

We briefly explain each stage in the following sections.

Encoding

The coding method that we used was inspired by Gaussian Receptive Field population-based sparse coding proposed in Bohte, Kok and La Poutre (2002); Bohte, La Poutré and Kok (2002). This method codes each continuous value from a time-based feature to spikes emitted at different times by a neuron population. The whole feature range is covered for the neurons and the time for generating the spikes depends on the distance from the current value to the center of a Gaussian receptive field covering each value interval. We used a population of five neurons per feature, in which only a neuron from the group spikes at the current time step. Figure 6.7 shows an example of coding the mouth length feature. Note that the dimension of feature is 64 and the temporal dimension of each spikes train is 129, because zeros are inserted between the spikes.



(a) blue: low valence, red: high valence, black: low and high valence

(b) time course of mouth length feature – low valence (blue) and one high valence (red)

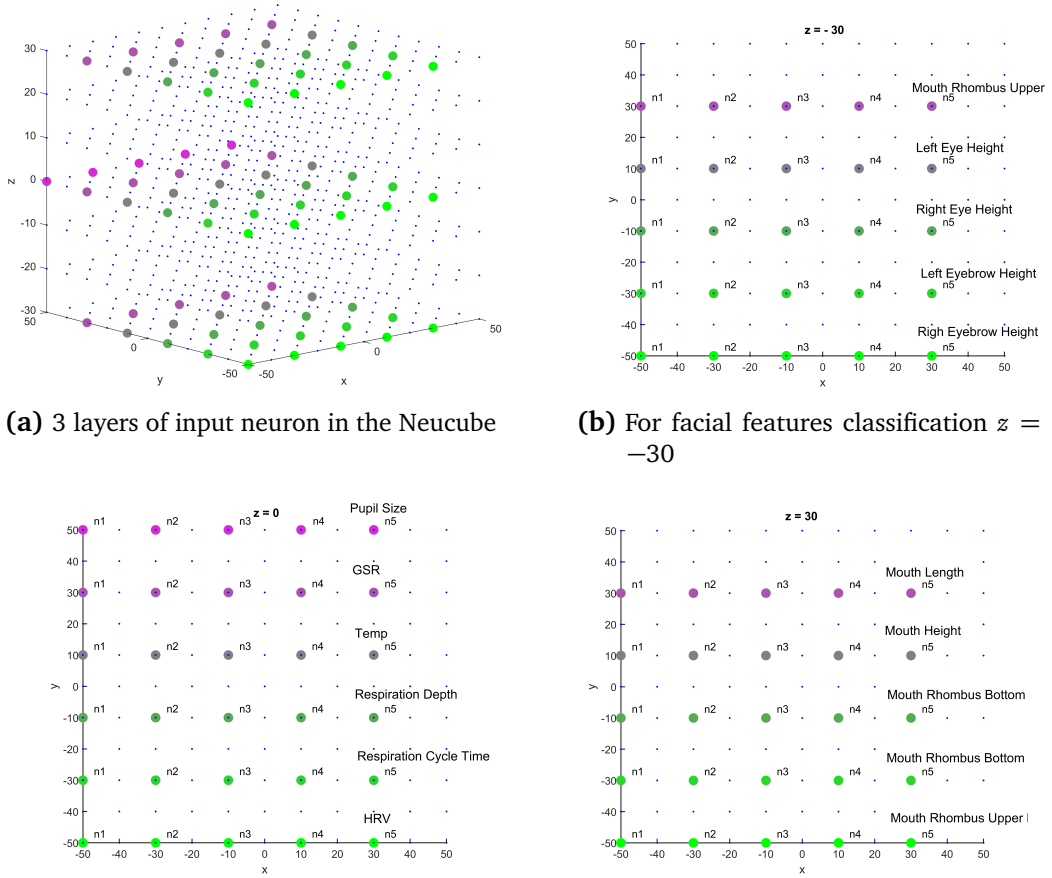
(c) low valence in blue, high valence in red

Figure 6.7: Encoding Continuous feature values to five neurons spiking.

It can be noted from the distribution of mouth length feature (Figure 6.7a, left plot; blue: low valence, red: high valence, black: low and high valence), that there are two peaks in the distribution indicating the separation between the two class. In the middle plot (Figure 6.7b), the time course of mouth length feature in a low valence event (blue) and one high valence event (red) for the subject 1 are shown. In the right plot (Figure 6.7c), spikes generated for these two events are shown (low valence in blue, high valence in red). The levels that define the receptive fields or range for exciting each neuron are defined using the feature distribution in the data from all detected events for all analyzed subjects. Levels for each five neuron population are automatically obtained by analyzing the histogram in such a way that the five ranges have the same count of value occurrences. The levels are shown as gray lines (left and middle plots in Figure 6.7b). Note that each feature value in time produces a spike in only one neuron from the population. Eighty input neuron are allocated in NeuCube network, as we have ten facial features and six peripheral signals.

Construction of SNNr

When brain imaging data, such as EEG, Are used, the SNNr can be built with a shape resembling the human brain N. Kasabov (2014) and the input neurons can be located based on the anatomical location of the EEG electrodes. However, in this study, as we are building a general classifier of facial features, we chose to build an $11 \times 11 \times 7$ array of neurons (equally spaced in x and y axes), as shown in Figure 6.8. Each five neuron population are spatially arranged in NeuCube structure in lines, as illustrated in Figure 6.8; this way neighbor neurons code similar feature values favoring spatial neuron specialization.



(a) 3 layers of input neuron in the Neucube

(b) For facial features classification $z = -30$

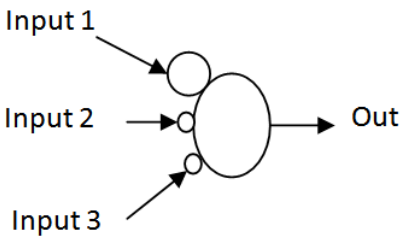
(c) For peripheral features classification $z = 0$

(d) For facial features classification $z = 30$

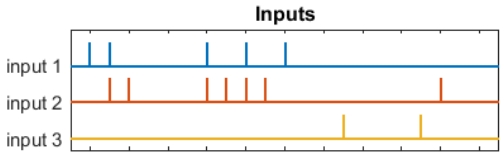
Figure 6.8: Input neurons location for facial and peripheral features classification. n1 means for the neuron coding the lowest values and n5 the highest ones for each feature. Note there are 3 layers of input neuron in the cube, located at $z = -30$ (facial), $z = 0$ (peripheral), and $z = 30$ (facial).

The SNNr was made with leaky integrate and fire model (LIFM) spiking neurons with recurrent connections. In this neuron model, the post-synaptic potential (PSP) increases or decreases with every input spike from pre-synaptic neurons. The effect of each spike is modulated by the corresponding synaptic connection weight. If PSP reaches a specific threshold (0.5 in this work), then the neuron emits an output spike toward its connected neighbours and the PSP resets to a reference value. The PSP can leak between spikes with a predefined time constant τ, if we are using an exponential model or a constant leak time. The latter is used in this work and is set to 0.002. After a neuron spikes, the absolute refractory time (equal to 1 in this work) is simulated by disabling it to increase the PSP until a certain unit time has passed. Figure 6.9 shows an example of LIFM neuron simulation with a refractory time that is equal to three seconds, potential leak rate equal to 0.02, a threshold of firing that is equal to 0.5 and synapses weights of 0.1, 0.1, and 0.35. It can be noted in Figure 6.9 that the accumulation of spikes in time leads to an increase of PSP until a spike is generated and the effect of disregarding input spikes immediately after a spike is generated.

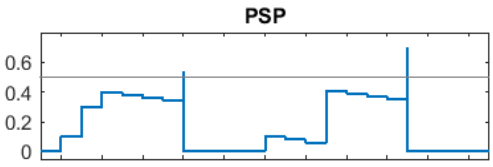
We set the initial connections (synapses strength) between neurons in SNNr using small-world connectivity Braitenberg and Schüz (2013); Bullmore and Sporns (2009). The connection probability was set, such that neurons were more likely to be connected to neighboring neurons than to the distant ones. It has been shown that such an approach brings some advantages with regard to learning speed, parallel processing, and also favors the linking of specialized processing cluster units Simard et al. (2005). Additionally, we defined a radius r to be the maximum distance of connections of one neuron to another in the reservoir ($r = 25$ in this study). The initial weights were assigned as the product of random values $[-1, +1]$ divided by Euclidean distance between pre-synaptic and post-synaptic neuron, so that 80% of them were positive values (excitatory connection), while 20% of them were negative values (inhibitory connections). Neuron connections are unidirectional,



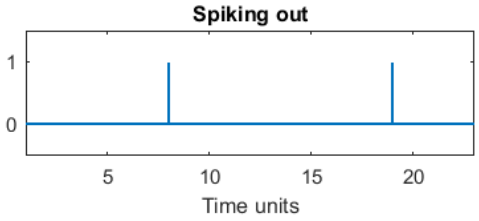
(a) Input neurons for synapses weights of 0.1, 0.1 and 0.35



(b) Each Input is converted into spikes



(c) Post-synaptic potential increases or decreases with the threshold of firing = 0.5



(d) Neuron emits an output spike toward its connected Neuron

Figure 6.9: Leaky integrate-and-fire model (LIFM) neuron model. Small circles at neuron inputs represent connection weights. Note that input 1 has a bigger weight and it produces a larger effect in PSP.

and the direction of communication was selected randomly. Connections between input neurons and other neuron are always positive and with doubled weight in comparison with other random connections. These connections were modified in the unsupervised learning stage in order to adapt to spatio-temporal patterns in input data.

Deep, Unsupervised SNN Training

We adjusted the connections between the neurons using the training data and a learning rule-based on Hebbian plasticity, called spike-time-dependent plasticity (STDP) S. Song et al. (2000). STDP learning modifies the neuronal connection weights while taking into account the time difference between post- and pre-synaptic firing. A connection is strengthened, if postsynaptic firing occurs after presynaptic firing; otherwise, it is decreased. After STDP learning, the spatio-temporal pattern

was saved in the value of connection weights in the SNNr. STDP learning rule is given as,

$$\Delta w = \text{sgn}(\Delta t)\frac{LR}{|\Delta t| + 1} \tag{6.1}$$

where LR is the STDP Learning Rate (0.001 in this work), sgn$(\cdot)$ is the function sign (-1 for negative values and 1 for positive), Δt is the difference between post- and pre-synaptic times ($\Delta t = t_{\text{post}} - t_{\text{pre}}$) and Δw is the change in the connection weight. The Hebbian relation Δw vs Δt is depicted in Figure 6.10. The learning results in the creation of deep structures of connections between neurons in the SNNr.

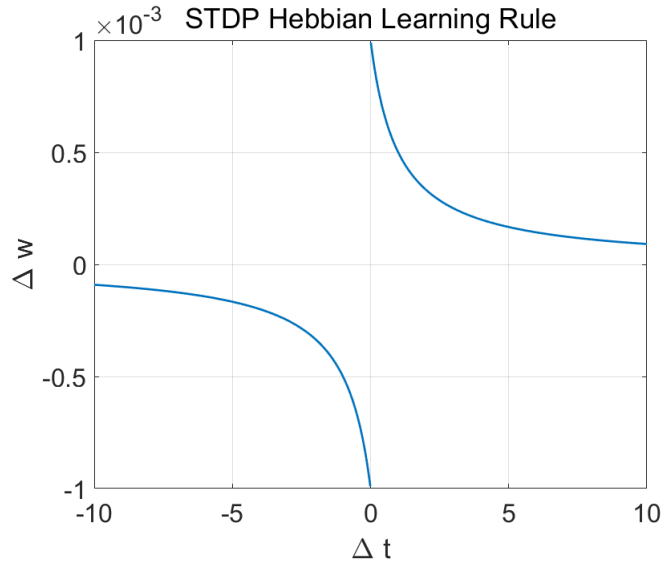


Figure 6.10: Hebbian Learning rule, connection (synaptic modification) vs difference between post- and pre-synaptic times.

Supervised Output Neurons Training

The deSNN is applied for supervised learning N. Kasabov, Dhoble et al. (2013). For every single training sample, an output neuron was created and connected to all of the neurons in the trained SNNr (see Figure 6.6). Each output neuron was trained using the corresponding training sample by propagating the signal through the network once more. The neuron's connections weights $w_{i,j}$ between neurons i (in the reservoir) and j (output neuron) were initially established using rank order (RO) rule N. Kasabov, Dhoble et al. (2013). The RO method ranks the order in

which the first spike arrives in the j neuron and the weights are given as,

$$w_{i,j}(0) = \alpha \mod^{\text{order}(i,j)} \tag{6.2}$$

where α is a learning parameter (in a partial case, equal to 1), mod is a modulation factor that defines how important the order of the spike is (0.8 in this study), order(i,j) represents the order (the rank) of the first spike at synapse (i,j) ranked among all of the spikes arriving from all synapses to the neuron j. Furthermore, order$(i,j) = 0$ for the first spike to neuron j and increases according to the input spike order at other synapses.

Once a synaptic weight $w_{i,j}$ is initialized, based on the order of the first spike from i to j, the synapse becomes dynamic. It increases its value with a small positive value (drift $= 0.005$) at any time t a new spike arrives at this synapse and decreases its value if there is no spike at this time, as described in the following formula,

$$w_{i,j}(t) = \begin{cases} w_{i,j}(t-1) + \text{drift}, & \text{if} \quad S_{i,j}(t) = 1 \\ w_{i,j}(t-1) - \text{drift}, & \text{if} \quad S_{i,j}(t) = 0 \end{cases}$$

where $S_{i,j}(t)$ describes the existence of spike from neuron i entering to neuron j at time t. Every generated output neuron was trained to recognize and classify spatio-temporal patterns of weights adjusted by a corresponding labeled input training sample.

Classification

At classification stage, the NeuCube is fed with validation data. For each sample, data synaptic weights for output neurons are calculated while using the same supervised rules used in supervised training procedure. The connection weights that are learned in this process are then classified using a K-nearest neighbor (KNN, with $K = 3$ neighbors) algorithm and the labels that are known for all of the samples.

We ran the whole NeuCube framework in a leave one subject out mode (LOSO) in order to test its capacity for learn spatio–temporal features from subjects and classify an unseen new subject.

Fusion of Multimodal Signals

Two schemas for the fusion of multimodal signals were explored—(1) features-level and (2) decision-level fusion. For features-level fusion, we coded all of the features (facial and peripheral) and included as input in NeuCube. Regarding decision-level fusion, for each subject, we calculated the accuracy of NeuCube classification in training data (rest of subjects) for separated modalities (facial and peripheral), and we chose the method with higher accuracy as the method for doing validation classification for the specific subject.

NeuCube Parameters

NeuCube performance in analyzing spatio-temporal data depends on several parameters. We chose a set of default parameter values that are equal to that used in the NeuCube development system publicly available online http://www.kedri.aut.ac.nz/neucube, with the exception in refractory time. We used one time unit for this parameter in order to increase neuron activity. The NeuCube parameters used in this work are given in Table 6.1.

Table 6.1: NeuCube parameters.

Small world radius (r)	25
STDP learning rate (LR)	0.001
Threshold of firing	0.5
Potential leak rate	0.002
Refractory time	1 s
mod	0.84
drift	0.005
K	3

6.7 Results

NeuCube framework was fed with coded data under a LOSO cross validation scheme, i.e., all of the data from a specific subject were excluded from the training set. All the parameters were fixed with values mentioned in Method section. Table 6.2 shows classification accuracy results in Mahnob-HCI dataset. We also included

Table 6.2: Video valence classification accuracy in Mahnob-HCI dataset using NeuCube.

Subject ID	Facial Features Accuracy (%)	Physiological Features Accuracy (%)	Fusion Detection Accuracy (%)	Fusion Features (%)
1	73.33	66.67	66.67	73.33
2	62.5	50	62.5	56.25
3	75	72.73	75	81.82
4	78.57	66.67	66.67	83.33
5	75	50	75	68.75
6	58.82	70.59	58.82	70.59
7	75	60	75	93.33
8	64.29	50	64.29	64.29
9	60	100	60	90
10	61.54	69.23	61.54	69.23
11	78.57	61.54	61.54	76.92
13	64.29	71.43	64.29	85.71
14	50	57.14	50	71.43
16	72.73	63.64	63.64	63.64
17	62.5	80	62.5	40
18	41.67	50	41.67	62.5
19	61.54	75	61.54	58.33
20	53.33	73.33	53.33	80
21	66.67	71.43	66.67	71.43
22	66.67	66.67	66.67	80
23	75	50	75	75
24	69.23	53.85	69.23	76.92
25	66.67	77.78	66.67	55.56
27	68.75	60	68.75	80
28	66.67	66.67	66.67	86.67
29	68.75	50	68.75	64.29
30	78.57	57.14	78.57	78.57
Total	66.67	63.84	65.11	73.15

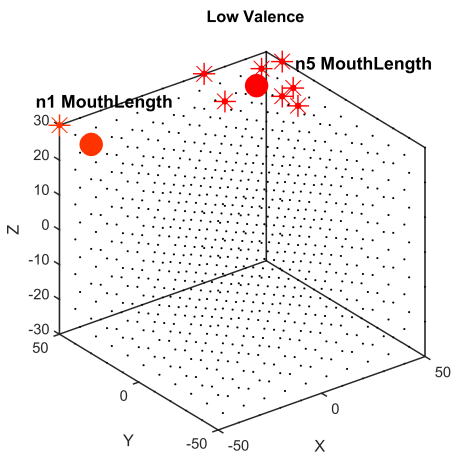
F1-score since some subject has imbalanced data, i.e., more than twice the number of sample for one class than the other. Total accuracies at the end of the table can be used for comparison with other works, because they are calculated using all the data which can be assumed as balanced, 390 videos were analyzed (207 : 53.07% low valence, 183 : 46.92% high valence).

Paired sample t-test when comparing the F1-score from Peripheral and Facial features result in no difference between them ($p < 0.05$). F1-score using decision-level features does not show a difference with facial nor peripheral ($p < 0.05$). Additionally, feature-level fusion F1-score (0.74) results in being better than decision-level fusion F1-score, 0.67 ($p < 0.01$).

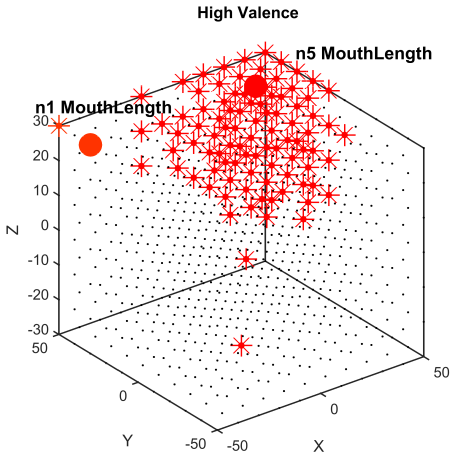
For decision-level fusion, we obtained a mean accuracy of 83.7% for classifying the training data using facial features and 80.94% using peripheral training data.

6.7.1 Clustering Spike Communication

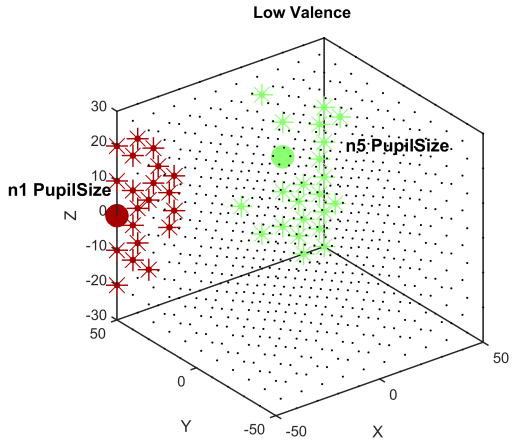
NeuCube framework has an option to analyze clusters of neuron-surrounding input neurons using the spike amount communicated between a pair of neurons. Figure 6.11 shows an example using this tool when the neuron reservoir is trained separately with one class (low valence) and the other one (high valence). For visualization purposes and taking into account that mouth length and pupil size have more discriminative power regarding the rest of features, only input neurons coding higher and lower values in mouth length and pupil size are shown. Figure 6.11 shows that neurons coding high values of mouth length and pupil size are more active for high valence and for this reason the cluster of spiking communication surrounding these neurons are bigger. Note that neuron coding low values of pupil size is more active for low valence. These results agree with features distribution in Figure 6.4.



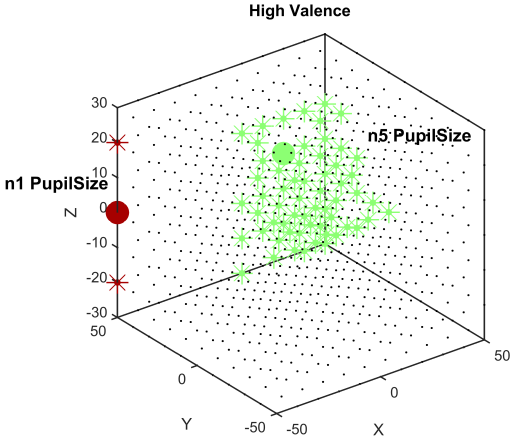
(a) Show lower spike values in mouth length



(b) Show higher spike values in mouth length



(c) Show lower spike values in pupil size



(d) Show higher spike values in pupil size

Figure 6.11: Neuron activity pattern example when NeuCube is trained using each Separate data (low and high valence).

6.8 Discussion

In this work, we developed an approach based on NeuCube N. Kasabov (2014), which is an eSNN framework, in order to classify emotional valence using multimodal dataset that included video and physiological signals. We used a population coding scheme, based on ROC to encode input data into spikes, which SNNs can handle. When tested on the benchmark dataset, the MAHNOB-HCI, our approach resulted in a accuracy about 73.15% for emotion classification. To the best of our knowledge, there has not been any other study to utilize SNN for affect recognition with multimodal data. In addition to the good accuracy of classification, the SNN

system can be incrementally trained on new data and new features in an adaptive way, allowing for the system to be used in an on-line applications N. K. Kasabov (2018b).

6.8.1 Related Work

The MAHNOB-HCI dataset has been used in several studies owing to its difficulty for the classification of spontaneous emotional responses from subjects. Several studies have resorted to a multimodal approach because the MAHNOB-HCI dataset also contains multimodal data in the form of physiological and audio signals.

In a study by Koelstra and Patras Koelstra and Patras (2013), EEG and facial expressions were fused to perform affect recognition and implicit tagging. In case of EEG, power spectral density (PSD) features were used and for facial expression, an AU detection method was used, which was originally proposed in Koelstra et al. (2010). Basically, the AU detection was performed using Free-form Deformations (FFDs) and Motion History Images too. For facial recognition, they trained the system using the MMI dataset Valstar and Pantic (2010) and obtained 64.5% of binary valence classification using only facial features and 74% by combining facial and EEG features. They performed a per-subject leave-one-trial-out cross-validation, where the classifier is trained on 19 trials from the same subject and tested on the 20th. As can be seen from their study, only using facial features result in low accuracies and fusion with EEG signal improved the classification accuracy.

Boxuan and colleagues developed a temporal information preserving framework by splitting signals into multiple stages in each video. They achieved a valence (unpleasant, neutral, pleasant) classification accuracy of 54% using only facial expression and 69% when fusing with physiological signals Zhong et al. (2017). They used Affdex SDK software McDuff et al. (2016), trained in 10,000 manually labeled facial images, which classify emotion-based on HOG features and support vector machine (SVM) classifier. Huang and colleagues obtained 50.57% for valence

classification using appearance descriptors based facial features (local binary pattern from three orthogonal planes, LBP-TOP) and 66.28% using fusion it with global EEG features Y. Huang et al. (2019). They used the LOSO cross-validation scheme in nine emotion categories. A convolution deep belief network (CDBN) was proposed in Ranganathan et al. (2016) in order to learn emotional features from multimodal datasets and the authors reported a classification accuracy of 58.5% with the MAHNOB-HCI dataset. Torres et al. Torres-Valencia, Álvarez-López and Orozco-Gutiérrez (2017) performed feature selection using discriminant-based algorithms, while using EEG and peripheral signals. Their results showed that EEG-related features show the highest discrimination ability. Furthermore, it was shown that EEG features, along with GSR, achieved the highest discrimination for arousal index, whereas for the valence index, EEG features are accompanied by the heart rate features in achieving the highest discrimination power. For the MAHNOB-HCI dataset, they obtain a classification accuracy of 66.09% and 69.59% in the valence and arousal dimension, respectively. Liu et al. J. Liu, Su and Liu (2017) tested a deep learning approach based on multi-layer Long short-term memory recurrent neural network (LSTM-RNN) for emotion recognition, which combined temporal attention and band attention. They achieved an accuracy of 74.5% in valence classification (9 class) fusing video and EEG analysis. They used 20 participants for training, four participants for validation, and three participants for testing. Huang et al. X. Huang et al. (2016) used transfer learning technique (pre-trained convolutional neural network, CNN) to obtain an 73.33% in binary valence accuracy in MAHNOB-HCI dataset using facial features and 75.21 fusing with EEG features.

Overall, the results that we have obtained from the MAHNOB-HCI dataset are comparable with the state of the art work learning methods applied on this database. We observe that, in some cases, the classification accuracy obtained using our SNN approach is better than the ANN approach, that have also used EEG signals, which we have excluded. It is also to be noted that it is difficult to establish a fair

Table 6.3: Comparison with related works on valence classification using Mahnob-HCI dataset.

Works	Features	Method	Classes	Cross-Validation	Accuracy %
Koelstra and Patras (2013)	Facial + EEG	Free-form Deformation and Motion History Images	Binary valence	Trained with MMI dataset, and data from the same subject	74
Zhong et al. (2017)	Facial + Physiological	Temporal Information Preserving Framework, SVM	Valence (3 classes)	LOSO	69
Y. Huang et al. (2019)	Facial + EEG	LBP-TOP, Transfer learning CNN, SVM	9 emotion categories	LOSO	62.28
Ranganathan et al. (2016)	Facial + Body + Physiologic	Convolutional deep belief network (CDBN) and SVM	Not mentioned	LOSO	58.5
Torres-Valencia et al. (2017)	EEG + Peripheral	Discriminant-based algorithm, SVM	Binary valence	80% train data-20% test data	66.09
J. Liu et al. (2017)	Facial + EEG	LSTM-RNN	Valence, 9 classes	24 subject training and 3 for testing	74.5
X. Huang et al. (2016)	Facial + EEG	Pretrained CNN	Binary valence	LOSO	75.21
ours	Facial + Peripheral	SNN, feature-level fusion	Binary valence	LOSO	73.15

comparison with most of the previous works, as we did not include EEG features and use pretrained models, as in X. Huang et al. (2016). We also disregarded all of the data related with the subject in a Leave-subject out validation scheme. Furthermore, in contrast with Deep Learning approaches, our SNN based method provides more interpretability of the model due to specialization of neurons clusters, needs for fewer data to train, and this can be done online with one pass of a new training sample.

6.8.2 Limitations

Our work has several limitations. First, we did not include any EEG features, because changes in EEG features that are associated with emotion are lumped features and we wanted to test NeuCube with temporal spatial patterns. The addition of raw EEG temporal signals into NeuCube would need for addition of much more input neurons into the model. We found this unfeasible to compute in a reasonable time for our experiments. Actually, temporal variations on EEG features for emotion detection are in a different scale than variation in other peripheral signals, we are using a 30 s window to analyze changes in physiological and facial features, this is a very short time to expect changes in lumped EEG features due to emotions. As discussed previously, several studies have shown that including EEG features considerably improves the classification accuracy. However, there are several challenges in using EEG for emotion recognition X. Hu, Chen, Wang and Zhang (2019), including the selection of robust features, continous decoding of affective states, reliable decoding of long-term reliability of EEG recordings for such studies, long preparation time, and, most importantly, adopting a proper model of emotion with regard to EEG and understanding the EEG representation of affective states. For an excellent overview of these challenges the reader is directed to X. Hu et al. (2019). Nonetheless, the possibility of using EEG with the NeuCube framework will be explored in our future studies. Second, other important features that could be utilized from the multimodal data could be speech and postures. Several studies have considered the implications of including speech in affect recognition, with pitch being considered to be an index into arousal Calvo and D'Mello (2010), although the classification accuracy is shown to be lower than facial expression. Nonetheless, this feature should definitely be considered in the future studies with SNN given the noninvasive and easy procedure to acquire voice. With regard to posture tracking, again it is a non-intrusive acquisition to the user's experience, but the equipment requires more expensive equipment as compared to speech. Additionally, there are some

constraints with regard to the user's position, for example the user should be sitting Calvo and D'Mello (2010).

We have also assumed that the face that is captured during the first two seconds after the stimulus is presented is neutral and consider it as the baseline. This could be problematic, especially if the participant is tired. Because we chose the last 30 s window for event selection in each video, we do not take into account the long-lasting facial expression. It could be interesting if long term facial variation inside the video could be considered as detected events. Additionally, it could be interesting to incorporate detecting facial micro-expressions in our framework but this is in general challenging due to limited availability of such data and as well as difficulties in analyzing minute changes in expression Y. Wang et al. (2015). Few methods have been proposed to address the problem of detecting micro-expressions using spatio-temporal local texture descriptor X. Li et al. (2013), Gabor filter with SVM classifier Q. Wu et al. (2011), and LBP-TOP with nearest neighbor classifier Guo et al. (2014), which can be incorporated to add more information for the SNN framework. Another improvement could be to normalize expressions between subjects by using pose estimation Murphy-Chutorian and Trivedi (2008), or correction of a 3D model Jourabloo and Liu (2015); X. Zhu et al. (2016). Further improvements could be made along the lines of detecting non-frontal head poses, identity bias, as well as illumination variation.

Although we studied the effect of varying certain NeuCube parameters, the performance of the proposed system may be affected by the choice of several other parameters. For instance, the effect of varying other NeuCube parameters such as radio, firing threshold, refractory time, and NeuCube resolution should be carefully investigated. The NeuCube framework also provides parameter optimization tool, which could be utilized instead of setting the parameters in an ad hoc manner.

6.9 Conclusions

Utilizing multimodal data to solve the problem of affect recognition with state of the art deep learning methods has gained a lot of popularity. SNNs offer an alternative to ANNs, where, in the former, is biologically more realistic model of neurons. In this work, we proposed a novel solution of using a variant of SNN, known as NeuCube, which is an eSNN, to solve affect recognition problem using multimodal data obtained from MAHANOB-HCI dataset. The eSNN is based on the ECOS principles which includes, efficient processing of spatio-temporal data and open evolving structure. Despite not including EEG, our approach provided results comparable to deep learning methods that utilize multimodal data, including EEG. In addition to the good accuracy of classification, the SNN system can be incrementally trained on new data and new features in an adaptive way, allowing for the system to be used in on-line applications.

7 Conclusion

This thesis has advanced the current status of affective computing research in many ways. The four research manuscripts outcome, relate to both the theory and practice in sensing human emotion using the third-generation artificial intelligence – Spiking Neural Network (SNN). The concluding chapter discusses the significant contributions towards the literature and the application of novel methods towards the research questions that were posed in the introduction section. This section highlights the fundamental limitations that have already been discussed and carefully examined in each of the manuscripts. The outcome of the research has also identified many open questions that should provide a further review for future research.

7.1 Research Questions and Contributions

7.1.1 How to design architectures of spiking neural networks that are capable of efficiently model the temporal, Spatio-temporal elements of the changing human emotions or expressions.

- **Spiking Neural Network (SNN) manuscript – refer to chapter 3.** The research paper reviews recent developments in the still-off-the-mainstream information and data processing area, design architectures of spiking neural networks (SNN) – the third generation of artificial neural networks. We provide background information about the functioning of biological neurons, discussing the most important and commonly used mathematical neural models. Most relevant information processing techniques, learning algorithms, and applications of spiking neurons are described and discussed, focusing

on feasibility and biological plausibility of the methods. Specifically, we describe in detail the functioning and organisation of the latest version of a 3D Spatio-temporal SNN-based data machine framework called NeuCube, as well as its SNN-related submodules. All described submodules are accompanied by formal algorithmic formulations. The architecture is highly relevant for the analysis and interpretation of various types of Spatio-temporal brain data (STBD), like EEG, NIRS, fMRI. However, we highlight some of the recent both STBD- and non-STBD-based applications. Finally, we summarise and discuss some open research problems that can be addressed in the future. These include, but are not limited to: application in the area of EEG-based BCI through transfer learning; application in the field of affective computing through the extension of the NeuCube framework which would allow for a biologically plausible SNN-based integration of central and peripheral nervous system measures.

7.1.2 How to perform neural encoding on facial expression and physiological data to represent human emotions as timings of spikes?

- **FacialSense manuscript – refer to chapter 4.** Affective computing aims at establishing smooth and harmonious interaction between humans and computers. Accurately detecting and interpreting facial emotions is a critical task in affective computing. Although several deep learning approaches based on artificial neural networks (ANNs) have been proposed and applied, building a robust facial emotion recognition (FER) system is still a challenge. Spiking neural networks (SNNs) represent the third generation of neural networks and employ biologically plausible models of neurons. SNNs have been shown to handle Spatio-temporal data, which is essentially the nature of data encountered in FER problem, efficiently. In this work, for the first time, we propose the application of SNNs to solve the FER problem. Specifically, we

use the NeuCube framework, which employs an evolving SNN architecture to classify emotional valence and evaluate the performance of our approach on the MAHNOB-HCI dataset. Our results show that, despite using only information from facial expression, the proposed method achieves a classification accuracy of 77.01%, which is comparable with other state-of-art deep learning approaches that utilise information both from facial expressions and physiological signals such as electroencephalogram (EEG). In conclusion, we have demonstrated that the SNN can be successfully used for solving the FER task and provide directions for future research utilising SNN for affective computing. In this work, we proposed a novel solution of using a variant of SNN known as NeuCube which is an eSNN (evolving SNN), to solve the FER problem and exploited some of the ECOS principles on which eSNN is based on, such as efficient processing of Spatio-temporal data and open evolving structure. We also proposed methods for feature extraction using event detection based on Otsu's thresholding algorithm. We showed that our approach, utilising only facial expressions, gives results comparable to deep learning methods that use multimodal data, including facial expression and physiological signals.

7.1.3 How to integrate spatial, temporal and event-related potential present in unimodal and multimodal human physiological data using spiking neural network architecture?

- **NeuroSense manuscript – refer to chapter 5.** Current EEG-based approaches to emotion recognition mostly rely on various handcrafted features extracted on relatively long-time windows of EEG during participants exposure to appropriate affective stimuli. In this research, we present a short-term emotion recognition framework based on spiking neural network (SNN) modelling of Spatio-temporal EEG patterns. Our method is based on event-related potential (ERP); the technique applies EEG signal segmentation based on detection

of short-term changes in facial landmarks and relies on no handcrafted EEG features. Differences between participants EEG properties are considered via subject-dependent spike encoding in the formulated subject-independent emotion recognition task. We test our methods on DEAP and MAHNOB-HCI databases due to the availability of both EEG and frontal face video data. Through exhaustive hyperparameter optimisation strategy, we show that SNN-based representation of EEG spiking patters provides valuable information for short-term emotion recognition. The obtained accuracies are 78.97% and 79.39% in arousal classification, and 67.76% and 72.12% on in valence classification, on the DEAP and MAHNOB-HCI datasets, respectively. Results encourage the use of presented EEG processing methodology as a complement to existing features and methods commonly used for EEG-based emotion recognition.

- **FusionSense Manuscript – refer to chapter 6.** Using multimodal signals to solve the problem of emotion recognition is one of the emerging trends in affective computing. Several studies have utilised state of the art deep learning methods and combined physiological signals such as the electrocardiogram (EEG), electroencephalogram (ECG), skin temperature along with facial expressions, voice, posture to name a few, to classify emotions. Spiking neural networks (SNNs) represent the third generation of neural networks and employ biologically plausible models of neurons. SNNs have been shown to handle efficiently Spatio-temporal data, which is essentially the nature of data encountered in emotion recognition problems. In this work, for the first time, we propose the application of SNNs to solve the emotion recognition problem with multimodal datasets. Specifically, we use the NeuCube framework, which employs an evolving SNN architecture, to classify emotional valence and evaluate the performance of our approach on the MAHNOB-HCI dataset. The multimodal data used in our work consist of facial expressions along

with physiological signals such as ECG, skin temperature, skin conductance, respiration signal, mouth length and pupil size. We perform classification under Leave-One-Subject-Out (LOSO) cross-validation mode. Our results show that the proposed approach achieves an accuracy of 73.15% for classifying binary valence when applying feature-level fusion, which is comparable to other deep learning methods. We accomplish this accuracy even without using EEG data, which other deep learning methods have relied on to achieve this level of accuracy. In conclusion, we have demonstrated that the SNN can be successfully used for solving the emotion recognition problem with multimodal data and also provide directions for future research utilising SNN for affective computing. In addition to the excellent accuracy, the SNN recognition system is incrementally trainable on new data in an adaptive way. It requires only one pass training, which makes it suitable for practical and on-line applications. These features are not manifested in other methods for this problem.

7.2 Future Direction and closing remarks

The limitations of the research work for each of the individual pieces of work have been discussed in detail in all the research manuscripts. Therefore, only the overall limitation of this work will be highlighted as a prelude to the discussion for the future research direction.

- **Towards learning long Spatio-temporal micro-expression patterns using the context of associative memory management.** The research with facial emotion recognition (FacialSense) has highlighted several limitations. However, at the same time, it also suggests some future research, especially in the area of long-lasting facial expression that was not taken into account. However, in order to incorporate and detect facial micro-expressions within the Spatio-temporal patterns, it requires research into the highly compressed

representation of the data using spike encoding and to augment associative memory management into the framework. Such research outcome would enable future implementation of an accurate and efficient emotion-sensing machine.

- **Towards integrating one emotion-sensing spiking encoding multimodal data cube, such as fusing facial, speech, posture and physiological signals.** SNNs have clear comparative advantages, including but not limited to, power consumption, computational speed and 3D biological potential. Although technical challenges continue to exist, it has nevertheless, provides researchers with many opportunities to advance the work of the next generation of brain-inspired spiking neural network. NeuCube's 3D SNN structure offers researchers with a better understanding and interpretation of Spatio-temporal brain data (STBD) as it allows for the integrative modelling of various STBD. Although the network structure is rather intuitive, the real challenge is to design an SNN structure for non-STDB-based classification and pattern recognition tasks such as for emotion-sensing system that involves fusing the evolving multimodalities cube structure with facial, voice, posture and physiological data. A significant research direction towards the emotion-sensing machine by fusing multimodal facial, posture, voice and physiological data into a connected multi-cube framework with communication capabilities between and within the cube.

Bibliography

Abbott, A., Sengupta, N. & Kasabov, N. (2016). Which method to use for optimal structure and function representation of large spiking neural networks: A case study on the neucube architecture. In *Neural networks (ijcnn), 2016 international joint conference on* (pp. 1367–1372). IEEE.

Abbott, L. F. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain research bulletin*, *50*(5-6), 303–304.

Adibuzzaman, M., Jain, N., Steinhafel, N., Haque, M., Ahmed, F., Ahamed, S. & Love, R. (2013). In situ affect detection in mobile devices: a multimodal approach for advertisement using social network. *ACM SIGAPP Applied Computing Review*, *13*(4), 67–77.

Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S. & Dutoit, T. (2018). The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Adrian, E. D. (1926). The impulses produced by sensory nerve endings. *The Journal of physiology*, *61*(1), 49–72.

Alam, F. & Riccardi, G. (2014). Predicting personality traits using multimodal information. In *Proceedings of the 2014 acm multi media on workshop on computational personality recognition* (pp. 15–18).

Alarcao, S. M. & Fonseca, M. J. (2017). Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*.

Albornoz, E. M., Milone, D. H. & Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, *25*(3), 556–570.

Alkawaz, M. H., Basori, A. H., Mohamad, D. & Mohamed, F. (2014). Realistic facial expression of virtual human based on color, sweat, and tears effects. *The Scientific World Journal*, *2014*.

Al-Nafjan, A., Alharthi, K. & Kurdi, H. (2020). Lightweight building of an electroencephalogram-based emotion detection system. *Brain Sciences*, *10*(11), 781.

Alu, D., Zoltan, E. & Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. *Science and Technology*, *20*(3), 222–240.

Atrey, P. K., Hossain, M. A., El Saddik, A. & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, *16*(6), 345–379.

Aung, M. S., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., ... & Elkins, A. C. (2015). The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset. *IEEE transactions on affective computing*, *7*(4), 435–451.

Avent, R. R., Ng, C. T. & Neal, J. A. (1994). Machine vision recognition of facial affect using backpropagation neural networks. In *Proceedings of 16th annual international conference of the ieee engineering in medicine and biology society* (Vol. 2, pp. 1364–1365). IEEE. Retrieved from https://towardsdatascience.com/how-wavenet-works-12e2420ef386

Badshah, A. M., Ahmad, J., Rahim, N. & Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (platcon)* (pp. 1–5). IEEE.

Balaban, J. (2019). How wavenet works. *Towards Data science*. Retrieved Dec 25, from https://towardsdatascience.com/how-wavenet-works-12e2420ef386

Balakrishnan, A. & Rege, A. (n.d.). Reading emotions from speech using deep neural networks. *Stanford Class Report*. Retrieved from web.stanford.edu%2Fclass%2Fcs224s%2Freports%2FAnusha_Balakrishnan.pdf&usg=AOvVaw1tTRdJ1bN-OyJRq20LQBMP

Barnea, O. & Shusterman, V. (1995). Analysis of skin-temperature variability compared to variability of blood pressure and heart rate. In *Proceedings of 17th international conference of the engineering in medicine and biology society* (Vol. 2, pp. 1027–1028). IEEE.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. & Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 2, pp. 568–573). IEEE.

Beck, A., Hiolle, A., Mazel, A. & Cañamero, L. (2010). Interpretation of emotional body language displayed by robots. In *Proceedings of the 3rd international workshop on affective interaction in natural environments* (pp. 37–42). ACM.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T. & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

Bi, G.-q. & Poo, M.-m. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, *18*(24), 10464–10472.

Bohte, S. M. (2004). The evidence for neural information processing with precise spike-times: A survey. *Natural Computing*, *3*(2), 195–206.

Bohte, S. M., Kok, J. N. & La Poutre, H. (2002). Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing*, *48*(1-4), 17–37.

Bohte, S. M., La Poutré, H. & Kok, J. N. (2002). Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer rbf networks. *IEEE Transactions on neural networks*, *13*(2), 426–435.

Borghini, G., Aricò, P., Di Flumeri, G., Cartocci, G., Colosimo, A., Bonelli, S., ... others (2017). Eeg-based cognitive control behaviour assessment: an ecological study with professional air traffic controllers. *Scientific reports*, *7*(1), 1–16.

Bota, P., Wang, C., Fred, A. & Silva, H. (2020). Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? *Sensors*, *20*(17), 4723.

Braitenberg, V. & Schüz, A. (2013). *Cortex: statistics and geometry of neuronal connectivity*. Springer Science & Business Media.

Brette, R. (2015). Philosophy of the spike: rate-based vs. spike-based theories of the brain. *Frontiers in systems neuroscience*, *9*, 151.

Breuer, R. & Kimmel, R. (2017). A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*.

Brigham, T. J. (2017). Merging technology and emotions: Introduction to affective computing. *Medical reference services quarterly*, *36*(4), 399–407.

Bullmore, E. & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, *10*(3), 186.

Buzsaki, G. (2006). *Rhythms of the brain*. Oxford University Press.

Cai, G. & Xia, B. (2015). Convolutional neural networks for multimedia sentiment analysis. In *Natural language processing and chinese computing* (pp. 159–167). Springer.

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

Camurri, A., Lagerlöf, I. & Volpe, G. (2003). Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International journal of human-computer studies*, *59*(1-2), 213–225.

Cao, Y., Chen, Y. & Khosla, D. (2015). Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, *113*(1), 54–66.

Capecci, E., Doborjeh, Z. G., Mammone, N., La Foresta, F., Morabito, F. C. & Kasabov, N. (2016). Longitudinal study of alzheimer's disease degeneration through eeg data analysis with a neucube spiking neural network model. In *2016 international joint conference on neural networks (ijcnn)* (pp. 1360–1366). IEEE.

Cassenaer, S. & Laurent, G. (2007). Hebbian stdp in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. *Nature*, *448*(7154), 709.

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X. & Bai, O. (2017). A fast, efficient domain adaptation technique for cross-domain electroencephalography (eeg)-based emotion recognition. *Sensors*, *17*(5), 1014.

Chan, C., Ginosar, S., Zhou, T. & Efros, A. A. (2019). Everybody dance now. In *Proceedings of the ieee international conference on computer vision* (pp. 5933–5942).

Chanel, G., Kronegg, J., Grandjean, D. & Pun, T. (2006). Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. In *International workshop on multimedia content representation, classification and security* (pp. 530–537). Springer.

Chen, L.-H., Ling, Z.-H. & Dai, L.-R. (2014). Voice conversion using generative trained deep neural networks with multiple frame spectral envelopes. In *Fifteenth annual conference of the international speech communication association.*

Chen, Y., Hu, J., Kasabov, N., Hou, Z. & Cheng, L. (2013). Neucuberehab: A pilot study for eeg classification in rehabilitation practice based on spiking neural networks. In *International conference on neural information processing* (pp. 70–77). Springer.

Chen, Z. J., He, Y., Rosa-Neto, P., Germann, J. & Evans, A. C. (2008). Revealing modular architecture of human brain structural networks by using cortical thickness from mri. *Cerebral cortex*, *18*(10), 2374–2381.

Cho, Y., Bianchi-Berthouze, N. & Julier, S. J. (2017). Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 seventh international conference on affective computing and intelligent interaction (acii)* (pp. 456–463). IEEE.

Cibau, N. E., Albornoz, E. M. & Rufiner, H. L. (2013). Speech emotion recognition using a deep autoencoder. *Anales de la XV Reunion de Procesamiento de la Informacion y Control*, *16*, 934–939.

Cid, F., Moreno, J., Bustos, P. & Núnez, P. (2014). Muecas: a multi-sensor robotic head for affective human robot interaction and imitation. *Sensors*, *14*(5), 7711–7737.

Cireşan, D., Meier, U. & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*.

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

Danelakis, A., Theoharis, T. & Pratikakis, I. (2015). A survey on facial expression recognition in 3d video sequences. *Multimedia Tools and Applications*, *74*(15), 5577–5615.

Darwin, C. & Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Delbruck, T. (2007). *jaer open source project*. Retrieved from http://jaer.wiki.sourceforge.net

De Meijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal behavior*, *13*(4), 247–268.

Deng, J., Zhang, Z., Marchi, E. & Schuller, B. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction* (pp. 511–516). IEEE.

Dhoble, K., Nuntalid, N., Indiveri, G. & Kasabov, N. (2012). Online spatio-temporal pattern recognition with evolving spiking neural networks utilising address event representation, rank order, and temporal spike learning. In *The 2012 international joint conference on neural networks (ijcnn)* (pp. 1–7). IEEE.

Diehl, P. U. & Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience*, *9*, 99.

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C. & Pfeiffer, M. (2015). Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *Neural networks (ijcnn), 2015 international joint conference on* (pp. 1–8). IEEE.

Diehl, P. U., Pedroni, B. U., Cassidy, A., Merolla, P., Neftci, E. & Zarrella, G. (2016). Truehappiness: Neuromorphic emotion recognition on truenorth. In *2016 international joint conference on neural networks (ijcnn)* (pp. 4278–4285). IEEE.

D'mello, S. K. & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 1–36.

Doborjeh, Z., Doborjeh, M., Taylor, T., Kasabov, N., Wang, G. Y., Siegert, R. & Sumich, A. (2019). Spiking neural network modelling approach reveals how mindfulness training rewires the brain. *Scientific reports*, *9*(1), 1–15.

Doborjeh, Z. G., Doborjeh, M. & Kasabov, N. (2018). Eeg pattern recognition using brain-inspired spiking neural networks for modelling human decision processes. In *2018 international joint conference on neural networks (ijcnn)* (pp. 1–7). IEEE.

Doborjeh, Z. G., Kasabov, N., Doborjeh, M. G. & Sumich, A. (2018). Modelling peri-perceptual brain processes in a deep learning spiking neural network architecture. *Scientific reports*, *8*(1), 8912.

Drubach, D. (2000). *The brain explained*. Prentice Hall Health Upper Saddle River, NJ.

Duo, S. & Song, L. X. (2012). An e-learning system based on affective computing. *physics Procedia*, *24*, 1893–1898.

Edwards, J., Jackson, H. J. & Pattison, P. E. (2002). Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review*, *22*(6), 789–832.

Ekman, P. (1992a). Are there basic emotions? *Psychological Review*.

Ekman, P. (1992b). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.

Ekman, P. & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*(1), 88–106.

Ekman, P. & Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, *1*(1), 56–75.

Ekman, P., Friesen, W. V. & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings* (Vol. 11). Elsevier.

El-Abbasy, K., Angelopoulou, A. & Towell, A. (2015). Affective computing to enhance e-learning in segregated societies. In *2015 imperial college computing student workshop (iccsw 2015)* (Vol. 49, pp. 13–20). OpenAccess Series in Informatics.

El Ayadi, M., Kamel, M. S. & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

El Kaliouby, R. & Robinson, P. (2005). Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction* (pp. 181–200). Springer.

Esser, S. K., Appuswamy, R., Merolla, P., Arthur, J. V. & Modha, D. S. (2015). Backpropagation for energy-efficient neuromorphic computing. In *Advances in neural information processing systems* (pp. 1117–1125).

Evans, A. C., Collins, D. L., Mills, S., Brown, E., Kelly, R. & Peters, T. M. (1993). 3d statistical neuroanatomical models from 305 mri volumes. In *Nuclear science symposium and medical imaging conference, 1993., 1993 ieee conference record.* (pp. 1813–1817). IEEE.

Fan, Y., Lu, X., Li, D. & Liu, Y. (2016). Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th acm international conference on multimodal interaction* (pp. 445–450). ACM.

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.-R. & Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural networks*, *22*(9), 1305–1312.

FitzHugh, R. (1961). Fitzhugh-nagumo simplified cardiac action potential model. *Biophys. J*, *1*, 445–466.

Fong, T., Nourbakhsh, I. & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, *42*(3-4), 143–166.

Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, *62*(10), 658–665.

Gachery, S. & Magnenat-Thalmann, N. (2001). Designing mpeg-4 facial animation tables for web applications. *MIRALab, University of Geneva*.

Ganapathy, N., Swaminathan, R. & Deserno, T. M. (2018). Deep learning on 1-d biosignals: a taxonomy-based survey. *Yearbook of medical informatics*, *27*(01), 098–109.

Gargesha, M., Kuchi, P. & Torkkola, I. (2002). Facial expression recognition using artificial neural networks. *Artif. Neural Comput. Syst*, *8*(4), 1–6.

Gautrais, J. & Thorpe, S. (1998). Rate coding versus temporal order coding: a theoretical approach. *Biosystems*, *48*(1-3), 57–65.

Gavrilescu, M. (2015). Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In *2015 23rd telecommunications forum telfor (telfor)* (pp. 720–723). IEEE.

Gerstner, W., Kempter, R., van Hemmen, J. L. & Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, *383*(6595), 76.

Gerstner, W. & Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press.

Gerstner, W., Ritz, R. & Van Hemmen, J. L. (1993). Why spikes? hebbian learning and retrieval of time-resolved excitation patterns. *Biological cybernetics*, *69*(5-6), 503–515.

Ghosh-Dastidar, S. & Adeli, H. (2007). Improved spiking neural networks for eeg classification and epilepsy and seizure detection. *Integrated Computer-Aided Engineering*, *14*(3), 187–212.

Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N. & Lester, J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational data mining 2013*.

Granholm, E. E. & Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*.

Greco, A., Valenza, G., Citi, L. & Scilingo, E. P. (2016). Arousal and valence recognition of affective sounds based on electrodermal activity. *IEEE Sensors Journal*, *17*(3), 716–725.

Grossman, P. & Wientjes, C. J. (2001). How breathing adjusts to mental and physical demands. In *Respiration and emotion* (pp. 43–54). Springer.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X. & Marsic, I. (2018). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *arXiv preprint arXiv:1805.08660*.

Gudi, A., Tasli, H. E., Den Uyl, T. M. & Maroulis, A. (2015). Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th ieee international conference and workshops on automatic face and gesture recognition (fg)* (Vol. 6, pp. 1–5). IEEE.

Gunes, H., Schuller, B., Pantic, M. & Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and gesture 2011* (pp. 827–834). IEEE.

Guo, Y., Tian, Y., Gao, X. & Zhang, X. (2014). Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *2014 international joint conference on neural networks (ijcnn)* (pp. 3473–3479). IEEE.

Guthier, B., Dörner, R. & Martinez, H. P. (2016). Affective computing in games. In *Entertainment computing and serious games* (pp. 402–441). Springer.

Guyonneau, R., VanRullen, R. & Thorpe, S. J. (2005). Neurons tune to the earliest spikes through stdp. *Neural Computation*, *17*(4), 859–879.

Hadjidimitriou, S. K. & Hadjileontiadis, L. J. (2012). Toward an eeg-based recognition of music liking using time-frequency analysis. *IEEE Transactions on Biomedical Engineering*, *59*(12), 3498–3510.

Hajlaoui, A., Chetouani, M. & Essid, S. (2018). Multi-task feature learning for eeg-based emotion recognition using group nonnegative matrix factorization. In *2018 26th european signal processing conference (eusipco)* (pp. 91–95). IEEE.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hanson Robotics. (2019). *The Making of Sophia: Facial recognition, expressions and the loving ai project.* Retrieved Dec 25, from https://www.hansonrobotics.com/the-making-of-sophia-facial-recognition-expressions-and-the-loving-ai-project/

Harper, R. & Southern, J. (2019). End-to-end prediction of emotion from heartbeat data collected by a consumer fitness tracker. In *2019 8th international conference on affective computing and intelligent interaction (acii)* (pp. 1–7). IEEE.

Healey, J. A. (2009). Affect detection in the real world: Recording and processing physiological signals. In *2009 3rd international conference on affective computing and intelligent interaction and workshops* (pp. 1–6). IEEE.

Healey, J. A. & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, *6*(2), 156–166.

Herz, A. V., Gollisch, T., Machens, C. K. & Jaeger, D. (2006). Modeling single-neuron dynamics and computations: a balance of detail and abstraction. *science*, *314*(5796), 80–85.

Hindmarsh, J. L. & Rose, R. (1984). A model of neuronal bursting using three coupled first order differential equations. *Proc. R. Soc. Lond. B*, *221*(1222), 87–102.

Hirsch, H.-G. & Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Asr2000-automatic speech recognition: Challenges for the new millenium isca tutorial and research workshop (itrw)*.

Hjorth, B. (1970). Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, *29*(3), 306–310.

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U. & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, *92*(1-2), 84–89.

Hodgkin, A. L. & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, *117*(4), 500–544.

Homma, I. & Masaoka, Y. (2008). Breathing rhythms and emotions. *Experimental physiology*, *93*(9), 1011–1021.

Hoque, M., McDuff, D. J., Morency, L.-P. & Picard, R. W. (2011). Machine learning for affective computing. In *International conference on affective computing and intelligent interaction* (pp. 567–567). Springer.

Horlings, R., Datcu, D. & Rothkrantz, L. J. (2008). Emotion recognition using brain activity. In *Proceedings of the 9th international conference on computer systems and technologies and workshop for phd students in computing* (pp. II–1).

Hornyak, T. (2018). How humanlike will ai robots like sophia become? *Now. The intersection of technology, innovation & creativity*. Retrieved Dec 25, from https://now.northropgrumman.com/how-humanlike-will-ai-robots-like-sophia-become/

Hu, J., Hou, Z.-G., Chen, Y.-X., Kasabov, N. & Scott, N. (2014). Eeg-based classification of upper-limb adl using snn for active robotic rehabilitation. In *Biomedical robotics and biomechatronics (2014 5th ieee ras & embs international conference on* (pp. 409–414). IEEE.

Hu, X., Chen, J., Wang, F. & Zhang, D. (2019). Ten challenges for eeg-based affective computing. *Brain Science Advances*, *5*(1), 1–20.

Huang, C., Gong, W., Fu, W. & Feng, D. (2014). A research of speech emotion recognition based on deep belief network and svm. *Mathematical Problems in Engineering*, *2014*.

Huang, T.-Y., Li, J.-L., Chang, C.-M. & Lee, C.-C. (2019). A dual-complementary acoustic embedding network learned from raw waveform for speech emotion recognition. In *2019 8th international conference on affective computing and intelligent interaction (acii)* (pp. 83–88). IEEE.

Huang, X., Kortelainen, J., Zhao, G., Li, X., Moilanen, A., Seppänen, T. & Pietikäinen, M. (2016). Multi-modal emotion analysis from facial expressions and electroencephalogram. *Computer Vision and Image Understanding*, *147*, 114–124.

Huang, Y., Yang, J., Liu, S. & Pan, J. (2019). Combining facial expressions and electroencephalography to enhance emotion recognition. *Future Internet*, *11*(5), 105.

Huang, Z., Dong, M., Mao, Q. & Zhan, Y. (2014). Speech emotion recognition using cnn. In *Proceedings of the 22nd acm international conference on multimedia* (pp. 801–804).

Huang, Z.-w., Xue, W.-t. & Mao, Q.-r. (2015). Speech emotion recognition with unsupervised feature learning. *Frontiers of Information Technology & Electronic Engineering*, *16*(5), 358–366.

Hubel, D. H. & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, *148*(3), 574–591.

Huguenard, J. R. (2000). Reliability of axonal propagation: The spike doesn't stop here. *Proceedings of the National Academy of Sciences*, *97*(17), 9349–9350.

Hunsberger, E. & Eliasmith, C. (2015). Spiking deep networks with lif neurons. *arXiv preprint arXiv:1510.08829*.

Ionescu, R. T., Popescu, M. & Grozea, C. (2013). Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on challenges in representation learning, icml*.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks*, *14*(6), 1569–1572.

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE transactions on neural networks*, *15*(5), 1063–1070.

Jacob, V., Brasier, D. J., Erchova, I., Feldman, D. & Shulz, D. E. (2007). Spike timing-dependent synaptic depression in the in vivo barrel cortex of the rat. *Journal of Neuroscience*, *27*(6), 1271–1284.

Jenke, R., Peer, A. & Buss, M. (2014). Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, *5*(3), 327–339.

Jia, X., Li, K., Li, X. & Zhang, A. (2014). A novel semi-supervised deep learning framework for affective state recognition on eeg signals. In *2014 ieee international conference on bioinformatics and bioengineering* (pp. 30–37). IEEE.

Jianbo, S. & Tomasi, C. (1994). Good features to track. In *Ieee computer society conference on computer vision and pattern recognition* (pp. 593–600).

Jourabloo, A. & Liu, X. (2015). Pose-invariant 3d face alignment. In *Proceedings of the ieee international conference on computer vision* (pp. 3694–3702).

Jung, H., Lee, S., Yim, J., Park, S. & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the ieee international conference on computer vision* (pp. 2983–2991).

Jung, T.-P., Sejnowski, T. J. et al. (2018). Multi-modal approach for affective computing. In *2018 40th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 291–294). IEEE.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., … others (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 543–550). ACM.

Kasabov, N. (1998). Evolving fuzzy neural networks-algorithms, applications and biological motivation. *Methodologies for the conception, design and application of soft computing, World Scientific*, *1*, 271–274.

Kasabov, N. (2010). To spike or not to spike: A probabilistic spiking neuron model. *Neural Networks*, *23*(1), 16–19.

Kasabov, N. (2012). Neucube evospike architecture for spatio-temporal modelling and pattern recognition of brain signals. In *Iapr workshop on artificial neural networks in pattern recognition* (pp. 225–243). Springer.

Kasabov, N. (2014). Neucube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. *Neural Networks*, *52*, 62–76.

Kasabov, N. (2017). From multilayer perceptrons and neurofuzzy systems to deep learning machines: Which method to use?-a survey. *International Journal on Information Technologies & Security*, *9*(2).

Kasabov, N. & Capecci, E. (2015). Spiking neural network methodology for modelling, classification and understanding of eeg spatio-temporal data measuring cognitive processes. *Information Sciences*, *294*, 565–575.

Kasabov, N., Dhoble, K., Nuntalid, N. & Indiveri, G. (2013). Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition. *Neural Networks*, *41*, 188–201.

Kasabov, N., Hu, J., Chen, Y., Scott, N. & Turkova, Y. (2013). Spatio-temporal eeg data classification in the neucube 3d snn environment: methodology and examples. In *International conference on neural information processing* (pp. 63–69). Springer.

Kasabov, N., Scott, N. M., Tu, E., Marks, S., Sengupta, N., Capecci, E., … others (2016). Evolving spatio-temporal data machines based on the neucube neuromorphic framework: design methodology and selected applications. *Neural Networks*, *78*, 1–14.

Kasabov, N., Zhou, L., Doborjeh, M. G., Doborjeh, Z. G. & Yang, J. (2016). New algorithms for encoding, learning and classification of fmri data in a spiking neural network architecture: a case on modeling and understanding of dynamic cognitive processes. *IEEE Transactions on Cognitive and Developmental Systems*, *9*(4), 293–303.

Kasabov, N. K. (2007). *Evolving connectionist systems: the knowledge engineering approach*. Springer Science & Business Media.

Kasabov, N. K. (2018a). Evolving spiking neural networks. In *Time-space, spiking neural networks and brain-inspired artificial intelligence* (p. 169-199). Springer.

Kasabov, N. K. (2018b). *Time-space, spiking neural networks and brain-inspired artificial intelligence* (Vol. 7). Springer.

Kasabov, N. K., Doborjeh, M. G. & Doborjeh, Z. G. (2016). Mapping, learning, visualization, classification, and understanding of fmri data in the neucube evolving spatiotemporal data machine of spiking neural networks. *IEEE transactions on neural networks and learning systems*, *28*(4), 887–899.

Katsigiannis, S. & Ramzan, N. (2017). Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, *22*(1), 98–107.

Katsumata, S., Sakai, K., Toujoh, S., Miyamoto, A., Nakai, J., Tsukada, M. & Kojima, H. (2008). Analysis of synaptic transmission and its plasticity by glutamate receptor channel kinetics models and 2-photon laser photolysis. In *Proc. of iconip*.

Kaulard, K., Cunningham, D. W., Bülthoff, H. H. & Wallraven, C. (2012). The mpi facial expression database—a validated database of emotional and conversational facial expressions. *PloS one*, *7*(3), e32321.

Kazemi, V. & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1867–1874).

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A. & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In *Social media and machine learning*. IntechOpen.

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J. & Masquelier, T. (2018). Stdp-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, *99*, 56–67.

Kim, J. & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, *30*(12), 2067–2083.

Kim, K. (2014). Emotion modeling and machine learning in affective computing. *Unpublished manuscript*.

Ko, B. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., ... Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31.

Koelstra, S., Pantic, M. & Patras, I. (2010). A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, *32*(11), 1940–1954.

Koelstra, S. & Patras, I. (2013). Fusion of facial expressions and eeg for implicit affective tagging. *Image and Vision Computing*, *31*(2), 164–174.

Kosti, R., Alvarez, J. M., Recasens, A. & Lapedriza, A. (2017). Emotion recognition in context. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1667–1675).

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

Kukolja, D., Popović, S., Horvat, M., Kovač, B. & Ćosić, K. (2014). Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *International journal of human-computer studies*, *72*(10-11), 717–727.

Kumarasinghe, K., Kasabov, N. & Taylor, D. (2020). Deep learning and deep knowledge representation in spiking neural networks for brain-computer interfaces. *Neural Networks*, *121*, 169–185.

Lan, Z., Sourina, O., Wang, L., Scherer, R. & Müller-Putz, G. R. (2018). Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems*, *11*(1), 85–94.

Lapicque, L. (1907). Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal de Physiologie et de Pathologie Generalej*, *9*, 620–635.

Laybounr, J. (2018). Meet the robot that can mimic human emotion. *CambridgeshireLive*. Retrieved Dec 25, from https://www.cambridge-news.co.uk/news/cambridge-news/cambridge-university-robot-human-emotion-14431300

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., ... Narayanan, S. (2004). Emotion recognition based on phoneme classes. In *Eighth international conference on spoken language processing*.

Lestienne, R. (2001). Spike timing, synchronization and information processing on the sensory side of the central nervous system. *Progress in neurobiology*, *65*(6), 545–591.

Li, L. & Chen, J.-h. (2006). Emotion recognition using physiological signals. In *International conference on artificial reality and telexistence* (pp. 437–446). Springer, Berlin, Heidelberg.

Li, P., Liu, H., Si, Y., Li, C., Li, F., Zhu, X., ... others (2019). Eeg based emotion recognition by combining functional connectivity network and local activations. *IEEE Transactions on Biomedical Engineering*.

Li, R., Liang, Y., Liu, X., Wang, B., Huang, W., Cai, Z., ... Pan, J. (2021). Mindlink-eumpy: An open-source python toolbox for multimodal emotion recognition. *Frontiers in Human Neuroscience*, *15*.

Li, S. & Deng, W. (2018). Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*.

Li, X., Pfister, T., Huang, X., Zhao, G. & Pietikäinen, M. (2013). A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th ieee international conference and workshops on automatic face and gesture recognition (fg)* (pp. 1–6). IEEE.

Lichtsteiner, P. & Delbruck, T. (2005). A 64x64 aer logarithmic temporal derivative silicon retina. In *Research in microelectronics and electronics, 2005 phd* (Vol. 2, pp. 202–205). IEEE.

Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R. & Chen, J.-H. (2010). Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, *57*(7), 1798–1806.

Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O. & Alvarez, K. (2003). Developing multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human-Computer Studies*, *59*(1-2), 245–255.

Lisetti, C. L. & Nasoz, F. (2002). Maui: a multimodal affective user interface. In *Proceedings of the tenth acm international conference on multimedia* (pp. 161–170). ACM.

Liu, J., Su, Y. & Liu, Y. (2017). Multi-modal emotion recognition with temporal-band attention based on lstm-rnn. In *Pacific rim conference on multimedia* (pp. 194–204). Springer.

Liu, J., Su, Y. & Liu, Y. (2018). Multi-modal emotion recognition with temporal-band attention based on lstm-rnn. In B. Zeng, Q. Huang, A. E. Saddik, H. Li, S. Jiang & X. Fan (Eds.), *Pacific rim conference on multimedia* (pp. 194–204). Springer International Publishing.

Liu, P., Han, S., Meng, Z. & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1805–1812).

Liu, S.-C. & Delbruck, T. (2010). Neuromorphic sensory systems. *Current opinion in neurobiology*, *20*(3), 288–295.

Liu, Y., Sourina, O. & Nguyen, M. K. (2011). Real-time eeg-based emotion recognition and its applications. In *Transactions on computational science xii* (pp. 256–277). Springer.

Lucas, B. D., Kanade, T. et al. (1981). An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia.

Luneski, A., Konstantinidis, E. & Bamidis, P. (2010). Affective medicine. *Methods of information in medicine*, *49*(03), 207–218.

Maass, W. (1997a). Fast sigmoidal networks via spiking neurons. *Neural Computation*, *9*(2), 279–304.

Maass, W. (1997b). Networks of spiking neurons: the third generation of neural network models. *Neural networks*, *10*(9), 1659–1671.

Maass, W. & Markram, H. (2004). On the computational power of circuits of spiking neurons. *Journal of computer and system sciences*, *69*(4), 593–616.

Mansouri-Benssassi, E. & Ye, J. (2021). Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural networks. *Soft Computing*, *25*(3), 1717–1730.

Mastebroek, H. A., Vos, J. E. & Vos, J. (2001). *Plausible neural networks for biological modelling* (Vol. 13). Springer Science & Business Media.

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J. & Kaliouby, R. e. (2016). Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 3723–3726). ACM.

Meftah, B., Lezoray, O. & Benyettou, A. (2010). Segmentation and edge detection based on spiking neural network model. *Neural Processing Letters*, *32*(2), 131–146.

Mehrabian, A. (1968). Communication without words. *Psychology today*, *2*(4).

Merolla, P., Arthur, J., Akopyan, F., Imam, N., Manohar, R. & Modha, D. S. (2011). A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm. In *Custom integrated circuits conference (cicc), 2011 ieee* (pp. 1–4). IEEE.

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., ... Modha, D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, *345*(6197), 668–673.

Mijić, I., Šarlija, M. & Petrinović, D. (2019). Mmod-cog: A database for multimodal cognitive load classification. In *2019 11th international symposium on image and signal processing and analysis (ispa)* (pp. 15–20). IEEE.

Mohemmed, A., Schliebs, S., Matsuda, S. & Kasabov, N. (2011). Method for training a spiking neuron to associate input-output spike trains. In *Engineering applications of neural networks* (pp. 219–228). Springer.

Morris, C. & Lecar, H. (1981). Voltage oscillations in the barnacle giant muscle fiber. *Biophysical journal*, *35*(1), 193–213.

Mosca, A. (2000). A review essay on antonio damasio's the feeling of what happens: Body and emotion in the making of consciousness. *Psyche*, *6*(10), 1–13. doi: https://doi.org/10.1353/jsp.2001.0038

Mu, Y. & Poo, M.-m. (2006). Spike timing-dependent ltp/ltd mediates visual experience-dependent plasticity in a developing retinotectal system. *Neuron*, *50*(1), 115–125.

Murphy-Chutorian, E. & Trivedi, M. M. (2008). Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *31*(4), 607–626.

Nakasone, A., Prendinger, H. & Ishizuka, M. (2005). Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th international workshop on biosignal interpretation* (pp. 219–222). Citeseer.

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

Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S. & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE transactions on affective computing*.

Nuntalid, N., Dhoble, K. & Kasabov, N. (2011). Eeg classification with bsa spike encoding algorithm and evolving probabilistic spiking neural network. In *International conference on neural information processing* (pp. 451–460). Springer.

Nwe, T. L., Foo, S. W. & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, *41*(4), 603–623.

Nykliček, I., Thayer, J. F. & Van Doornen, L. J. (1997). Cardiorespiratory differentiation of musically-induced emotions. *Journal of Psychophysiology*.

O'Connor, P., Neil, D., Liu, S.-C., Delbruck, T. & Pfeiffer, M. (2013). Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in neuroscience*, *7*, 178.

O'reilly, C., Gosselin, N., Carrier, J. & Nielsen, T. (2014). Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, *23*(6), 628–635.

Othman, M., Kasabov, N., Tu, E., Feigin, V., Krishnamurthi, R., Hou, Z., ... Hu, J. (2014). Improved predictive personalized modelling with the use of spiking

neural network system and a case study on stroke occurrences data. In *Neural networks (ijcnn), 2014 international joint conference on* (pp. 3197–3204). IEEE.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62–66.

Ouyang, W. & Wang, X. (2013). Joint deep learning for pedestrian detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2056–2063).

Pan, J. & Tompkins, W. J. (1985). A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*(3), 230–236.

Pang, B., Lee, L. et al. (2006). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *1*(2), 91–231. doi: https://doi.org/10.1561/1500000001

Pantic, M. & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390.

Parhi, K. K. & Unnikrishnan, N. K. (2020). Brain-inspired computing: Models and architectures. *IEEE Open Journal of Circuits and Systems*, *1*, 185–204.

Partala, T., Jokiniemi, M. & Surakka, V. (2000). Pupillary responses to emotionally provocative stimuli. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 123–129).

Paulun, L., Wendt, A. & Kasabov, N. K. (2018). A retinotopic spiking neural network system for accurate recognition of moving objects using neucube and dynamic vision sensors. *Frontiers in Computational Neuroscience*, *12*, 42.

Petro, B., Kasabov, N. & Kiss, R. M. (2019). Selection and optimization of temporal spike encoding methods for spiking neural networks. *IEEE transactions on neural networks and learning systems*, *31*(2), 358–370.

Piana, S., Stagliano, A., Odone, F., Verri, A. & Camurri, A. (2014). Real-time automatic emotion recognition from body gestures. *arXiv preprint arXiv:1402.5047*.

Picard, R. W. (1995). Affective computing-mit media laboratory perceptual computing section technical report no. 321. *Cambridge, MA*, *2139*.

Picard, R. W. & Scheirer, J. (2001). The galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings 9th int. conf. on hci.*

Picard, R. W., Vyzas, E. & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, *23*(10), 1175–1191.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, *89*(4), 344–350.

Poria, S., Cambria, E., Bajpai, R. & Hussain, A. (2017a). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Poria, S., Cambria, E., Bajpai, R. & Hussain, A. (2017b). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Poria, S., Cambria, E., Hussain, A. & Huang, G.-B. (2015). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, *63*, 104–116.

Poria, S., Peng, H., Hussain, A., Howard, N. & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, *261*, 217–230.

Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K. C., Dimitropoulos, K. & Daras, P. (2016). Multimodal affective state recognition in serious games applications. In *2016 ieee international conference on imaging systems and techniques (ist)* (pp. 435–439). IEEE.

Ranganathan, H., Chakraborty, S. & Panchanathan, S. (2016). Multimodal emotion recognition using deep learning architectures. In *2016 ieee winter conference on applications of computer vision (wacv)* (pp. 1–9). IEEE.

Rani, P., Liu, C., Sarkar, N. & Vanman, E. (2006). An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications*, *9*(1), 58–69.

Rank, E. & Pirker, H. (1998). Generating emotional speech with a concatenative synthesizer. In *Fifth international conference on spoken language processing*.

Rao, K. S. (2011). Role of neural network models for developing speech systems. *Sadhana*, *36*(5), 783–836.

Rifai, S., Bengio, Y., Courville, A., Vincent, P. & Mirza, M. (2012). Disentangling factors of variation for facial expression recognition. In *European conference on computer vision* (pp. 808–822). Springer.

Rivera-Hernández, R., Stoyanov, D., Tsolaki, M. & Ramón, G. L. (2013). Affective computing applied in elderly with depression and alzheimer's disease. *Psychopathology: Theory, Perspectives and Future Approaches*.

Roffo, G. (2016). Feature selection library (matlab toolbox). *arXiv preprint arXiv:1607.01327*.

Roffo, G., Melzi, S. & Cristani, M. (2015). Infinite feature selection. In *Proceedings of the ieee international conference on computer vision* (pp. 4202–4210).

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rudovic, O. O. (2016). Machine learning for affective computing and its applications to automated measurement of human facial affect. In *2016 international*

symposium on micro-nanomechatronics and human science (mhs) (pp. 1–1). IEEE.

Rueckauer, B., Lungu, I.-A., Hu, Y. & Pfeiffer, M. (2016). Theory and tools for the conversion of analog to spiking convolutional neural networks. *arXiv preprint arXiv:1612.04052*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei1, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Russell, J. A. (1979). Affective space is bipolar. *Journal of personality and social psychology*, *37*(3), 345.

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

Russell, J. A., Lewicka, M. & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of personality and social psychology*, *57*(5), 848.

Saha, S., Datta, S., Konar, A. & Janarthanan, R. (2014). A study on emotion recognition from body gestures using kinect sensor. In *2014 international conference on communication and signal processing* (pp. 056–060). IEEE.

Santhoshkumar, R. & Geetha, M. K. (2019). Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks. *Procedia Computer Science*, *152*, 158–165.

Sarkar, C., Bhatia, S., Agarwal, A. & Li, J. (2014). Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 acm multi media on workshop on computational personality recognition* (pp. 11–14).

Sarkar, P. & Etemad, A. (2019). Self-supervised learning for ecg-based emotion recognition. *arXiv preprint arXiv:1910.07497*.

Šarlija, M., Jurišić, F. & Popović, S. (2017). A convolutional neural network based approach to qrs detection. In *Image and signal processing and analysis (ispa), 2017 10th international symposium on* (pp. 121–125). IEEE.

Scheirer, J., Fernandez, R. & Picard, R. W. (1999). Expression glasses: a wearable device for facial expression recognition. In *Chi'99 extended abstracts on human factors in computing systems* (pp. 262–263).

Schiano, D. J., Ehrlich, S. M., Rahardja, K. & Sheridan, K. (2000). Face to interface: facial affect in (hu) man and machine. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 193–200). ACM.

Schuller, B., Rigoll, G. & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. I–577). IEEE.

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90–99.

Schwenker, F., Boeck, R., Schels, M., Meudt, S., Siegert, I., Glodek, M., ... Krell, G. (2017). Multimodal affect recognition in the context of human-computer interaction for companion-systems. *Companion technology: a paradigm shift in human-technology interaction*, 387–408.

Sengupta, N. & Kasabov, N. (2017). Spike-time encoding as a data compression technique for pattern recognition of temporal data. *Information Sciences*, *406*, 133–145.

Sengupta, N., McNabb, C. B., Kasabov, N. & Russell, B. R. (2018). Integrating space, time, and orientation in spiking neural networks: A case study on multimodal brain data modeling. *IEEE Transactions on Neural Networks and Learning Systems*(99), 1–15.

Shami, M. T. & Kamel, M. S. (2005). Segment-based approach to the recognition of emotions in speech. In *2005 ieee international conference on multimedia and expo* (pp. 4–pp). IEEE.

Shi, J. et al. (1994). Good features to track. In *1994 proceedings of ieee conference on computer vision and pattern recognition* (pp. 593–600). IEEE.

Shon, S., Ali, A. & Glass, J. (2017). Mit-qcri arabic dialect identification system for the 2017 multi-genre broadcast challenge. In *2017 ieee automatic speech recognition and understanding workshop (asru)* (pp. 374–380). IEEE.

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., ... Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors*, *18*(7), 2074.

Siddharth, S., Jung, T.-P. & Sejnowski, T. J. (2019). Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Transactions on Affective Computing*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484.

Simard, D., Nadeau, L. & Kröger, H. (2005). Fastest learning in small-world neural networks. *Physics Letters A*, *336*(1), 8–15.

Soleymani, M., Lichtenauer, J., Pun, T. & Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*(1), 42–55.

Soleymani, M., Lichtenauer, J., Pun, T. & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*(1), 42–55.

Soleymani, M., Villaro-Dixon, F., Pun, T. & Chanel, G. (2017). Toolbox for emotional feature extraction from physiological signals (teap). *Frontiers in ICT*, *4*, 1.

Song, S., Miller, K. D. & Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, *3*(9), 919.

Song, T., Lu, G. & Yan, J. (2020). Emotion recognition based on physiological signals using convolution neural networks. In *Proceedings of the 2020 12th international conference on machine learning and computing* (pp. 161–165).

Sreeja, P. & Mahalakshmi, G. (2017). Emotion models: a review. *International Journal of Control Theory and Applications*, *10*(8), 651–657.

Sreeja, P. & Mahalaksmi, G. (2015). Applying vector space model for poetic emotion recognition. *Advances in Natural and Applied Sciences*, *9*(6 SE), 486–491.

Stam, C. J. (2004). Functional connectivity patterns of human magnetoencephalographic recordings: a 'small-world'network? *Neuroscience letters*, *355*(1-2), 25–28.

Sun, W., Zhao, H. & Jin, Z. (2017). An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing*, *267*, 385–395.

Szirtes, G., Orozco, J., Petrás, I., Szolgay, D., Utasi, Á. & Cohn, J. F. (2017). Behavioral cues help predict impact of advertising on future sales. *Image and Vision Computing*, *65*, 49–57. Retrieved from https://doi.org/10.1016/j.imavis.2017.03.002

Taherkhani, A., Belatreche, A., Li, Y., Cosma, G., Maguire, L. P. & McGinnity, T. (2019). A review of learning in biologically plausible spiking neural networks. *Neural Networks*.

Taherkhani, A., Belatreche, A., Li, Y., Cosma, G., Maguire, L. P. & McGinnity, T. M. (2020). A review of learning in biologically plausible spiking neural networks. *Neural Networks*, *122*, 253–272.

Talairach, J. & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system. *An approach to cerebral imaging*.

Tan, C., Ceballos, G., Kasabov, N. & Puthanmadam Subramaniyam, N. (2020). Fusionsense: Emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network. *Sensors*, *20*(18), 5328.

Tan, C., Šarlija, M. & Kasabov, N. (2020). Spiking neural networks: Background, recent development and the neucube architecture. *Neural Processing Letters*, 1–27.

Tan, C., Šarlija, M. & Kasabov, N. (2021). Neurosense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio-temporal eeg patterns. *Neurocomputing*, *434*, 137–148.

Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*.

Tao, J. & Tan, T. (2005). Affective computing: A review. In *International conference on affective computing and intelligent interaction* (pp. 981–995). Springer.

Tavanaei, A. & Maida, A. S. (2017). A spiking network that learns to extract spike signatures from speech signals. *Neurocomputing*, *240*, 191–199.

Taylor, D., Scott, N., Kasabov, N., Capecci, E., Tu, E., Saywell, N., ... Hou, Z.-G. (2014). Feasibility of neucube snn architecture for detecting motor execution and motor intention for use in bciapplications. In *Neural networks (ijcnn), 2014 international joint conference on* (pp. 3221–3225). IEEE.

Thorpe, S. & Gautrais, J. (1998). Rank order coding. In *Computational neuroscience* (pp. 113–118). Springer.

Thorpe, S. J. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. *Parallel processing in neural systems*, 91–94.

Torres-Valencia, C., Álvarez-López, M. & Orozco-Gutiérrez, Á. (2016). Svm-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, *1*(11), 9–23.

Torres-Valencia, C., Álvarez-López, M. & Orozco-Gutiérrez, Á. (2017). Svm-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, *11*(1), 9–23.

Trentin, E., Schwenker, F., El Gayar, N. & Abbas, H. M. (2018). Off the mainstream: Advances in neural networks and machine learning for pattern recognition. *Neural Processing Letters*, *48*(2), 643–648.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B. & Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5200–5204). IEEE.

Tzirakis, P., Zhang, J. & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5089–5093). IEEE.

Uddin, M. Z., Hassan, M. M., Almogren, A., Alamri, A., Alrubaian, M. & Fortino, G. (2017). Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access*, *5*, 4525–4536.

Valstar, M. & Pantic, M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd intern. workshop on emotion (satellite of lrec): Corpora for research on emotion and affect* (p. 65). Paris, France.

Ververidis, D. & Kotropoulos, C. (2005). Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *2005 ieee international conference on multimedia and expo* (pp. 1500–1503). IEEE.

Viola, P., Jones, M. et al. (2001). Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, *1*(511-518), 3.

Vitay, J. & Hamker, F. H. (2011). A neuroscientific view on the role of emotions in behaving cognitive agents. *KI-Künstliche Intelligenz*, *25*(3), 235–244. doi: 10.1007/s13218-011-0106-y

Walk, R. D. & Homan, C. P. (1984). Emotion and dance in dynamic light displays. *Bulletin of the Psychonomic Society*, *22*(5), 437–440.

Wang, S., Zhu, Y., Wu, G. & Ji, Q. (2014). Hybrid video emotional tagging using users' eeg and video content. *Multimedia tools and applications*, *72*(2), 1257–1283.

Wang, S.-H., Phillips, P., Dong, Z.-C. & Zhang, Y.-D. (2018). Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm. *Neurocomputing*, *272*, 668–676.

Wang, W., Pedretti, G., Milo, V., Carboni, R., Calderoni, A., Ramaswamy, N., … Ielmini, D. (2018). Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses. *Science advances*, *4*(9), eaat4752.

Wang, Y., See, J., Phan, R. C.-W. & Oh, Y.-H. (2015). Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PloS one*, *10*(5), e0124674.

Wei, W., Jia, Q. & Chen, G. (2016). Real-time facial expression recognition for affective computing based on kinect. In *2016 ieee 11th conference on industrial electronics and applications (iciea)* (pp. 161–165). IEEE.

Wen, G., Li, H., Huang, J., Li, D. & Xun, E. (2017). Random deep belief networks for recognizing emotions from speech signals. *Computational intelligence and neuroscience*, *2017*.

Weymar, M. & Schwabe, L. (2016). Amygdala and emotion: the bright side of it. *Frontiers in neuroscience*, *10*, 224.

Wiem, M. B. H. & Lachiri, Z. (2017). Emotion classification in arousal valence model using mahnob-hci database. *International Journal of Advanced Computer Science and Applications*, *8*(3).

Wilson, C. & Callaway, J. (2000). Coupled oscillator model of the dopaminergic neuron of the substantia nigra. *Journal of neurophysiology*, *83*(5), 3084–3100.

Wolfkühler, W., Majorek, K., Tas, C., Küper, C., Saimed, N., Juckel, G. & Brüne, M. (2012). Emotion recognition in pictures of facial affect: Is there a difference between forensic and non-forensic patients with schizophrenia? *The European Journal of Psychiatry*, *26*(2), 73–85.

Wu, C.-H., Huang, Y.-M. & Hwang, J.-P. (2016). Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology*, *47*(6), 1304–1323.

Wu, Q., Shen, X. & Fu, X. (2011). The machine knows what you are hiding: an automatic micro-expression recognition system. In *international conference on affective computing and intelligent interaction* (pp. 152–162). Springer.

Wysoski, S. G., Benuskova, L. & Kasabov, N. (2010). Evolving spiking neural networks for audiovisual information processing. *Neural Networks*, *23*(7), 819–835.

Xu, J., Broekens, J., Hindriks, K. & Neerincx, M. A. (2014). Robot mood is contagious: effects of robot body language in the imitation game. In *Proceedings of the 2014 international conference on autonomous agents and multi-agent systems* (pp. 973–980). International Foundation for Autonomous Agents and Multiagent Systems.

Xu, J., Broekens, J., Hindriks, K. & Neerincx, M. A. (2015). Mood contagion of robot body language in human robot interaction. *Autonomous Agents and Multi-Agent Systems*, *29*(6), 1216–1248.

Yadegaridehkordi, E., Noor, N. F. B. M., Ayub, M. N. B., Affal, H. B. & Hussin, N. B. (2019). Affective computing in education: A systematic review and future research. *Computers & Education*, *142*, 103649.

Yamasaki, T., Fukushima, Y., Furuta, R., Sun, L., Aizawa, K. & Bollegala, D. (2015). Prediction of user ratings of oral presentations using label relations. In *Proceedings of the 1st international workshop on affect & sentiment in multimedia* (pp. 33–38).

Yeasin, M., Bullot, B. & Sharma, R. (2006). Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, *8*(3), 500–508.

Yeni-Komshian, G. H. & Benson, D. A. (1976). Anatomical study of cerebral asymmetry in the temporal lobe of humans, chimpanzees, and rhesus monkeys. *Science*, *192*(4237), 387–389.

Zacharatos, H., Gatzoulis, C. & Chrysanthou, Y. L. (2014). Automatic emotion recognition based on body movement analysis: a survey. *IEEE computer graphics and applications*, *34*(6), 35–45.

Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y. & Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, *273*, 643–649.

Zhang, Q., Chen, X., Zhan, Q., Yang, T. & Xia, S. (2017). Respiration-based emotion recognition with deep learning. *Computers in Industry*, *92*, 84–90.

Zhang, W., Yin, Z., Sun, Z., Tian, Y. & Wang, Y. (2020). Selecting transferrable neurophysiological features for inter-individual emotion recognition via a shared-subspace feature elimination approach. *Computers in Biology and Medicine*, *123*, 103875.

Zhang, X., Xu, C., Xue, W., Hu, J., He, Y. & Gao, M. (2018). Emotion recognition based on multichannel physiological signals with comprehensive nonlinear processing. *Sensors*, *18*(11), 3886.

Zhao, K., Chu, W.-S. & Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3391–3399).

Zheng, W.-L. & Lu, B.-L. (2015). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175.

Zheng, W.-L., Zhu, J.-Y. & Lu, B.-L. (2017). Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions on Affective Computing*.

Zheng, W.-L., Zhu, J.-Y., Peng, Y. & Lu, B.-L. (2014). Eeg-based emotion classification using deep belief networks. In *2014 ieee international conference on multimedia and expo (icme)* (pp. 1–6). IEEE.

Zhong, B., Qin, Z., Yang, S., Chen, J., Mudrick, N., Taub, M., ... Lobaton, E. (2017). Emotion recognition with facial expressions and physiological signals. In *2017 ieee symposium series on computational intelligence (ssci)* (pp. 1–8). IEEE.

Zhu, X., Lei, Z., Liu, X., Shi, H. & Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 146–155).

Zhu, Y., Wang, S. & Ji, Q. (2014). Emotion recognition from users' eeg signals with the help of stimulus videos. In *2014 ieee international conference on multimedia and expo (icme)* (pp. 1–6). IEEE.