



الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

جامعة غرداية

University of Ghardaia

Registration number

كلية العلوم والتكنولوجيا

/...../...../...../...../

Faculty of Science and Technology

قسم الرياضيات الإعلام لآلي

Department of Mathematics and Computer Science

Memory

For obtaining the master's degree

Field: Mathematics and Computer Science

Sector: computer science

Specialty: Intelligent Systems for Knowledge Extraction

Theme : **Emotion recognition in video using deep learning**

Presented by:

Zohra Hiba

Soumia Taleb Ahmed

Examined by the jury composed of

M: **Slimane BELLAOUAR**

MCA

Univ.Ghardaia

President

M: **Abderrahmane ADJILA**

MAA

Univ.Ghardaia

Examiner

M: **Slimane OULADNAOUI**

MCB

Univ.Ghardaia

Supervisor

University Year 2020/2021



Dedicace



I am dedicating this thesis :
To my dear mother and father.
for their moral support, their encouragement throughout my studies, they
gave everything they could for me for nothing in return.
To my brothers and sisters
Who are my support in life and are waiting for my success.
to all my family
Those who love me and are happy with my success.
To all my friends and everyone who loves me
Thank you for your constant encouragement.

Soumia





Dedicace



I am dedicating this thesis :

To my loving parents.

whose love for me knew no bounds and, who taught me the value of hard work !

To my grandparents

Although they are no longer of this world, their memories continue in my life.



To my sisters and brothers

who has been a constant source of support and encouragement during the challenges of graduate

to all beloved people

who have meant and continue to mean so much to me

Zohra



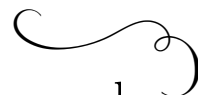
Acknowledgements



First and foremost we are extremely grateful to our supervisors Mr, **OULAD NAOUI SLIMANE** for his invaluable supervision, support and tutelage as we do this humble work, you have all our praise and appreciation.



We also extend our sincere thanks to the members of the jury for for their interest our modest work , as it took their time and effort.



We thank everyone who encouraged us to complete this work.



Abstract

Affective computing aims to implement methods and technologies to recognize and synthesize human emotions. Understanding human facial expressions is essential to the success of this new branch of AI.

Emotions can be conveyed through various channels, the most prominent are facial expressions, speech, texts and various other physiological signals. This topic has occupied researchers for a long time due to the difficulty of understanding and categorizing these expressions.

In this work, we explore the different techniques carried out to recognize facial emotions in videos. We experiment on the AFEW dataset with two models based on deep learning. The first uses TCNs and the second uses CNNs.

The experience with the first model was very hard since it belongs to recent sequential models and was not completed due to difficulty of implementation and limited resources. The second model achieved good accuracy of up to 91%.

Keywords: Facial Emotion recognition, Deep Learning, TCN, CNN, AFEW database.

ملخص

تهدف الحوسبة العاطفية إلى تنفيذ أساليب وتقنيات للتعرف على المشاعر البشرية وتوليدها. يعد فهم تعابير الوجه البشري أمراً ضرورياً لنجاح هذا الفرع الجديد من الذكاء الاصطناعي. يمكن نقل المشاعر من خلال مجموعة متنوعة من القنوات، أبرزها تعابير الوجه، الصوت، النصوص ومختلف الإشارات الفسيولوجية الأخرى. شغل هذا الموضوع الباحثين لفترة طويلة بسبب صعوبة فهم هذه التعبيرات وتصنيفها .

في هذا العمل نستكشف الأعمال المختلفة التي تم تنفيذها للتعرف على مشاعر الوجه في مقاطع الفيديو. نجرب نموذجين يعتمدان على التعلم العميق ، الأول باستخدام الشبكة العصبية التلافيفية المتزامنة *TCN* والثاني باستخدام الشبكة العصبية التلافيفية العادية *CNN* . تم إجراء الاختبارات على مجموعة بيانات *AFEW* .

لقد كانت التجربة مع النموذج الأول صعبة للغاية نظرًا لأنه ينتمي إلى النماذج المتسلسلة الحديثة ولم يكتمل بسبب صعوبة التنفيذ والموارد المحدودة. في حين حقق النموذج الثاني دقة جيدة تصل إلى ٩١% .

كلمات مفتاحية

التعرف على مشاعر الوجه ، التعلم العميق، الشبكة العصبية التلافيفية المتزامنة ، الشبكة العصبية التلافيفية ، قاعدة بيانات *AFEW* .

Résumé

L'informatique affective vise à mettre en oeuvre des méthodes et des technologies permettant de reconnaître et synthétiser les émotions humaines. Bien comprendre les expressions faciales humaines est un élément essentiel dans la réussite de cette nouvelle branche de l'IA.

Les émotions peuvent être véhiculées par divers canaux dont les plus prépondérants sont les expressions faciales, la voix, les textes, et divers autres signaux physiologiques. Ce sujet a occupé les chercheurs depuis longtemps en raison de la difficulté de comprendre et de catégoriser ces expressions.

Dans ce travail, nous explorons les différents travaux menés pour reconnaître les émotions faciales dans les vidéos. Nous expérimentons deux modèles basés sur l'apprentissage profond. Le premier utilise les TCN et le deuxième les CNN. Les tests ont porté sur l'ensemble de données AFEW.

L'expérience avec le premier modèle a été très dure étant donné qu'il appartient aux modèles séquentiels récents et n'a pas été achevée pour difficulté d'implémentation et limite de nos ressources. Le deuxième modèle a atteint une bonne précision allant jusqu'à 91%.

Mots clés : Reconnaissance des émotions faciales, Apprentissage profond, TCN, CNN, Base de données AFEW.

CONTENTS

List of Figures	iv
List of Tables	vi
List of Abbreviations	vi
Introduction	1
1 Background	3
1.1 Introduction	3
1.2 Artificial intelligence, Machine learning and Deep learning	3
1.2.1 Artificial Intelligence	3
1.2.2 Machine learning	4
1.2.3 Deep Learning	5
1.3 Emotions and their models	13
1.3.1 Definition	13
1.3.2 The Basic Emotion Model	14
1.3.3 The Dimensional Model	16
1.3.4 The Componential Appraisal Model	17
1.4 Emotion Recognition	18
1.4.1 Feature of emotion	18
1.4.2 Facial Action Coding System (FACS)	21
1.5 Applications	22
1.5.1 Software engineering	22

1.5.2	Education and e-education	22
1.5.3	Websites customization	23
1.5.4	Health care	23
1.5.5	Automotive industry	23
1.5.6	Video game	24
1.5.7	Others	24
1.6	Conclusion	24
2	State of the art	25
2.1	Introduction	25
2.2	Earlier work	25
2.3	Machine Learning based approaches	26
2.3.1	Face Tracking and Detection	26
2.3.2	Feature Extraction	31
2.3.3	Emotion Recognition/Classification	37
2.4	Deep Learning-based Approaches	41
2.4.1	Convolutional Neural networks (CNN)	42
2.4.2	3D-CNN	42
2.4.3	CNN with RNN	43
2.4.4	TCN	44
2.5	Datasets	45
2.5.1	C-K (Cohn-Kanade) database	45
2.5.2	Cohn-Kanade Dataset (CK+)	45
2.5.3	Japanese Female Facial Expressions (JAFFE)	46
2.5.4	MMI dataset	46
2.5.5	IEMOCAP dataset	46
2.6	Evaluation of emotion recognition system	46
2.7	Conclusion	47
3	Experiment	48
3.1	Introduction	48
3.2	Network Architecture	48
3.3	Implementation Setup	53
3.3.1	Dataset	53
3.3.2	Data processing	54
3.3.3	Environment	56

3.4 Results and Discussion	58
3.5 Conclusion	60
4 Conclusion	62

LIST OF FIGURES

1.1	Biological neuron (Liu, 2020)	6
1.2	Artificial neuron (Liu, 2020)	6
1.3	Simple perceptron (Gupta and Raza, 2019)	6
1.4	Before and after convolution filter applied to an image (Kölbl, 2017)	7
1.5	CNN classification process	7
1.6	Typical architecture of 3D CNN (Singh et al., 2020)	8
1.7	The architecture VRNN (Saud and Shakya, 2020)	9
1.8	Architecture of LSTM (Saud and Shakya, 2020)	10
1.9	An example of a combined architecture of CNN and RNN (Huang et al., 2017)	11
1.10	Differences between (a) standard convolutional network, (b) causal convolutional network, and (c) dilated causal convolutional network (Lara-Benítez et al., 2020)	12
1.11	The six basic emotions (Yao, 2014)	14
1.12	Plutchik’s wheel of emotions (Plutchik, 1982)	15
1.13	Basic emotions on the Valance Arousal Dimensional Model (Jerritta et al., 2011)	17
1.14	Anger	18
1.15	Fear	19
1.16	Disgust	19
1.17	Happiness	20
1.18	Sadness	20
1.19	Surprise	21
1.20	Some examples of AUs (upper and lower face) adapted from (Ko, 2018)	21
2.1	Procedure in conventional face expression approaches. (Khan, 2013)	27
2.2	KLT algorithm flow chart used in (Boda et al., 2016)	28
2.3	frame 0 (Wagener and Herbst, 2002)	28
2.4	frame 1 (Wagener and Herbst, 2002)	28

2.5	frame 2 (Wagener and Herbst, 2002)	28
2.6	frame 3 (Wagener and Herbst, 2002)	28
2.7	frame 4 (Wagener and Herbst, 2002)	29
2.8	frame 5 (Wagener and Herbst, 2002)	29
2.9	The facial mesh and the associated 16 Bézier volumes (Tao and Huang, 2002)	29
2.10	Results of the real-time tracking system (Tao and Huang, 2002)	30
2.11	Person-Spotter's model graph and background suppression (Steffens et al., 1998)	30
2.12	Face detection using the VJ algorithm (Alionte and Lazar, 2015)	31
2.13	Other example for face detection using the VJ algorithm (Abdulsalam et al., 2019)	31
2.14	facial landmarks in Pantic work (Khan, 2013)	32
2.15	Gabor filter transformation (Lyons et al., 1998)	33
2.16	The basic LBP operator (Ahonen et al., 2004)	34
2.17	Extraction of LBP histogram from a facial image (Huang et al., 2019)	34
2.18	The basic haar-like feature template (Khan, 2013)	35
2.19	The process of Haar-like feature extraction (Khan, 2013)	35
2.20	Feature point displacement (Cohn et al., 1998).	36
2.21	Applications of optical flow-based methods on facial images (Sánchez et al., 2011)	36
2.22	A network structure of TCN-based model (Yang and Liu, 2019)	45
3.1	Visualization of a stack of dilated causal convolutional layers	49
3.2	Non-Causal TCN - ks = 3, dilations = [1, 2, 4, 8], 1 block	50
3.3	TCN Architecture arguments	50
3.4	CNN Architecture	51
3.5	CNN architecture part 1	51
3.6	CNN architecture part 2	52
3.7	CNN architecture part 3	52
3.8	CNN architecture part 4	53
3.9	One sampled frame from each of the 7 classes in AFEW	54
3.10	Train dataset	55
3.11	Test dataset	55
3.12	Data processing code	56
3.13	anaconda navigator	56
3.14	jupyter notebook in Google Colab	57
3.15	jupyter notebook interface	57
3.16	The training and test accuracy curve	59
3.17	The training and test loss curve	59
3.18	The confusion matrix for the	60
3.19	Precision,Recall,F1-score,Support	60

LIST OF TABLES

1.1	Selected lists of basic emotions (Sumpeno et al., 2011)	16
1.2	The prototypical AUs observed in basic and compound emotion category, proposed in (Fabian Benitez-Quiroz et al., 2016)	22
2.1	Subject-independent comparison with AlexNet results (accuracy%) (Mollahosseini et al., 2016)	42
3.1	AFEW database attributes (Dhall et al., 2012)	54

LIST OF ABBREVIATIONS

AI	Artificial intelligence
ML	Machine learning
DL	Deep learning
GPU	Graphics Processing Units
CPU	Central Processing Units
NLP	Natural Language Processing
ANN	Artificial Neural Networks
SLP	Single-Layer Neural Networks
MLP	Multiple Layered Perceptrons
CNN	convolutional Neural Networks
3D-CNN	3D Convolutional Neural Networks
RNN	Recurrent Neural Networks
TCN	Temporal Convolutional Network
VRNN	Vanilla Recurrent Neural Networks
LSTM	Long Short-Term Memory
FLs	Facial Landmarks
FACs	Facial Action Coding System

AUs	Facial Action Units (AUs)
LBP	Local Binary Pattern
SVM	Support Vector Machine
ASMs	Active Shape Models
AMM	Active Appearance Mode
AR	Attention-Rejection
PU	Pleasantness-Unpleasantness
KLT	Kanade-Lucas-Tomasi tracking algorithm
PBVD	Piecewise Bezier Volume Deformation tracker
VJ	Viola Jones algorithm
HS	Horn-Schunck optical flow
SIFT	Scale Invariant Feature Transform
DT	Decision Trees
HMM	Hidden Markov Models
Adaboost	Adaptive Boosting
KNN	k-Nearest Neighbours
ARLBP	Symmetric Region Local Binary Pattern method
TAN	Tree-Augmented-Naive Bayes
MMI	Man-Machine Interface
MCGM	Multi-Channel Gradient-Model
FAR	False Acceptance Rate
ASDs	Autism Spectrum Disorders
PCA	Principal Component Analysis
JAFFE	Japanese Female Facial Expression
MUFE	Mevlana University Facial Expression
FAPs	Facial Animation Parameters(FAPs)

MUs	Motion Units
ML-HMM	Multi-Level Hidden Markov Models
MS-HMM	Multi-Stream Hidden Markov Models
KFTL	Kinect Face Tracking Library
DTAGN	Deep Temporal Appearance-Geometry Network
DPM	Deformable Part Model
MAOP-DL	Multi-Angle Optimal Pattern-Based Deep Learning
CK	Cohn-Kanade database
IEMOCAP	Interactive Emotional Motion Capture database
AFEW	Acted Facial Expressions In The Wild AFEW

Context and motivation

Today, machines can recognize emotions by analyzing multiple data sources, including human voice, facial expressions, and body gestures. More beneficial information can be found in face expressions. Researchers are increasingly describing emotions using various models like basic model, dimensional model, and Componential Appraisal Model. In basic model, human emotions are divided into six discrete classes: happiness, sadness, fear, neutral, disgust, and surprise.

However, dimensional models and Componential Appraisal Model of emotion have greater expressive power than discrete categories of emotion, but they are more complex and ambiguous. So, it is easier to measure the basic model than to evaluate the dimension model and Componential Appraisal Model.

Detecting emotions with technology is quite a challenging task, that has been gaining increased attention due to its applicability to various domains, such as software engineering, health care, education, and other basic uses in human daily life. This is what makes studying this field so important.

Studying and developing methods for identifying feelings is very important in daily life, as they use them in several areas, for example, in trading through websites, identifying the feelings of the buyer and providing him with the products he wants.

The topic of emotion recognition through facial expressions goes back to 1862 when Duchenne was interested in how the muscles in the human face form facial emotions. After, Charles Darwin in 1872, when he studied facial expressions and body gestures in mammals...etc. To date, research is still ongoing in this field.

Goals and approach

Our goal in this work is to use a deep learning model for emotion recognition and classification in videos because it reduces reliance on feature extraction by using “end-to-end” learning directly from the input data to the classification result. Also, it has the ability to manage large amounts of data compared to the classic methods. That’s why we developed a TCN-based model and a normal convolutional neural network model for emotion recognition with a data set of video . This set called Acted Facial Expressions In The Wild (AFEW) dataset. It contains the videos of six basic emotions: anger, disgust, fear, happiness, sadness, surprise and neutrality. We used python to program this module.

Organization

Besides the present Introduction, this document includes three main chapters:

Chapter 1 introduces some preliminaries and background. It is divided into two sections: The first introduces the basic concepts of machine learning and deep learning, the second concerns the psychological part that contain some definitions about emotions and their models, and various applications of emotion recognition.

Chapter 2 presents the state of the art in facial emotion recognition in video. It overviews the earlier work in this field, and two adopted approach that are machine learning based approach and deep learning based approach.

Chapter 3 contains the implementation part with an experiment, we will discuss the obtained result.

We conclude with a summary and some open directions for further research.

CHAPTER 1

BACKGROUND

1.1 Introduction

Emotions have a significant impact on our daily life, they affect the way we interact with others. Our choices, our perceptions and the activities that we perform are all ruled by the emotions we experience at any given time.

Nowadays, with the advancement of technology, people resorts to use it in their dealings with others, which led to a great interest in the design and improvement of the interaction between humans and machines.

In this part, we are going to begin with making a brief reminder of the development of artificial intelligence from its basic techniques to deep learning. After that, we will define the meaning of emotion and mention the different theories that classify them. Later we will mention the different reasons that make our subject important by mentioning different areas of use of emotions.

1.2 Artificial intelligence, Machine learning and Deep learning

1.2.1 Artificial Intelligence

Artificial intelligence is an old field of computer science concerned with all systems and research that aim to solve human problems in an automatic and spontaneous way in an attempt to simulate human intelligence.

The term was defined for the first time by a group of computer scientists at the Dartmouth Conferences in 1956. It is applied when a machine imitates cognitive capacities that people relate with human minds, such as learning and problem-solving.

After many years, the field reach its limit because the research in AI use a large amount of data, which need a very long processing time.

To solve that issue, it needs a parallel processing using Graphics Processing Units (GPU) that also allows a huge memory bandwidth compared to Central Processing Units (CPU) (Ongsulee, 2017).

AI research subjects include the study of: pattern recognition, Natural Language Processing (NLP), Automatic Reasoning and Game Theory just to name a few, which are essentially rule based. Later, learning theory was involved and resulted in another philosophy for intelligent systems called Machine Learning (Ongsulee, 2017).

1.2.2 Machine learning

Machine learning (ML) is a subfield of AI that aims to develop software that are used to solve a real-world problem with the ability of learning from sample data to construct a model that can predict the response for new unseen data.

In 1959 Arthur Samuel defined machine learning as the "field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). ML is usually divided into three categories according to the tackled problem: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning

Supervised learning algorithms are trained using labeled data. That is, a sample set of pairs input-output. They use patterns to make prediction of the label values on additional unlabeled data. Supervised methods include: classification and regression, which make use of several techniques such as Decision Trees, k Nearest Neighbors, Naive Bayes, and Artificial Neural Networks (Ongsulee, 2017).

Unsupervised learning

In this type of learning, the algorithm receive a set of inputs without them being labeled. In that sense, the model tries to learn patterns from the data by exploring them. Popular unsupervised learning tasks are: clustering, association discovery, and dimensionality reduction. They employ many techniques such as: k-means algorithm, Apriori, Principal Component Analysis, and Singular Value Decomposition. (Spiers, 2016; Ongsulee, 2017)

Reinforcement learning

This type of learning has three primary components: the learner (an agent), the environment (everything the learner interacts with), and actions (what the learner can do). The principle is to reward or punish the agent accordingly to the decisions it took in order to achieve a goal (Spiers, 2016; Ongsulee, 2017).

ML methods gain a lot of interest due essentially to two main factors. In one hand, the increasing amount of data available in the world, and in the other because computational devices have become more powerful and cheaper. But, ML has a primary drawback in the sense that it requires an ongoing human intervention in many stages: data preparation, feature selection, and result interpretation. Recently, a more interesting approach has revolutionized the learning field and the IA scene in general: Deep Learning.

1.2.3 Deep Learning

The emergence of Deep Learning (DL) was linked with the development of artificial neural networks. It is a subfield of machine learning (also known as deep structured learning, hierarchical learning or deep machine learning) (Deng and Yu, 2014). One of its most important characteristics is the use of the hierarchical feature extraction and the unsupervised feature learning algorithms in the place of handcrafted features selection (Song and Lee, 2013).

Artificial Neural Networks

Artificial neural networks (ANN) are a technology inspired by a biological observation and a simulation of the activity of the brain and nervous system, exactly the biological neuron. Figures 1.1 and ?? illustrate the biological and the formal neuron.

After a great deal of research and according to the idea of combining multiple processing elements into a network which is attributed to McCulloch and Pitts in the early 1940s, the first and the simplest ANN called perceptron was presented in the research paper of Frank Rosenblatt in the late 1950 (Walczak, 2019). This architecture is illustrated by Figure 1.3.

The perceptron was a single-layer neural network (SLP). In few years later the multiple layered perceptrons (MLP) was proposed. DL was concerned specially with more elaborated form of these multilayered ANN (Spiers, 2016).

Many researches adopted DL, where it has been shown to produce state-of-the-art results on various tasks. Different architectures have been developed and applied such as: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Temporal Convolutional Network (TCN), Auto-Encoder, etc. In the sequel, we will recall briefly the core of a selection of them.

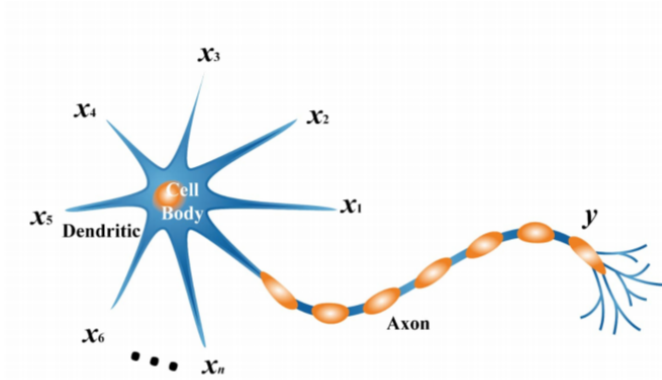


Figure 1.1: Biological neuron (Liu, 2020)

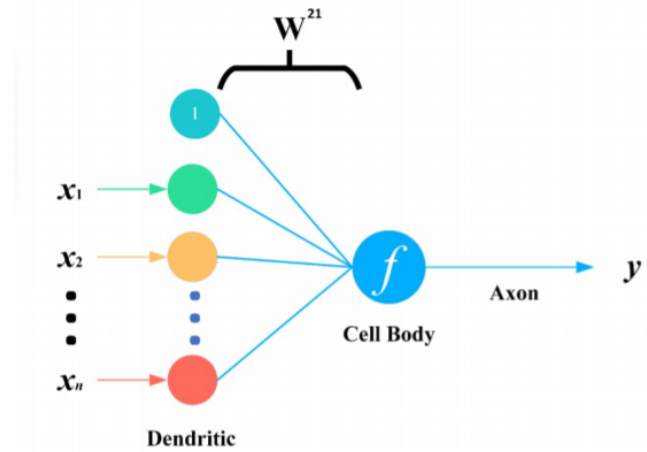


Figure 1.2: Artificial neuron (Liu, 2020)

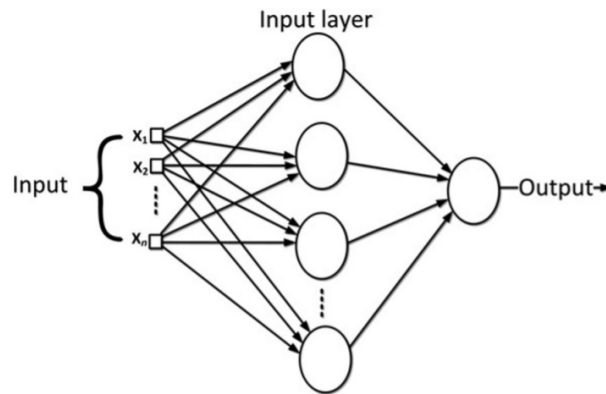


Figure 1.3: Simple perceptron (Gupta and Raza, 2019)

Convolutional Neural Networks

Many research was inspired from the biological observation, including the architecture of CNN which is inspired by the study focused on the organization of the animal visual cortex. It was presented by Hubel and Wiesel (1964) (Hubel and Wiesel, 1965). CNN was introduced for the first time in 1998 by Yann LeCun (LeCun et al., 1998).

CNN is a multi-layer neural network that consists of at least one convolutional layer, which can generate a feature map from the input data using a convolution filter, see for more explanation (Dettmers, 2015).

After every convolutional layer there is an activation function. Next to convolutional layers there are also pooling layers and fully connected layers. At the end of the network, a linear classifier computes the network output (Guo et al., 2016).

Convolution layer: A filter goes all over the input, multiply its elements with original image matrix, sum up to extract features from the original input. The output is called a feature map. An example is illustrated in Figure 1.4.

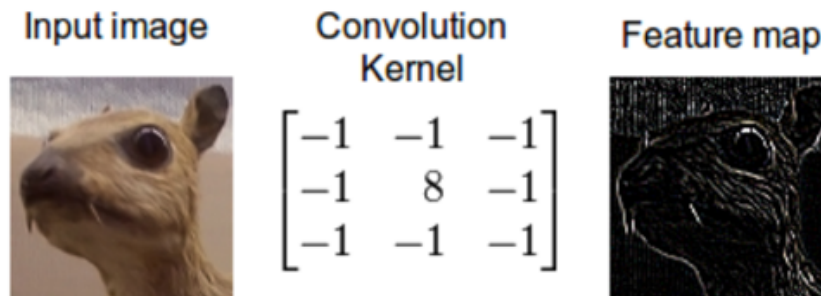


Figure 1.4: Before and after convolution filter applied to an image (Kölbl, 2017)

Pooling layer: This layer keeps the most important information and remove the rest. It can reduce the size of the feature map and thus minimize the computation cost. Generally, two types of pooling are used: max and average, by taking the maximum/average value of the pooling local area respectively.

Fully connection layer: Connects all features and sends the output value to the classifier.

Softmax layer: Maps the output of multiple neurons to the interval of [0,1] which can be considered as a probability.

The whole process of a typical CNN process is shown in figure 1.5.

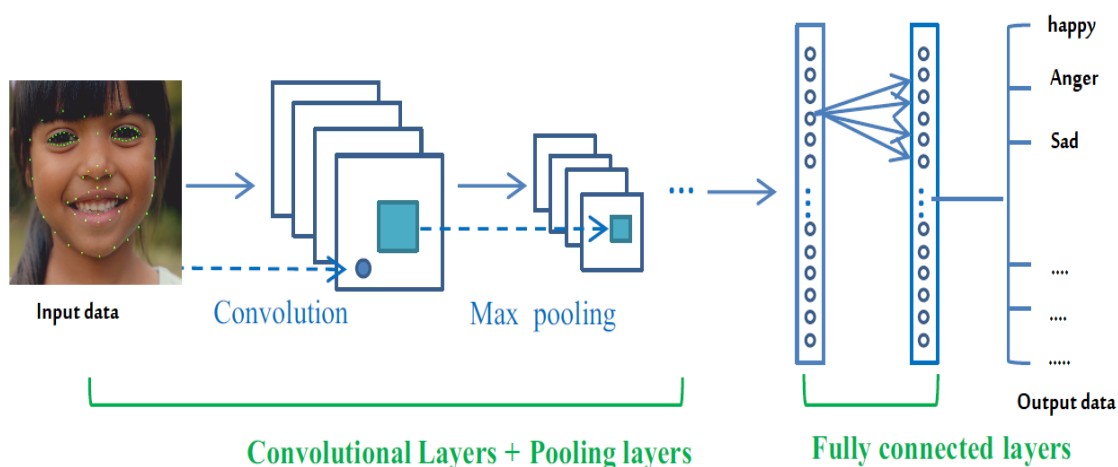


Figure 1.5: CNN classification process

There are other CNN structures that have proven a good classification ability such as: Alex-Net, Google-Net, Res-Net, and VGG. These structures have been built and tested in many

important tasks (He et al., 2019).

One of the major restrictions of standard CNN is that they only extract spatial relations of the input data while the temporal relations of them are neglected if they are part of a sequential data (Hasani and Mahoor, 2017), such as text, audio, and video. this kind of data is used in various tasks like speech recognition or time-series prediction which require a system to store and use context information. (Mallya, 2017).

To overcome this limitation, many architectures have been developed to deal with these aspects, such as 3D Convolutional Neural Networks (3D-CNN), Recurrent Neural Networks (RNN) and their multiple variants.

Three Dimensional Convolutional Neural Networks (3D-CNN)

1D CNN can extract only the spectral features from the data, while 2D CNN can extract spatial features from the input data. However, 3D CNNs can take advantage of both 1D and 2D CNNs by extracting both spectral and spatial features simultaneously from the input data (Singh et al., 2020).

3D CNNs are formed of 3D convolution throughout the whole architecture. In 3D convolution, filters are designed in 3D, and channels and temporal information are represented as different dimensions (Kalfaoglu et al., 2020), a typical architecture of 3D CNN is shown in the figure 1.6 below.

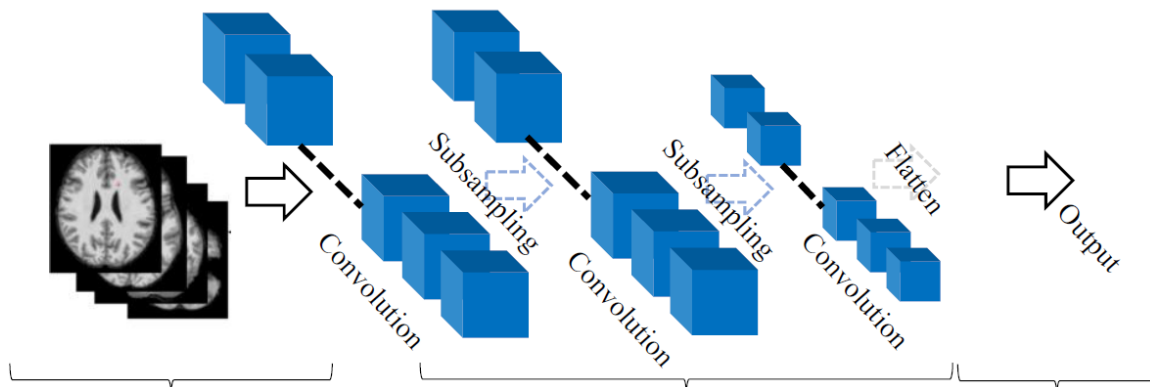


Figure 1.6: Typical architecture of 3D CNN (Singh et al., 2020)

Recurrent Neural Networks

The introduction to recurrent neural network was the discovery of the feedback (closed loop) connections, which was introduced by Hopfield in 1983. The first and simplest model of RNN was the Vanilla RNN (VRNN), or sometimes called Elman RNN. VRNN representation is shown in figure 1.7. The tanh activation function is used in the hidden recurrent layer and

the activation function for the output layer is selected according to the problem to be solved (Saud and Shakya, 2020). The operation of vanilla RNN can be expressed mathematically as below:

$$H(x) = f(W_x h X_t + W_h h H_t - 1) \quad (1.1)$$

$$O_t = g(W_x o C_t) \quad (1.2)$$

Where X_t is input at time , H_t state info at time t.

W_{xh}, W_{hh} and W_{xo} are weight of matrices.

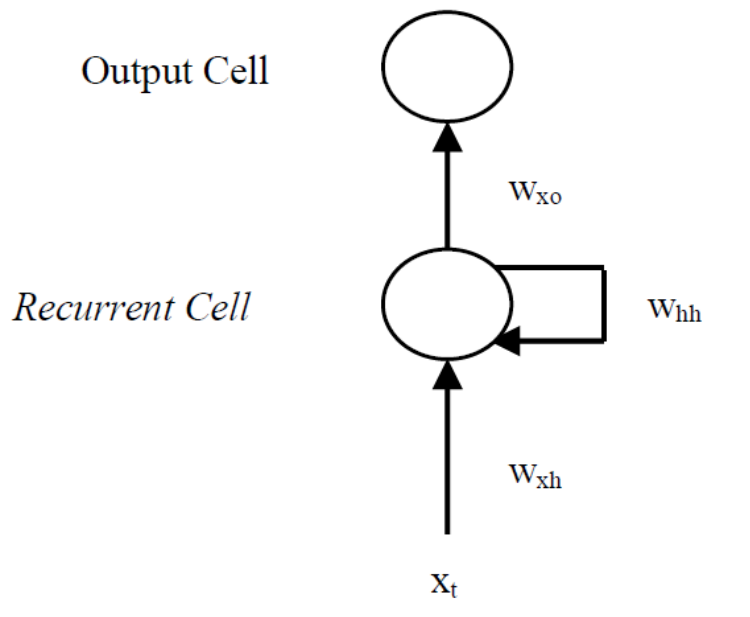


Figure 1.7: The architecture VRNN (Saud and Shakya, 2020)

This model has a main problem called vanishing gradient problem, for more understanding read this paper (Hochreiter, 1998).

Since RNN is used to deal with sequential data, and in order to handle the long-term dependencies and avoiding the vanishing/exploding gradient problem (Yu et al., 2019).

Long Short-Term Memory

The LSTM network is capable of learning long-term dependencies. It is well-suited to classify and/or predict sequential data. The common architecture of LSTM units is composed of a memory cell, an input gate, an output gate and a forget gate (Hochreiter and Schmidhuber, 1997), as depicted in Figure 1.8.

Every LSTM cell computes new values of hidden state and cell state, with the mathematical formulation is given below.

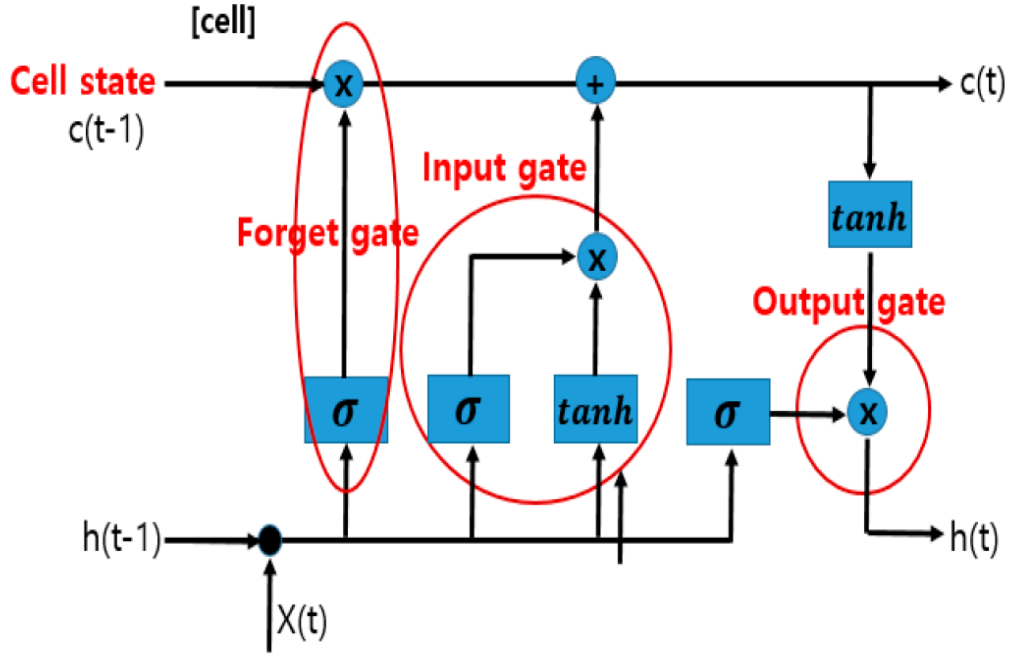


Figure 1.8: Architecture of LSTM (Saud and Shakya, 2020)

$$f_t = \alpha(x_t W_f + H_{t-1} U_f) \quad (1.3)$$

$$i_t = \alpha(x_t W_i + H_{t-1} U_i) \quad (1.4)$$

$$o_t = \alpha(x_t W_o + H_{t-1} U_o) \quad (1.5)$$

$$H_t = \tanh(x_t W_g + H_{t-1} U_g) \quad (1.6)$$

$$C_t = \alpha(c_{t-1} * f_t + i_t * H_t) \quad (1.7)$$

$$H_t = \tanh(C_t) * o_t \quad (1.8)$$

Where i, f and o are input, forget and output respectively .

H and C are hidden state and memory state respectively.

Many researches have combined two or three model of neural network, such us CNN, RNN or other types of ANN, to benefit from the advantages of them, and they achieved a good accuracy.

CNN with RNN

This neural network architecture takes advantage of the construction of convolutional neural network (CNN) and recurrent neural network (RNN) and joint them together for making higher results in different tasks. CNN is able to extract temporal or spatial features from data, but lacks the ability of learning sequential relations. On the other side, RNN is trained for

sequential modelling but incapable to extract features in a parallel way (Wang et al., 2016). An example of a combined architecture of CNN and RNN is shown in the Figure 1.9.

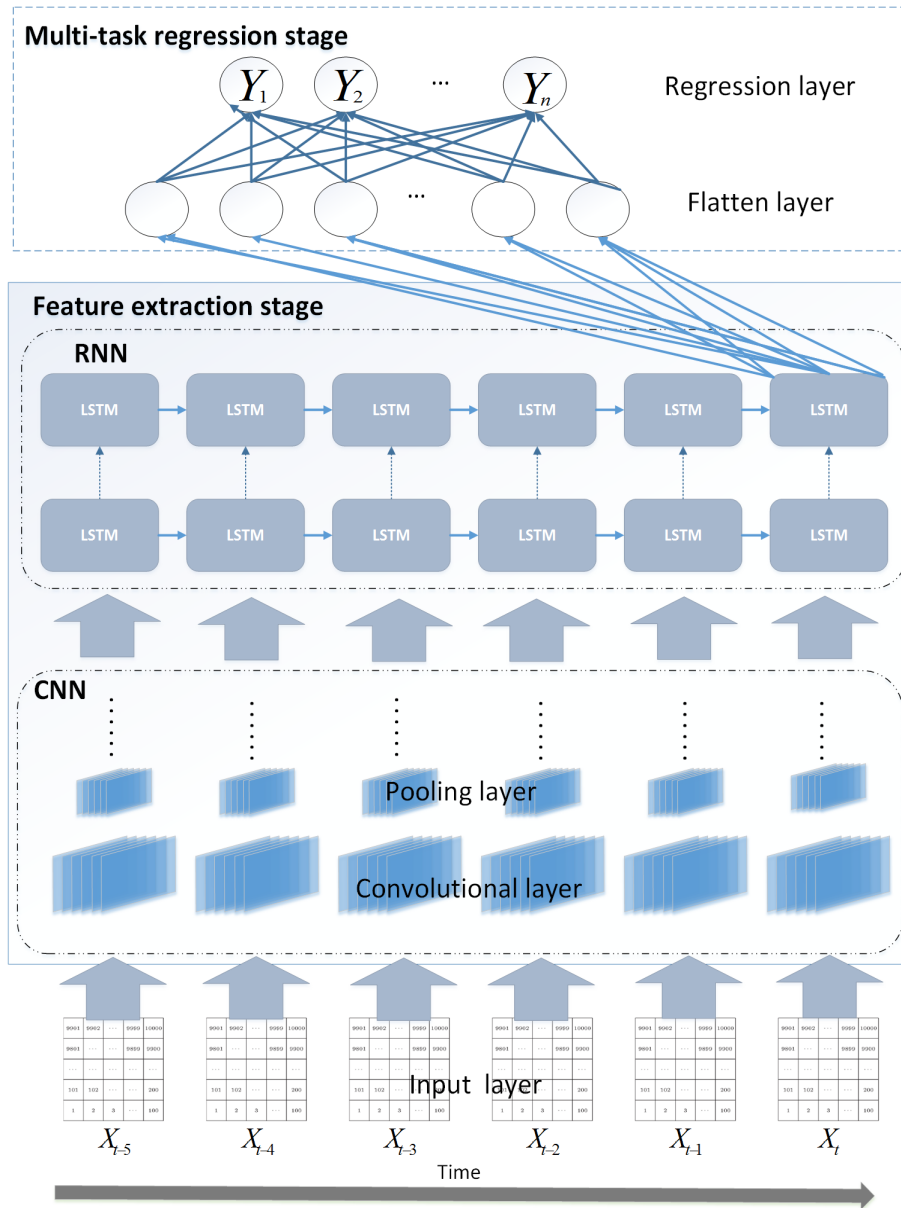


Figure 1.9: An example of a combined architecture of CNN and RNN (Huang et al., 2017)

The idea of combining RNN and CNN network have been used in many tasks, where it has got better results than prior methods. It has been adopted by many researches: 3D object classification (Socher et al., 2012), Sentiment analysis (Wang et al., 2016) (Wang et al., 2016), video copy detection (Hu and Lu, 2018), acoustic scene classification (Bae et al., 2016), Sector Stock Price Analysis (Zhang et al., 2018), Face Anti-spoofing (Xu et al., 2015).

Temporal Convolutional Network

Another neural network architecture was developed to deal with sequence data, called Temporal Convolutional Networks (TCN). While they avoid the drawbacks of CNN and RNN, TCN outperforms these approaches in many datasets and various recent applications, while preserving an effective calculation and memory resources (Bai et al., 2018).

TCN consists of three parts including: causal convolution, dilated convolution and residual layers. It is based on two major characteristics: The first one is that causal convolution in the architecture means that there is no information leakage from future to past; secondly, the architecture can take a sequence of any length and map it to an output sequence of the same length, just as with an RNN (Feng, 2019).

The difference between standard convolution and causal convolution is the fact that convolutional operation in causal convolution performed to obtain the output at time t does not take future values as inputs. This means that, using a kernel size k , the output O_t is obtained using the values of $X_{t(k-1)}, X_{t(k-2)}, \dots, X_{t-1}, X_t$, (Figure 1.10) (Lara – Bentezet et al., 2020).

TCNs use one dimensional dilated convolutions that increase the receptive field of the network without using pooling operations, hence there is no loss of resolution (Yu and Koltun, 2015).

The differences between standard convolutional network, causal convolutional network, and dilated causal convolutional network are shown in the Figure 1.10.

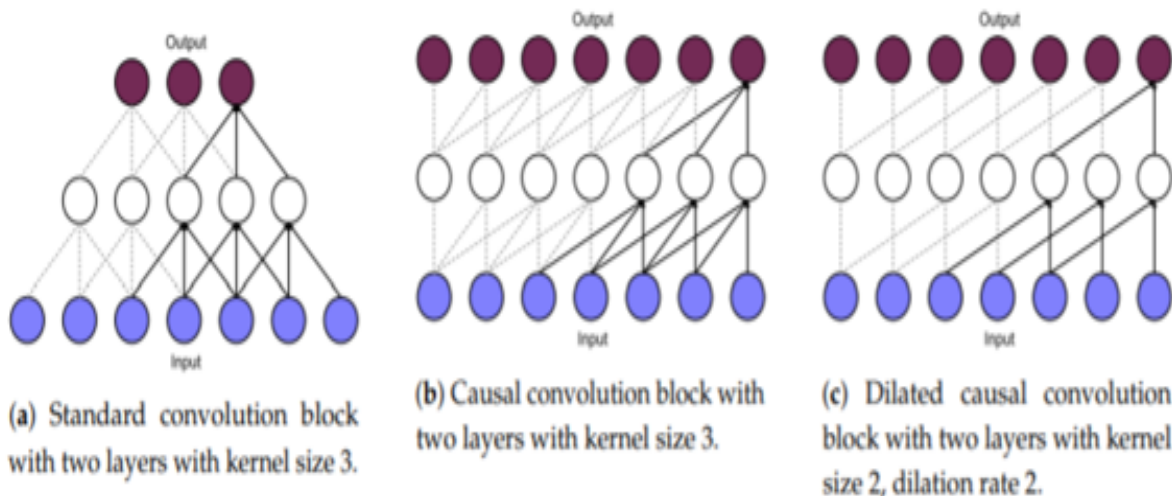


Figure 1.10: Differences between (a) standard convolutional network, (b) causal convolutional network, and (c) dilated causal convolutional network (Lara-Benítez et al., 2020)

1.3 Emotions and their models

At various times, a person's face reveals how he or she feels or what is his/her mood. Humans can produce during communication thousands of facial movements ranging in complexity and meaning. People's interactions are influenced by their emotions; even though everyone intuitively understands emotions, defining them can be difficult because there are some facial expressions which have similar general shape, yet they convey a different message with varying degrees of intensity (Singh, 2012).

It is remarkable that the study of emotion has been seriously neglected throughout much of psychology's brief history. Ironically, psychologists have been the last to recognize that emotions lie at the center of human experience (Sloboda and Juslin, 2001).

1.3.1 Definition

As pointed out by Fehr and Russell (Fehr and Russell, 1984), "everyone knows what an emotion is, until asked to give a definition".

There is not a conclusive definition of emotion, but a few psychologists propose numerous of them. Among them Paul Ekman who is an American psychologist and teacher emeritus at the College of California, San Francisco. He may be a pioneer within to ponder of feelings and their connection to facial expression. He gave the following definition: "emotions are a process, a particular kind of automatic appraisal influenced by our evolutionary and personal past, in which we sense that something important to our welfare is occurring, and a set of psychological changes and emotional behaviors begins to deal with the situation"¹. Kleinginna and Kleinginna (Kleinginna and Kleinginna, 1981) proposed the following consensual definition: "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can give rise to affective experiences such as feelings of arousal, pleasure/displeasure; generate cognitive processes such as perceptually relevant effects, appraisals, labeling processes; activate widespread physiological adjustments to the arousing conditions; and lead to behavior that is often, but not always, expressive, goal-directed, and adaptive." This definition is based upon a review of 92 definitions found in textbooks, articles, dictionaries, and other sources (Sloboda and Juslin, 2001).

A few different theories or models have emerged to categorize and clarify the emotions that people experience: the basic emotion model, the dimensional model, and the componential appraisal model.

¹<https://www.paulekman.com/universal-emotions/>

1.3.2 The Basic Emotion Model

The idea of "basic emotions" dates back to the works of Descartes (1649/1988). However, the real debate on "emotional basicness" was presented in Darwin book (1872/1998) entitled *The Expression of the Emotions in Man and Animals*, which was interpreted by Tomkins in 1962 and 1963 (Yao, 2014)

P. Ekman and his colleagues represented the theoretical proposals of the basic emotion model in their research on universal recognition of emotion from facial expression. According to their experiments, judging the static images with facial expressions of human, there are six basic emotions as shown in the Figure 1.11 that can be recognized universally. These emotions are: happiness, sadness, surprise, fear, anger, and disgust. From his research the other emotions (higher level emotions) can be combined from six basic emotions (Yao, 2014)



Figure 1.11: The six basic emotions (Yao, 2014)

Plutchik coincided with Ekman's theory and developed the "wheel of emotions" in 1980 shown in the Figure 1.12. There are eight emotions arranged in opposite pairs (joy and sadness; anger and fear; disgust and acceptance; surprise and anticipation) with the strength of the emotions described in distinct colors. According to Plutchik's research, human beings cannot experience opposite emotions at the same time. Complex emotions, which could arise from a cultural condition or association with basic emotion, can be formed by just modifying some basic emotions. Though some researchers have proposed a different number of basic emotions

which can range from 2 to 18 such as Ortony and Turner in 1990 or Wierzbicka in 1992.

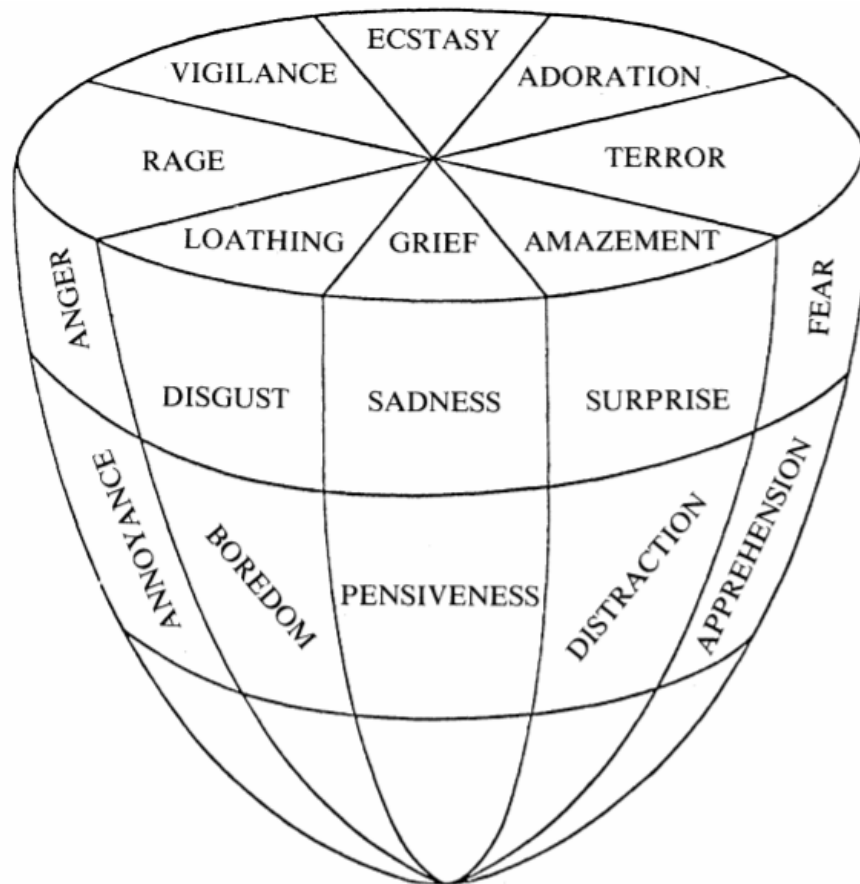


Figure 1.12: Plutchik's wheel of emotions (Plutchik, 1982)

The notion of emotional basicness has been criticized, because different researchers have come up with different sets of basic emotions (Sloboda and Juslin, 2001). The difference is shown in Table 1.1.

Ortony and Turner in 1990 argued that the view of existing basic emotions can build or explain all other emotions, and the expression of emotions is not the same as the emotions themselves. For example, specific facial expressions that are recognized around the world and seem universal are not linked to emotions, but rather to certain conditions that also elicit emotions (Yao, 2014). Ekman also reported that there is some confusion from the judgment study of the six basic emotions. For example, anger and disgust, fear and surprise, Surprise is also confused with the emotion of interest (Yao, 2014). Some authors such as Izard (2011), Levenson (2011), Panksepp and Watt (2011) state that "pure" basic emotions are rarely experienced by adults (because they interact with higher order cognitive processes to produce more complex emotional states (Piórkowska and Wrobel, 2017). Overall, despite some differences, the theories of basic emotions offer a conceptual framework for future research in the field (Piórkowska and Wrobel, 2017).

Table 1.1: Selected lists of basic emotions (Sumpeno et al., 2011)

Psychologist	Basic emotion
Plutchik	Anger, anticipation, trust, disgust joy, fear, sadness, surprise
Ekman, Friesen, Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Frijda	Desire, happiness, interest, surprise, wonder, sorrow
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
Mowrer	Pain, pleasure,
Oatley and Johnson-Laird	Anger, disgust, anxiety, happiness, sadness

1.3.3 The Dimensional Model

Emotions in a dimensional framework can be mapped by two or three variables, such as Valence, arousal and energy or control. Valence dimension usually represents the positive or negative degree of the emotion, and the range is from uncomfortable feelings to comfortable feelings. The arousal dimension represents how excited the emotion is, and it ranges from low to high. While the energy or control dimension represents the degree of the energy or control over the emotion (Yao, 2014).

The dimensional approach dates back to Spencer in 1980, cited in the work of Izard in 1977, the same notion was developed and provided by Wundt in 1897, Woodworth in 1938, and Schlosberg in 1941 (Sloboda and Juslin, 2001).

Emotion recognition tasks usually uses only two-dimensional model, which is the valence-arousal model. This model is very intuitive to represent emotions on some continuous scale, but it will cause a loose of some information. Basic emotions can still be represented in a dimensional model as a point or an area. But, some of them such as anger and disgust are hard to distinguish, and some emotions cannot even be described as shown in the figure ?? (Yao, 2014).

The figure 1.13 shows the representation of the six basic emotions in on the Valance Arousal Dimensional Model.

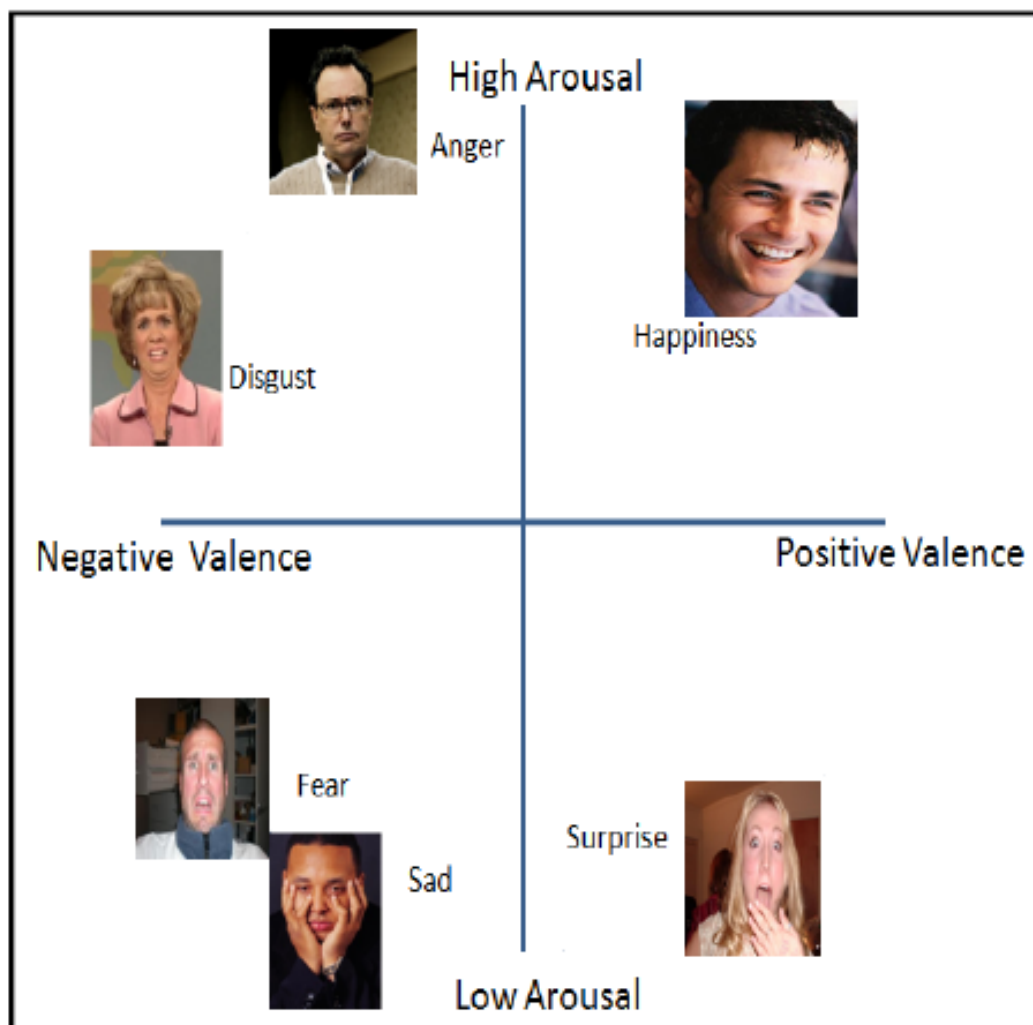


Figure 1.13: Basic emotions on the Valence Arousal Dimensional Model (Jerritta et al., 2011)

1.3.4 The Componential Appraisal Model

This model can be seen as an extension of the dimensional model proposed by Scherer and his colleagues (Scherer and Moors, 2019). He defined emotions as complex, multi-componential, dynamic process, and there is no limitation on their number. Emotion differentiation of this model allowed Scherer to model individual differences and emotional disorders. Though, it is still an open area for emotion recognition with it, because the measurement of emotional states changing in this model is complicated (Yao, 2014).

Thus, in most of recent emotion recognition research, the basic emotion model is chosen, because a higher number of dimensions provided by other models cannot be relied upon for estimation. We will also choose it in our study and try to recognize the six fundamental emotions in video (Anger, Fear, Disgust, Happiness, Sadness and Surprise).

1.4 Emotion Recognition

Emotion plays an important role in human beings daily lives. Understanding and recognizing emotions is an important research area that many researchers work on in recent years using various methods (Singh and Fang, 2020). They have used many sources including text, speech, hand, EEG/ECG signals, and body gesture as well as facial expression. Presently, most of the emotion recognition methods only use one of these sources (Yao, 2014).

Recognizing emotions from facial expressions or recognizing facial emotions is a method that uses human facial features (eyes, eyebrows, and mouth) to identify feelings, as these features change during our behavior with others.

The figures: 1.14 to 1.19 shows the change in the features of the human face for the six basic emotions.

1.4.1 Feature of emotion

Anger

Eyebrows pulled down, upper lids pulled up, lower lids pulled up, merging of lips rolled in, lips may be tightened.

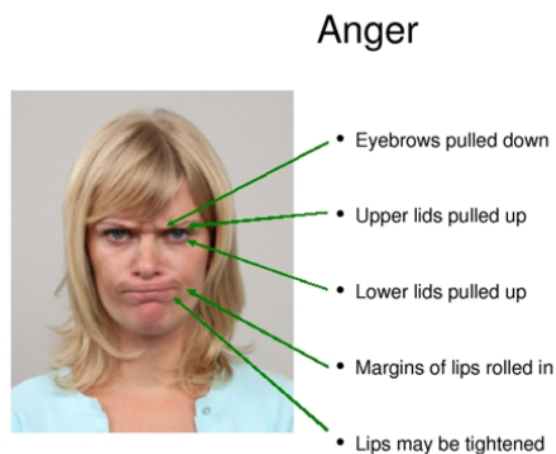


Figure 1.14: Anger

Fear

With the characteristics below eyebrows pulled up and together, upper eyelids pulled up, mouth stretched.



Figure 1.15: Fear

Disgust

Eyebrows pulled down, nose wrinkled, upper lip pulled up, lips loose.

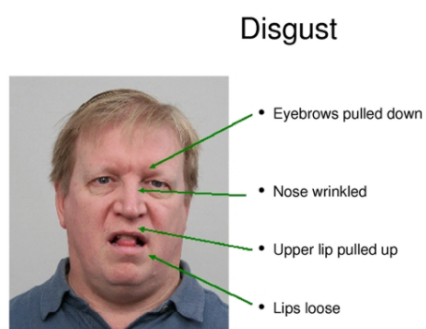


Figure 1.16: Disgust

Happiness

Muscle around the eye tightened, crows feet wrinkles around eyes, cheeks raised, lip corners raised diagonally.

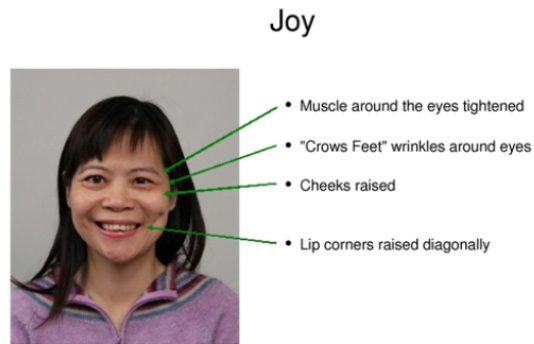


Figure 1.17: Happiness

Sadness

Inner corners of eyebrows raised, eyelids loose, lip corners pulled down.

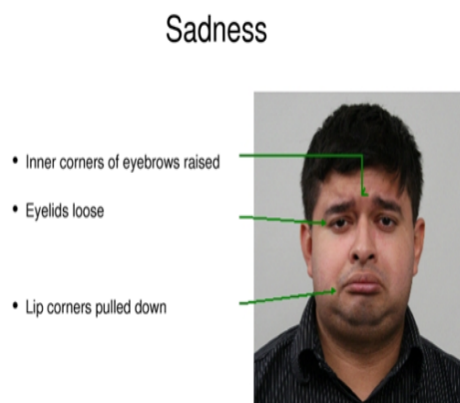


Figure 1.18: Sadness

Surprise

Entire eyebrows pulled up, eyelids pulled up, mouth hangs open.

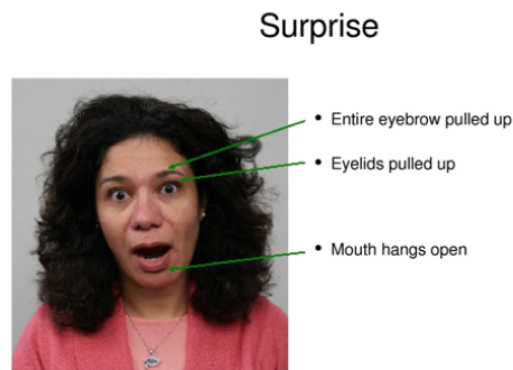


Figure 1.19: Surprise

There are other emotions such as shame, pride, jealousy and guilt. While these emotions are important ones, they are still not considered part of the basic emotions set.

These features are visually salient points in facial regions, it is called Facial Landmarks (FLs). Ekman use it to develop the Facial Action Coding System (FACS)

1.4.2 Facial Action Coding System (FACS)

FACS is a system developed by Ekman and Friesen ([Ekman, 1977](#)) based on facial muscle changes for describing facial expressions by action units (AUs). Facial action units (AUs) code the fundamental actions (46 AUs) of individual or groups of muscles that are typically observed when a facial expression produces a particular emotion ([Tian et al., 2001](#)). Figure 1.20 illustrates some examples. Any facial emotions can be uniquely described by a combination of AUs. See Table 1.2 for a summary.

AU1	AU2	AU5	AU9	AU15	AU23	AU25	AU27
Inner Brow Raiser	Outer Brow Raiser	Upper Lid Raiser	Nose Wrinkler	Lip Corner Depressor	Lip Tightener	Lip Parts	Mouth Stretch

Figure 1.20: Some examples of AUs (upper and lower face) adapted from ([Ko, 2018](#))

Table 1.2 shown the typical AUs seen in each of the basic and compound emotion categories.

Table 1.2: The prototypical AUs observed in basic and compound emotion category, proposed in (Fabian Benitez-Quiroz et al., 2016)

Category	AUs	Category	AUs
Happy	12,25	Sadly disgusted	4,10
Sad	4,15	Fearfully angry	4,20,25
Fearful	1,4,20,25	Fearfully supr	1,2,5,20,25
Angry	4,7,24	Fearfully disgd	1,4,10,20,25
Surprised	1,2,25,26	Angrily surprised	4,25,26
Disgusted	9,10,17	Disgd.surprised	1,2,5,10
Happily sad	4,6,12,25	Happy fearful	1,2,12,25,26
Happily supr.	1,2,12,25	Angrily disgusted	4,10,17
Happily disgd.	10,12,25	Awed	1,2,5,25
Sadly fearful	1,4,15,25	Appelled	4,9,10
Sadly angry	4,7,15	Hatred	4,7,10
Sadly surprised	1,4,25,26		

1.5 Applications

Recognizing emotions and use it in many domains is a part of affective computing, which was defined by Rosalind Picard (Picard, 1999). This part presents some applications in many areas of life.

1.5.1 Software engineering

It is clear that human emotions influence interactions with software products. which can influence human feelings that make people buy or not, as shown in Hill record of investigation (Hill, 2009). In the other hand, the emotional state of programmers have significant impact on program quality and developers efficiency (Wrobel, 2013). Therefore, recognize emotions is interesting in software engineering.

Recognizing emotions can extend the software usability by measuring users satisfaction and enhance software quality and developers productivity (Kołakowska et al., 2014).

1.5.2 Education and e-education

To improve the learning environment, many systems have been developed to recognize students' sentiments.

Shen et al. (Shen et al., 2009) build an Emotion Integrated e-learning architecture model and explore the machine's capacity to understand emotions from physiological information.

A lot of study has gone into measuring the affective dimension of student participation in learning. The work proposed in (Grafsgaard et al., 2013) presents an automated facial recognition approach to analyze student facial movements during tutoring using the Computer Expression Recognition Toolbox (CERT), which tracks well-defined facial movements. The predictive models highlighted relationships between facial expression and aspects of engagement, frustration, and learning.

In (Whitehill et al., 2014), another attempt is made to establish real-time automatic engagement recognition from students' facial expressions. Teachers regularly analyze their pupils' levels of involvement, and facial expressions play a key role in these evaluations. The result show that automated engagement detectors perform with comparable accuracy to humans.

1.5.3 Websites customization

With the development of the Internet, service providers are collecting more information about their customers. The layout and advertisements are displayed based on the user's profile. Adding information about users' emotions can provide more accurate knowledge of users, thus service providers can influence their behavior on websites, as well as trigger their emotions through different types of online advertising (Kołakowska et al., 2014).

1.5.4 Health care

The mental status of patients reflects the behavioral and cognitive functioning. This increases the interest of recognizing emotion in health. Through wellness monitoring, and using an emotion recognition system we can benefit the mental health and well-being of individuals that are stressed, anguished, or depressed (Hasnul et al., 2021).

Emotion recognition system can also be used by medical companies and service providers to improve their services, based on the emotional feedback of patients (Hossain and Muhammad, 2019).

1.5.5 Automotive industry

Since the emotional state such as the level of vigilance, drowsiness, fatigue or bad mood of drivers plays a significant role on driving performance and safety of the roadways, the automotive industry integrated an interactive emotion recognition technology in new cars (Jones and Jonsson, 2007).

Using facial emotion detection with a car-voice can impact driving performance (Nass et al., 2005), resulting in an increased roadways safety.

1.5.6 Video game

Video games are created with a specific target audience in mind, with the goal of inducing a specific set of behaviors and feelings in the players. Users are requested to play the game for a certain amount of time during the testing process, and their feedback is used to improve the final product. Facial expression detection can help determine which feelings a user is experiencing in real time while playing without having to manually analyze the entire movie (Kołakowska et al., 2014).

1.5.7 Others

We have mentioned above the famous domain of research in emotion recognition, but there are other diverse applications which is substantiated by research publications such us:

- Analyze emotions to display personalized messages in smart environments.
- Help decision-making of recruiters
- Identify uninterested candidates in a job interview, monitor moods and attention of employees.
- Lie detectors and smart border control.
- Predictive screening of public spaces to identify emotions triggering potential terrorism threat.
- Analyzing footage from crime scenes to indicate potential motives in a crime.

1.6 Conclusion

In this chapter, we have provided a brief reminder of the key concepts important to understanding our subject. Emotion recognition has been in the center of attention for many years. Several researchers have proposed enormous papers on the field. A selection of these work will be mentioned in the next chapter.

CHAPTER 2

EMOTION RECOGNITION: STATE OF THE ART

2.1 Introduction

Emotion recognition is an important research area that many researchers work on for a long time. That is why the literature on the field is too voluminous to review here. Emotions can be expressed through uni-modal social behaviors, such as speech, facial expressions, and gestures, or bi-modal behavior such as speech and facial expressions, or they can be expressed through multi-modal parameters such as audio, video and physiological signals.

In our research, we try to focus on some of the methods that facial expressions are used in the video frame to identify emotions, and since the latter is a collection of sequential images over time, this does not prevent us from mentioning some research in this field about images. We will survey the research development in this task through three parts: classical methods, ML based methods, and DL based methods.

2.2 Earlier work

The study and understanding of human facial expressions has been a long-standing problem since years. Initially, it was the interest of psychologists and later the focus of computer scientists, who relied on psychological research to embody it in computer systems. These are some important former work.

Duchenne, 1862: The first scientific study on facial expression analysis that was interested in how the muscles in the human face form facial emotions was published ([Eleftheriadis, 2016](#)).

Charles Darwin, 1872: He studied facial expressions and body gestures in mammals ([Darwin, 2015](#)), and the relevance of face expressions in communication and characterized the

various emotions expressed through facial expressions.

Schlosberg, 1954: He proposed a basis with three dimensions for describing emotions: AR(attention-rejection), PU(pleasantness-unpleasantness), and level of activation (Schlosberg, 1954).

Paul Ekman, 1971: His work is considered as the most important and comprehensive work presented in the field of facial expression analysis. It describes a group of six basic emotions universal in terms of expression and understanding (Ekman and Friesen, 1971).

Sown et al, 1978: They used a series of movie frames to make a preliminary investigation on automatically expressions analysis (Sown, 1978).

Terzopoulos and Waters, 1990: Depended on Ekman and Friesen work, they derived a 3D computer model to display human facial expressions, and also to analyze the expressions from a video sequence (Terzopoulos and Waters, 1990).

Yacoob and Davis, 1994: They presented a study of facial expression recognition using statistical properties with only very weak models of facial shape (Yacoob and Davis, 1994).

With the advent of artificial intelligence, the research on this topic has increased by finding algorithms to identify and classify emotions which we have divided into two groups according to whether the features are manually extracted (ML-based approaches) or generated by a deep neural network (DL-based Approaches).

2.3 Machine Learning based approaches

The ML conventional emotion recognition approaches comprise three important steps: (1) face detection/tracking, (2) feature extraction, and (3) expression classification.

First, we proceed by detect the face from an input image or video frame using a face detection technique. Second, one should extract various spatial and temporal features from the detected face using feature extraction methods such Local Binary Pattern (LBP), or Gabor Filters. Third, machine-learning algorithms such as Support Vector Machine (SVM) or Random Forests are applied for expression classification using extracted features. This process is shown in Figure 2.1.

To more understand this approach, we present below a simplified explanation of the three stages in this traditional approach.

2.3.1 Face Tracking and Detection

In facial expression analysis, the initial stage is to locate the face in the image or video sequence. Face detection or face localization refers to detecting the face within an image, whereas face

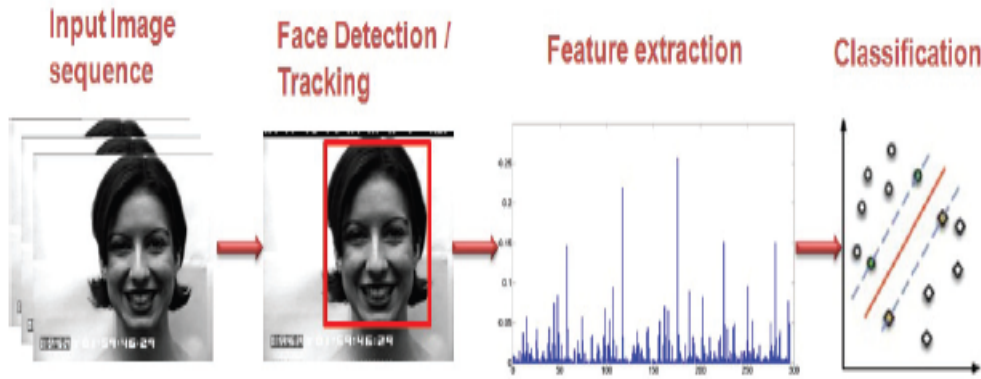


Figure 2.1: Procedure in conventional face expression approaches. (Khan, 2013)

tracking refers to locating the face and tracking it through multiple frames of a video sequence. This aids in the acquisition of face features while ignoring other objects and items (Khan, 2013; Boda et al., 2016).

There has been a lot of work in face detection and tracking research. Of course, we cannot survey all the methods. We mention some of them in the rest of the section.

Kanade-Lucas-Tomasi Tracking algorithm

The Kanade-Lucas-Tomasi tracker was one of the early 1990s systems for detecting and tracking faces. Lucas and Kanade proposed the first framework (Lucas et al., 1981). Their work was later extended by Tomasi and Kanade. This technique is used to find scattered feature points with enough texture to allow for accurate tracking of the needed points (Sethi and Aggarwal, 2011).

In (Boda et al., 2016), the Kanade-Lucas-Tomasi (KLT) algorithm is used to track human faces in a video frame. From one frame to the next, they calculate the displacement of the tracked points. Calculation of the head movement using this displacement computation is simple. The optical flow tracker is used to track the feature points of a human face. The algorithm tracks the face in two easy steps: first, it locates the traceable feature points in the first frame, and then it uses calculated displacement to follow the discovered features in subsequent frames, show Figure 2.2.

The suggested project's goal in (Wagener and Herbst, 2002) is to create a reliable and efficient face (head) tracker using the Kanade-Lucas-Tomasi (KLT) tracking algorithm. Figures 2.3. to 2.8 show a sequence of images produced by the KLT tracking algorithm.

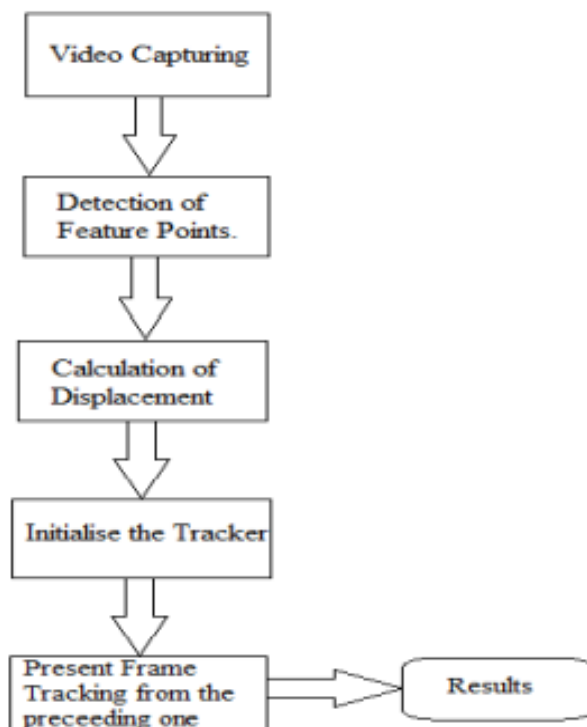


Figure 2.2: KLT algorithm flow chart used in (Boda et al., 2016)

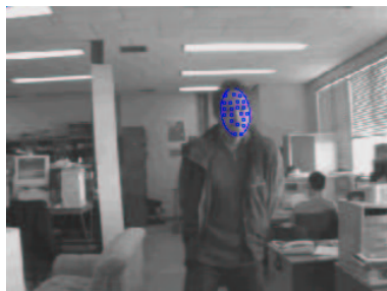


Figure 2.3: frame 0 (Wagener and Herbst, 2002)



Figure 2.4: frame 1 (Wagener and Herbst, 2002)



Figure 2.5: frame 2 (Wagener and Herbst, 2002)



Figure 2.6: frame 3 (Wagener and Herbst, 2002)



Figure 2.7: frame 4 (Wagner and Herbst, 2002)



Figure 2.8: frame 5 (Wagner and Herbst, 2002)

Piecewise Bezier Volume Deformation (PBVD) tracker

The Piecewise Bezier Volume Deformation (PBVD) tracker is a face tracker widely used by face expression recognition researchers. It is developed by Tao and Huang (Tao and Huang, 2002). This tracker uses a generic 3D wireframe model of the face associated with 16 Bezier volumes as shown in Figure 2.9. Figure 2.10 shows the wireframe model and the associated

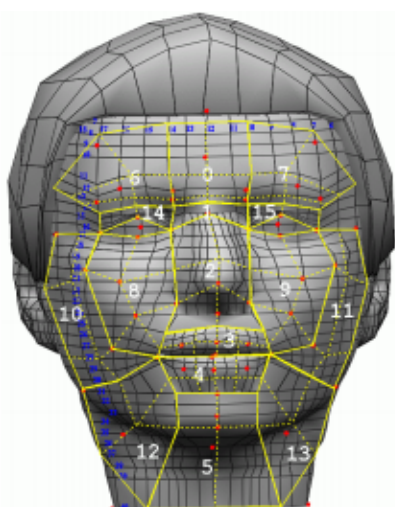


Figure 2.9: The facial mesh and the associated 16 Bézier volumes (Tao and Huang, 2002)

real-time face tracking.

The Person-Spotter's tracking

This system, developed in 1998 by Steffens et al., achieves fast and robust tracking. Along with face recognition, the system also has modules for face detection and tracking. The tracking algorithm locates regions of interest which contain moving objects by forming difference images. Skin and convex detectors are then applied to this region in order to detect and track the face. The Person-Spotter's tracking system has been demonstrated to be robust against considerable background motion. It applies a graph model on the face to automatically ignore

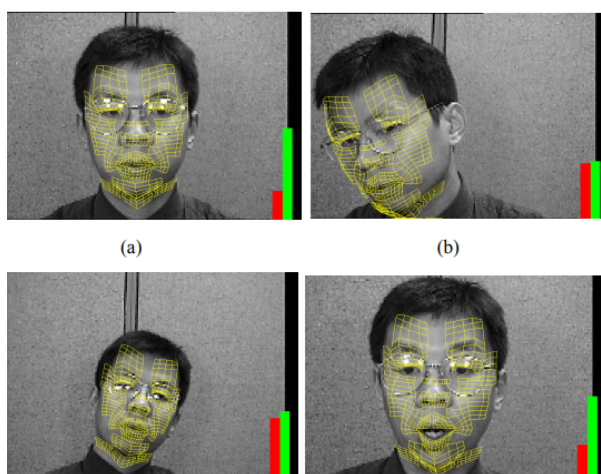


Figure 2.10: Results of the real-time tracking system ([Tao and Huang, 2002](#))

the background ([Steffens et al., 1998](#)). An illustration is shown in Figure 2.11.

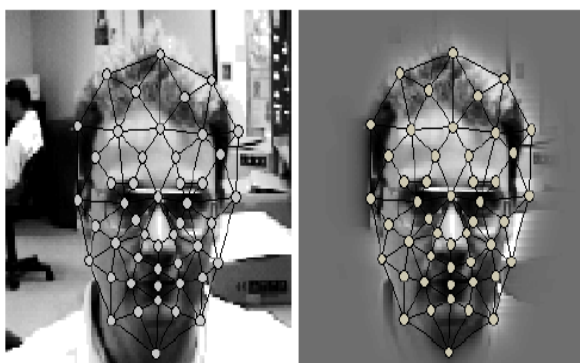


Figure 2.11: Person-Spotter's model graph and background suppression ([Steffens et al., 1998](#))

Viola Jones algorithm

The Viola Jones algorithm, was proposed by Paul Viola and Michael Jones in 2001. It was the first framework to provide competitive detection rates for objects. The framework has a high detection rate, making it a very fast and robust algorithm ([Dang and Sharma, 2017](#)). Figure 2.12 shows an example of the use of VJ algorithm in detecting the face. It has four stages: Haar feature Selection, creating an integral image, AdaBoost training and Cascading classifiers. The Viola-Jones (VJ) algorithm is used also to detect faces in video frames ([Abdulsalam et al., 2019](#)) as shown in Figure 2.13.



Figure 2.12: Face detection using the VJ algorithm ([Alionte and Lazar, 2015](#))

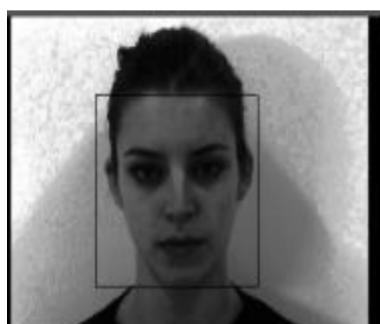


Figure 2.13: Other example for face detection using the VJ algorithm ([Abdulsalam et al., 2019](#))

2.3.2 Feature Extraction

The second step of the conventional face expression process is extracting useful data or information (features) from the image (detected face image) and represent it in a lower dimensional space. This step may directly influence the performance of the algorithms; therefore, it is necessary to choose the suitable feature extraction method relative to its applicability and feasibility.

Feature extraction is an important step in the process of conventional approach, and aims at the extraction of the least amount of relevant information, without getting later into the problem of curse of dimensionality and over-fitting when the number of extracted features is overwhelming ([Huang et al., 2019](#)). Three types of features can be extracted with different techniques : geometric features, appearance features, and motion features.

Geometric features

Facial feature points (such as: mouth, eyes, brows, nose) are extracted to form a feature vector that represents the face geometry, which include the shape and locations of facial components ([Tian et al., 2005](#)).

The motivation for employing a geometry-based method is that the position and size of various facial points differs when expressing, by measuring this movement, the underlying facial

expression can be determined (Khan, 2013).

One of this methods that extract geometric features is Active Shape Model(ASM).

Active Shape Model(ASM)/Active Appearance Model(AAM)

The Active Shape Model(ASM) proposed in (Cootes et al., 1995) is a statistical model. It matches the initial shape of the human face using global shape model, and establish a local texture model to acquire the contour features of the target(expression) more precisely (Huang et al., 2019).

ASM use only shape constraints with some information about the image structure, and have not benefit of all the available information: the texture across the target (Khan, 2013).

Therefore, the Active Appearance Model(AAM) (Cootes et al., 2001) is developed for matching a statistical model based on both shape and texture. It could be considered as the hybrid methods because it used both geometric and appearance features (Khan, 2013).

Cristinacce and his colleagues (Cristinacce et al., 2004) uses AAM to detect feature points of local edges such as facial organs, in a Multi-Stage Approach that they propose to Facial Feature Detection.

Saatci and Town (Saatci and Town, 2006) combine AAM with the SVM classifier to improve the recognition rates. The use of AAM is referred to the higher recognition rate in fitting texture features (Huang et al., 2019).

The typical examples of geometric-feature-based methods are those of Pantic and her colleagues (Pantic et al., 2007) in facial expression recognition task. Figure 2.14 shows the tracked landmarks in Pantic work.

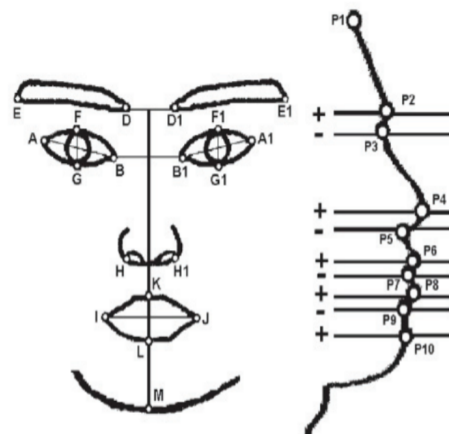


Figure 2.14: facial landmarks in Pantic work (Khan, 2013)

The appearance features

The appearance features describe the appearance (skin texture) changes of the face, such as wrinkles and furrows (Tian et al., 2005). Some of the methods that are usually applied to extract appearance information are: Gabor Features, Local Binary Pattern (LBP) features, and Haar-like features.

Gabor Features

This method depends on filtering the input image with a Gabor filter to extract features. It gives the highest response at edges and at points where texture changes. A Gabor filter can be represented in the space domain using complex exponential notation as:

$$F_{k_0} = \frac{k_0^2}{\sigma^2} \exp\left(-\frac{(k_0)^2 x^2}{2\sigma^2}\right) \left(\exp(ik_0 \cdot x) - \exp\left(-\frac{\sigma^2}{2}\right) \right) \quad (2.1)$$

where $x = (x, y)$ is the image location and k_0 is the peak response frequency (Khan, 2013). The Figure 2.15 and shows the transformation after the Gabor filter is applied on two facial expression images. Lyong and his partners applied a set of Gabor filters, which are multi-

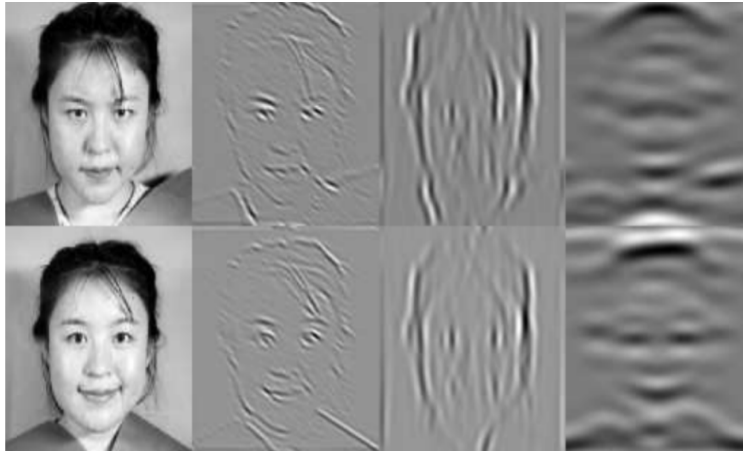


Figure 2.15: Gabor filter transformation (Lyons et al., 1998)

orientation and multi-resolution to code facial expression images (Lyons et al., 1998).

In the same task, Yu et al. (Yu and Bhanu, 2006) used Gabor features in their proposition of a novel genetically inspired learning method. Whereas Mattela and colleagues (Mattela and Gupta, 2018) use Gabor Mean Discrete Wavelet Transform as a reduction technique to decrease the dimension and redundancy of Gabor filters.

Local Binary Pattern (LBP)

The LBP operator calculates the brightness relationship between each pixel of an image and its local neighborhood, by thresholding the 3 x 3 neighborhood of each pixel with the center

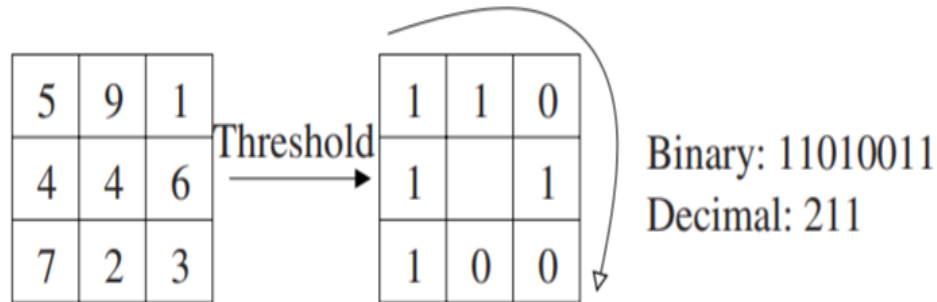


Figure 2.16: The basic LBP operator (Ahonen et al., 2004)

value (Khan, 2013), as shown in the Figure 2.16. After this operation, it encodes a binary sequence to form a local binary pattern, then the histogram is used as a feature description of the image (Huang et al., 2019), as shown in Figure 2.17.

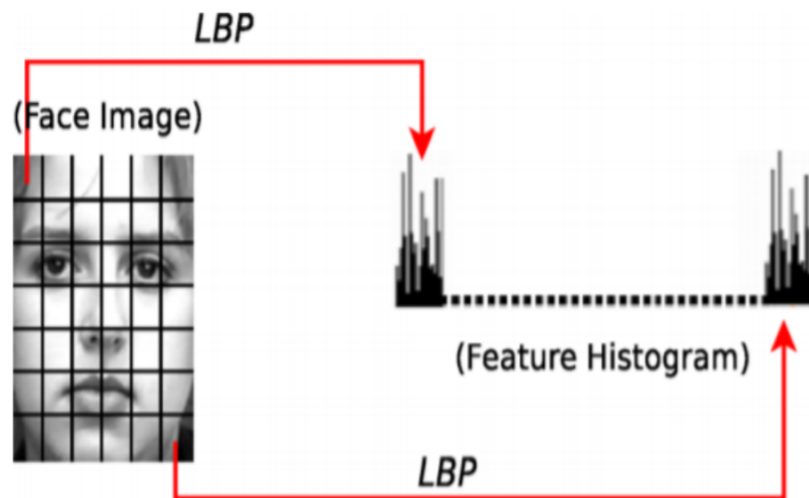


Figure 2.17: Extraction of LBP histogram from a facial image (Huang et al., 2019)

Haar-like Feature Extraction

The Haar-like Feature template was introduced by Viola and Jones (Viola and Jones, 2001). It combined edge, linear, center and diagonal features. We called them Haar-like because, they all follow the step function proposed by Alfred Haar. In two dimensions, The feature template is divided into a pair of adjacent rectangles, one light and one dark, shown in the Figure 2.18. The feature values of the template are defined as the differences between value of light rectangle pixels and dark rectangle pixels. The Haar eigenvalue reflects the gray-scale variation of the image (Huang et al., 2019). The process of Haar-like feature extraction on a face image is shown in the Figure 2.19. The advantage of a Haar-like feature is its calculation

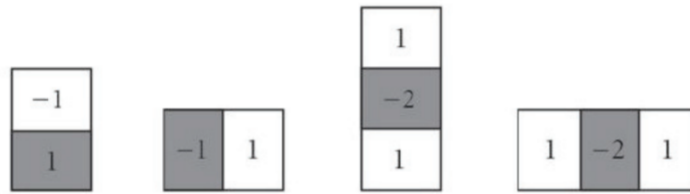


Figure 2.18: The basic haar-like feature template (Khan, 2013)



Figure 2.19: The process of Haar-like feature extraction (Khan, 2013)

speed, and in a constant time (Khan, 2013), that make researchers also applied it on facial expression analysis as (Yang et al., 2010) and (Whitehill and Omlin, 2006).

The motion features

The main purpose of this method is extracting some feature points or motion information from the regions of the features using sequential images. Feature point tracking and Optical flow are widely used and to understand more see (Mallick et al., 2016).

Optical Flow Method

Optical flow methods try to extract the features of the continuous moving(displacement) face image sequence (spatial information), an example is displayed in the Figure 2.20, using Horn–Schunck (HS) optical flow (Horn and Schunck, 1981), to combine the two-dimensional velocity field and the gray-scale (Huang et al., 2019). Optical flow have been used to extract features motion caused by human facial expressions in Yacoob work (Yacoob and Davis, 1996). Also, Cohn and his colleagues (Cohn et al., 1998) developed an optical flow-based approach

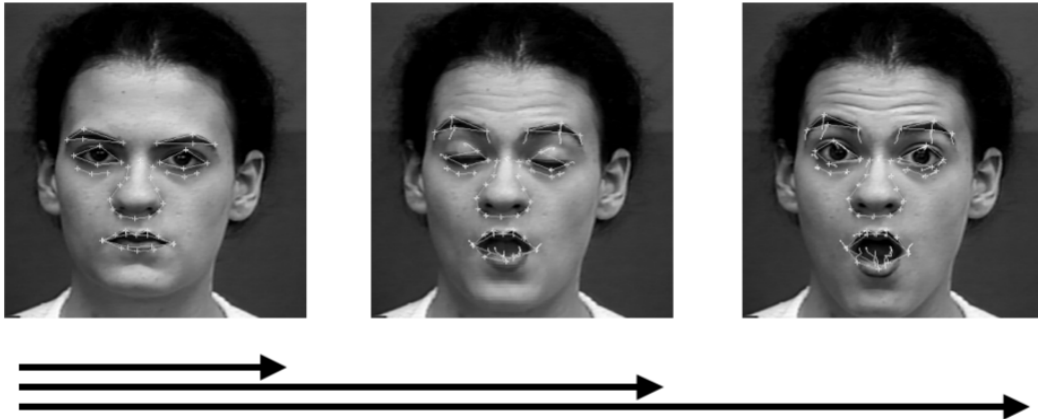
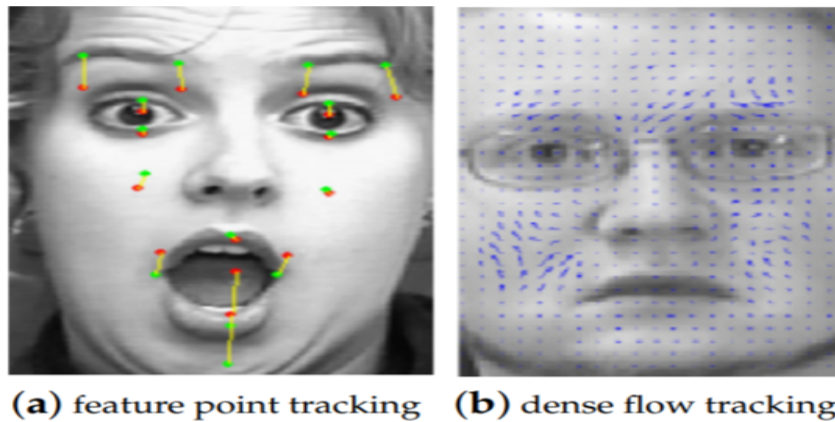


Figure 2.20: Feature point displacement (Cohn et al., 1998).

to capture emotional expression by automatically recognizing subtle changes in facial expressions. Sánchez et al. (Sánchez et al., 2011) systematically compare two optical flow-based FER methods, respectively referred to as feature point tracking and dense flow tracking, as shown in Figure 2.21.



(a) feature point tracking (b) dense flow tracking

Figure 2.21: Applications of optical flow-based methods on facial images (Sánchez et al., 2011)

Feature Point Tracking

Feature point tracking method synthesizes the input emotional expressions in relation to the displacement of the feature points. A feature point tracking method that combined the Kanade-Lucas-Tomasi (KLT) and Scale Invariant Feature Transform (SIFT) algorithms was proposed by Liu et al. (Liu et al., 2011). This method tracks targets stably and accurately when they change in size and attitude.

Tie et al. (Tie and Guan, 2012) present a variable 3D expression recognition model from video sequences. It extracts over 26 points from the video stream as the feature points of the face

model, then track these feature points with multiple particle filters.

2.3.3 Emotion Recognition/Classification

In this step, emotions (face expression) will be classified after the face detection and feature extraction, by various methods such as: Support vector Machine(SVM), Decision Trees(DT), Hidden Markov Models(HMM)([Eddy, 1996](#)), Adaptive Boosting(Adaboost)([Margineantu and Dietterich, 1997](#)), k-Nearest Neighbors(KNN), Naïve Bayes (NB)([Rish et al., 2001](#)), and Artificial Neural Networks(ANNs). We will present a set of existing research that used various of these methods.

Decision Trees

A decision tree is a tree structure used in classification and regression models. It works by dividing the datasets into smaller groups and developing them sequentially into nodes and leaves. The decision tree's branches indicate the category of the datasets. ([Saad et al., 2018](#))

An automatic recognition system has been developed ([Thorat et al., 2015](#)) for the six basic facial expressions in video streams using a decision tree. The system uses Viola Jones algorithm to detect face and face components and classify them using a decision tree. The system archived in an average of 76.43% correct classification of six basic expressions from video streams, with an expression error rate of 23.56%.

Salman and al. ([Salman et al., 2016](#)) proposed a method for emotion recognition from facial expressions. It is based on calculating six distances using Euclidean, Manhattan, or Minkowski distance between four parts of the face: eyebrow, eyes, nose, and mouth which better describe a facial expression. Decision tree is applied on two databases(JAFEE and COHEN). This system uses as inputs the six distances previously calculated for each face in order to have a facial expression classifying system with seven possible classes (six basic emotions plus neutral). Their results achieved a recognition rate of 89.20% and 90.61% respectively in JAFFE and COHEN databases.

k-Nearest Neighbours

k-NN is the most basic of all machine learning and classification algorithms; it saves all available examples and categorizes new ones using a similarity measure. As a result, the value K is utilized to classify data using simple histogram similarities. ([Kambi Beli and Guo, 2017](#))

Goutami et al. ([Panchal and Pushpalatha, 2017](#)) suggested a face emotion detection and recognition system. In this system, Local binary pattern and Asymmetric region local binary pattern methods are used to detect and extract features from face images. These methods are also used to reduce the dimensionality after the preprocessing step, and kNN classifier is used to predict

emotion. The proposed method was tested on the JAFFE database, where it achieved an accuracy of 95.051%.

For solving face recognition problems such as luminance and position, Kambi and his colleagues (Kambi Beli and Guo, 2017) presented an approach that combine two algorithms: the local binary pattern (LBP) which is used to extract facial features, and k-nearest neighbor for image classifications. Their experiment was conducted on the CMU PIE facial database and the LFW (Labeled Faces in the Wild) dataset, which achieved an accuracy of 99.26% and 85.71%, respectively.

Naïve Bayes (NB)

Cohen et al.(Cohen et al., 2002) proposed a system for recognizing emotions in video sequences by facial expressions using the Tree-Augmented-Naive Bayes (TAN). They provide an algorithm for determining the optimum TAN structure by learning the connections between facial features. They observed that employing this TAN structure produces substantially better results than using simpler NB classifiers. The results achieved were as follows:

- For Person Independent Tests: The system can determinate now with about 77% accuracy determine whether a person is making a negative or positive facial expression.
- For Person dependent Tests: The system can determinate now with 87-92% accuracy whether a person displays a negative or a positive facial expression.

Next, Cohen and his partners (Cohen et al., 2003) make another experiment with continuous video input. For classifying the expressions from the video, they used different Bayesian classifiers: Naïve Bayes and Gaussian Tree-Augmented Naïve Bayes(TAN). They changed the distribution from Gaussian to Cauchy because of the ability of Cauchy to account for heavy tail distributions, and used Gaussian Tree-Augmented Naïve Bayes(TAN) classifiers to learn the dependencies among different facial motion features. Using their own database, the results for person-dependent were: NB (Gaussian): 79.36% | NB (Cauchy):80.05% |TAN: 83.31%, and for Person Independent Tests were: NB (Gaussian): 60.23%| NB (Cauchy): 64.77% | TAN: 66.53%. They also make Person Independent Tests on the Cohn-Kanade DB where they achieved the following accuracy: (Gaussian): 67.03%| NB (Cauchy): 68.14%| TAN: 73.22%.

Support Vector Machines (SVM)

The Support Vector Machine (SVM) is a supervised machine learning technique that is typically used in classification and regression tasks.

Bartlett and his partners (Bartlett et al., 2003) proposed a system for facial emotion recognition in video stream. As a feature extraction method, they applied Gabor Wavelets and use SVM and a combined method (SVM with Ada-Boost). The system have been tested on Cohn-Kanade

database, and achieved when using SVM 84.8%, and 88.8% when using SVM with Ada-Boost, the result arrived to 90.7% with RBF Kernel.

In the same task of real time approach for facial emotion recognition, Philipp Michel and Rana El Kaliouby ([Michel and El Kaliouby, 2003](#)) employed an automatic facial feature tracker to perform face localization and feature extraction, and a Support Vector Machine as a classifier. The proposed method was tested with tree different scenarios:

- Person independent: 71.8%
- Person dependent(train and test data supplied by expert): 87.5%
- Person dependent(train and test data supplied by 6 users during ad-hoc interaction): 60.7%.

A study was conducted in Massachusetts Institute of Technology by James P. Skelley ([Skelley, 2005](#)) to train a Man-Machine Interface (MMI) to recognize facial expressions, by using Optical Flow and texture-value Combined Feature to extract the feature from Cohn-Kanade and HID Databases. Then, classifying face expression by using Support Vector Machines. It achieved a 94.2% successful classification rate for groups of subjects, and a 92% average for individuals.

Anderson and McOwan ([Anderson and McOwan, 2006](#)) used spatial ratio template tracker for face locating, and performed optical flow on it using multichannel gradient-model (MCGM). The extracted motion signatures from CMU-Pittsburg AU coded database and a non-expressive database are then classified using SVM and MLPs. Both methods gave almost similar performance (absolute recognition rate: with MLP:81.82%, with SVM:80.52%), they decided to go with the use of SVMs because SVMs had a much lower False Acceptance Rate (FAR) when compared to MLPs.

The effect of partial occluding on facial expression recognition was investigated by Kotsis et al. ([Kotsia et al., 2008](#)) in order to specify which portion of the face contains more discriminant information for each facial expression. Gabor wavelets, the DNMF algorithm and shape-based SVMs have been used to achieve recognition on Cohn-Kanade and JAFFE databases. In non-occluded experiments, the recognition rate when using the Cohn-Kanade: 91.4% with SVM.

Youssef and his colleagues ([Youssef et al., 2013](#)) used Kinect depth video with SVM and kNN for detecting Autism Spectrum Disorders (ASDs) in children. They consider the six universal emotions and find that ?? has the best recognition rate of 39%.

Ghimire and Lee ([Ghimire and Lee, 2013](#)) extracted geometric features from the sequences of facial expression images. For facial expression recognition, they used two methods, the first is based on multi-class Ada-Boost algorithm, the second was SVM classifier, and it achieved an accuracy of 97.35% with the extended Cohn-Kanade (CK+)database. They found that the landmarks moved in comparable ways, with facial expressions changing throughout time,

independent of ethnicity, age, or gender.

Myunghoon Suk and Balakrishnan Prabhakaran ([Suk and Prabhakaran, 2014](#)) presented in their research paper a mobile application for real time facial expression recognition running on a smartphone with a camera. The proposed system used a set of Support Vector Machines (SVMs) for classifying six basic emotions and neutral expression along with checking mouth status. The facial expression features are extracted by Active Shape Model (ASM) and generated by the displacement between neutral and expression features. They report experimental results with 86% of accuracy with 10 folds cross validation in 309 video samples of the extended Cohn-Kanade (CK+) database, and 72% in their mobile application running on Samsung Galaxy S3 with 2.4 fps.

In the same field, an approach based on PCA and LBP algorithms was suggested by Abdulrahman and Eleyan ([Abdulrahman and Eleyan, 2015](#)) where an SVM have been used as classifier. The results of all trials conducted on the Japanese Female Facial Expression and Mevlana University Facial Expression datasets show that PCA+SVM has an average recognition rate of 87% and 77%, respectively.

Hidden Markov Models (HMM)

A video analysis system that aims at expression recognition was proposed by Pardas and Bonafonte ([Pardàs and Bonafonte, 2002](#)), and tested on Cohn-Kanade database. Its first step was extracting the Facial Animation Parameters (FAPs) using an improved Active Contour algorithm, then recognizing emotion with HMM classifier. Overall efficiency was of 84% (across 6 prototypic expressions) and experiments with joy, surprise and anger was 98%, and with joy, surprise and sadness was 95%. They found that FAPs convey the necessary information to extract emotions. The algorithm performs well when differentiating individual expressions and may also be used to extract expressions in long video sequences where speech is intermingled with silent frames, however with lesser accuracy.

Cohen et al. ([Cohen et al., 2003](#)) proposed a real-time system for emotion classification from video using PBVD Tracker to extract Motion Units (MUs). They suggest the use of HMMs to automatically split a video into different expression segments. Person-dependent facial expression recognition rate of 78.49% with emotion-specific HMM and 82.46% with multi-level HMM on their own database. For Person Independent Tests, HMM rate of 55.71% and ML-HMM rate of 58.63%.

A novel multi-stream-HMM(MS-HMM) automatic facial expression recognition system have been described by ([Aleksic and Katsagelos, 2006](#)). It uses MPEG-4 compliant facial features followed by PCA to reduce dimensionality before giving it to the HMM. They use 284 recordings of 90 subjects from Cohn-Kanade database. The accuracy result Using HMM with only eyebrow FAPs gave 58.8% rate, and with only outer lip FAPs rate of 87.32%, for Joint FAPs gave 88.73%

rate. After assigning stream weights and then using a MS-HMM, the system rate of 93.66% with outer lip having more weight than eyebrows.

Artificial Neural Networks (ANNs)

Kundu and his colleagues ([Kundu and Singh, 2013](#)) proposed a machine learning system for human emotion recognition based on facial expressions using geometrical features. After image acquisition, face and different regions are marked manually (segmentation). Then, features are extracted from the segmented face and neural network with one hidden layer has been used to classify these extracted features. The system has a moderately high recognition rate of 75% with a very small feature set using Japanese Female Facial Expression (JAFPE).

ANN was also used in emotion recognition based on speech signal. The extracted features are related to statistics of pitch, formants, and energy contours, as well as spectral, perceptual and temporal features, jitter, and shimmer. Without raising the system's complexity or computing time, a success rate of 85% or even higher can be attained ([Hendy and Farag, 2013](#)).

Tanwi Mallick et al. ([Mallick et al., 2016](#)) presented facial emotion recognition system using Kinect 1.0 data and the Kinect Face Tracking Library (KFTL). They detect various Action Units of the face from the feature points extracted by KFTL and then recognize emotions by multiple as well as single ANN. The system achieves an accuracy of 40% or more for five out of six emotions (excluding Fear), and over 64% for four of them (further excluding Happiness), they observed that single ANN behaves better, though the recognition rates of fear and happiness are unsatisfactory.

In conventional facial emotion recognition approaches, feature extraction and classification are designed manually and separately, which means that these two phases cannot be optimized simultaneously. These approaches are obviously less dependent on data and hardware, but it gives good results in small data, whereas researchers need to deal with big data, this became possible with the advent of deep learning.

2.4 Deep Learning-based Approaches

Deep neural networks have shown to outperform traditional methods that has been applied to the field of emotions recognition, among them CNNs and RNNs. Deep learning-based approaches greatly reduce the dependence on feature extraction by employing an "end-to-end" learning directly from input data to classification result. They also have the potential to manage large amounts of data.

Table 2.1: Subject-independent comparison with AlexNet results (accuracy%) (Mollahosseini et al., 2016)

	Proposed architecture	AlexNet
MultiPie dataset	94.7	94.8
MMI dataset	77.9	56.0
DISFRA dataset	55.0	56.1
FEBRA dataset	76.7	77.4
SFEW dataset	47.7	48.6
CK+ dataset	93.2	92.2
FER2013 dataset	66.4	61.1

2.4.1 Convolutional Neural networks (CNN)

Many researchers have adopted deep learning approach for recognizing facial expressions, one of them is Jung and his partners (Jung et al., 2015). Basically, they combined two models: The first use a CNN to extract temporal appearance data from image sequences, while the second employs a fully connected DNN to extract temporal geometry features from temporal face landmark points. This network is called the deep temporal appearance-geometry network (DTAGN), and it reached an accuracy of 97.25%.

Mollahosseini et al. (Mollahosseini et al., 2016) conducted a comprehensive experiment on seven available facial expression databases: MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013. Their proposed architecture consists of two convolutional layers each followed by max pooling and then four inception layers. The inception layers improve the depth and width of the network while maintaining the same computational budget. In terms of both accuracy and training time, the results displayed in Table 2.1 are comparable to or better than state-of-the-art approaches and classic convolutional neural networks.

The ordinary CNN extract only the spatial relations of the input data and ignore the temporal relations of them if they are part of a sequenced data. To overcome this problem, CNN was further extended to 3D Convolutional Neural Networks (3D-CNNs).

2.4.2 3D-CNN

According to psychologists, expressions are invoked by certain facial parts that contain the most descriptive information for representing them. This observation bring Liu et al. (Liu et al., 2014) to the same spirit of the Deformable Part Model (DPM) (Felzenszwalb et al., 2009). They incorporated a deformable facial parts learning module into a 3D CNN, that can detect a particular facial action part under structured spatial constraint, and obtain the deformable part detection maps to serve as the part-based representation for expression recognition. the

average expression recognition rates on CK+ , MMI and FERA datasets are 87.9%, 62.2%, and 56.3%, respectively.

3D-CNN was also adopted in combined architectures including the winner hybrid networks in the EmotiW 2016 Challenge (Fan et al., 2016) that combines 3D convolutional networks and RNN in a late-fusion fashion. RNN encodes motion after taking appearance features collected by CNN over individual video frames as input, while 3D CNN models appearance and motion of video simultaneously. This system achieved a recognition accuracy of 59.02%.

Another proposed study (Singh and Fang, 2020) that has explored different architectures of neural networks and demonstrated (CNN+RNN)+ 3D CNN multi-model architecture perform better. It complements each other by learning emotion features from the audio signal (CNN+RNN) and also learning emotion features from face expression in video frames (3D CNN). Using IEMOCAP dataset, this combined architecture gave an emotion prediction accuracy of 54.0% among four emotions and 71.75% among three emotions.

Many approaches have adopted a classical CNN directly for facial emotion recognition. However, because CNN-based methods cannot reflect temporal variations in the facial components, a recent hybrid approach combining a CNN with RNN.

2.4.3 CNN with RNN

Kahou et al in their work (Ebrahimi Kahou et al., 2015) presented a complete system for emotion recognition in video in the Wild (EmotiW2015) Challenge. They focus their presentation and experimental analysis on a hybrid CNN-RNN architecture that can outperform a previously applied CNN approach using temporal averaging for aggregation. The resulting test performance was only 49.907%, this could be explained by the fact that the entire dataset has been examined, resulting in overfitting.

Kim et al. (Kim et al., 2017) used an approach to video emotion recognition in which CNN was used to learn the spatial features of representative state frameworks, then LSTM was used for facial expression to learn the temporal features to represent the spatial feature. The proposed method was applied to the MMI dataset whereit achieved an accuracy of 78.61%.

Another network architecture to deal with the same task (Hasani and Mahoor, 2017) comprises of 3D inception ResNet layers followed by an LSTM module that extracts spatial and temporal correlations within facial images as well as between distinct frames in a video. Four known databases were used for evaluation are: CK+, MMI, FERA and DISFA. The accuracy results on the four databases were 67.52%, 54.76%, 41.93%, 40.51%, respectively.

For solving the problem of sudden changes in illumination and finding the proper alignment of the feature set, Jain and his partners proposed a multi-angle optimal pattern-based deep learning (MAOP-DL) method (Jain et al., 2020). Initially, this method begins by removing the

background from images and isolating the foreground, and then extracts the texture patterns and the relevant key features of the facial points. The relevant features are then selectively extracted, and an LSTM-CNN is employed to predict the required label for the facial expressions. The suggested MAOP-DL proves its effectiveness on CK+ and MMI databases, and confirms their assurance of wide applicability in recent applications.

2.4.4 TCN

Without the need for complex recurrent networks, the TCN aims to capture the temporal features of emotion classification. It was used by Manasi Gund et al. ([Gund et al., 2021](#)) where they were concerned in dynamic facial expression detection because some information better delivered through moving faces. Their TCN proposed model takes facial features as input, it was originally trained and tested on the CK+ dataset and produced 99.57% accuracy. They found that of all other models using the CK+ data set, the TCN model proves to be the simplest with the best results and small computational cost that incurs by taking only the features rather than the whole images.

TCN is used to recognize human emotion with various uni-model features such as physiological signals ([Yang and Liu, 2019](#)), by extracting high-level features from an electroencephalogram (EEG) while considering the temporal dependence.

Also, TCN used to extract a specific human emotion ([Bargshady et al., 2020](#); [Feng, 2019](#)). Detailed model of the network structure of TCN-based model for emotion recognition is presented in the Figure 2.22.

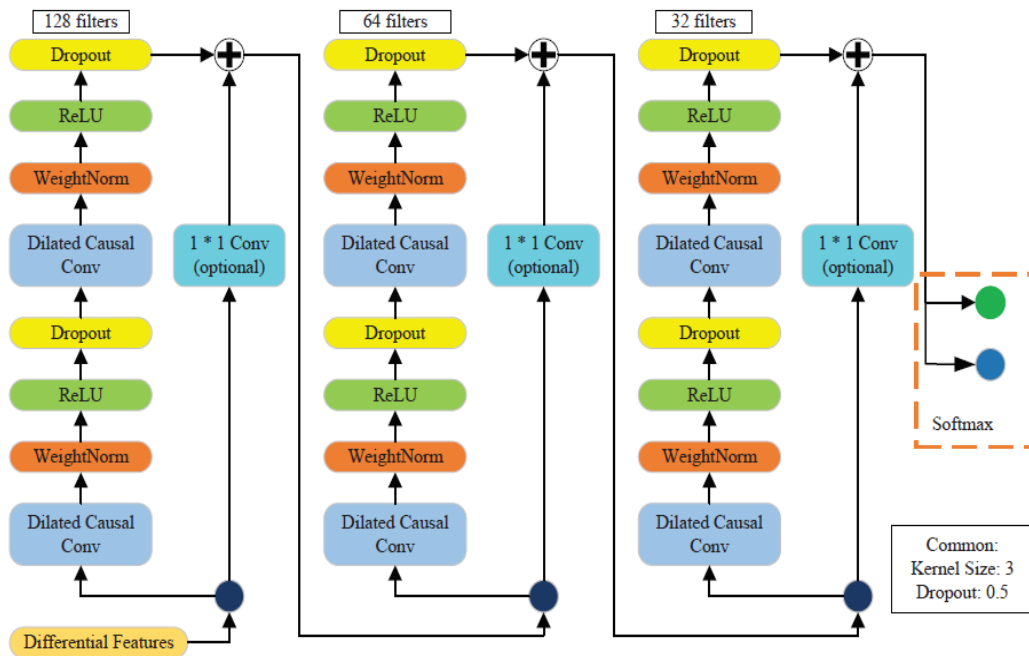


Figure 2.22: A network structure of TCN-based model (Yang and Liu, 2019)

2.5 Datasets

Training and testing on existing datasets is a frequently used method of expression recognition. We introduce some popular databases related to facial expression, consisting of 2D and 3D video sequences and still images.

2.5.1 C-K (Cohn-Kanade) database

It is a video-based encoded facial database of subtractive expressions, including 486 sequences from 97 poses. The peak expression is fully FACS coded. The number of subjects included is 100. Released in the year 2000, for the cause of advancing research into detecting individual facial expressions. It has become one of the most commonly used data sets. But the drawbacks found were that emotion labels were not validated and non-availability of common performance measure in order to compare with standard algorithms.(Lucey et al., 2010)

2.5.2 Cohn-Kanade Dataset (CK+)

CK+ is a dataset of 593 sequences from 123 subjects, among which 327 sequences have emotion labels. The dataset contains seven expressions including anger, disgust, fear, happy, sad, surprise, and contempt. The images have pixel resolutions of $640 * 480$ and $640 * 490$ with 8-bit precision for gray-scale values.

2.5.3 Japanese Female Facial Expressions (JAFFE)

This database contains 219 images of 7 facial expressions posed by 10 Japanese female models. Each image has been rated on 6 emotion classes on 60 Japanese subjects. As it involves multiple subjects exhibiting multiple emotions it has been extensively used in the field of research. The original size of each facial image is 256 pixels * 256 pixels (Lyons et al., 1998).

2.5.4 MMI dataset

A database which consists of more than 1500 samples of both static images and image sequences of faces from 19 male and female subjects of varying ethnic background. It is fully annotated for the presence of AUs in the video sequences. The original size of each facial image is 720 by 576 pixels (Pantic et al., 2005).

2.5.5 IEMOCAP dataset

The "interactive emotional dyadic motion capture database" (IEMOCAP) was collected by the SAIL Laboratory at the University of Southern California (USC) by the Speech Analysis and Interpretation team. This database was created using markers on the face, head, and hands to record ten actors in dyadic sessions, during scripted and spontaneous spoken communication scenarios, these markers provide detailed information about their face expression and hand movements. To elicit specific types of emotions (happiness, anger, sadness, frustration, and neutral state), the ten actors ran selected emotional scenarios and improvised scenarios. The corpus contains approximately twelve hours of data (Busso et al., 2008).

2.6 Evaluation of emotion recognition system

During the training process, the evaluation metric is quite important, and there are numerous measures to evaluate deep learning model quality. (Hossin and Sulaiman, 2015). To evaluate our emotion recognition model, we use the following metrics: precision, recall, accuracy, and F-measure.

- **Precision** is the fraction of automatic annotations of emotion i that are correctly recognized, It is defined as:

$$Precision = TP / (TP + FP) \quad (2.2)$$

- **Recall** is the number of correct recognition of emotion i over the actual number of images with emotion i . It is defined as:

$$Recall = TP / (TP + FN) \quad (2.3)$$

- **Accuracy** is the ratio of true outcomes (both true positive to true negative) to the total number of cases examined, and it is defined as:

$$Accuracy(ACC) = (TP + TN)/(TP + TN + FP + FN) \quad (2.4)$$

- **F-measure** represents the harmonic mean between recall and precision values, and it is defined as:

$$F - measure = 2 \times Precision \times Recall/(Precision + Recall) \quad (2.5)$$

Where:

- TP is the number of true positives in the dataset.
- FN is the number of false negatives in the dataset.
- FP is the number of false positives in the dataset.
- TN is the number of true negative in the dataset.

2.7 Conclusion

In this chapter, we mentioned some of the first research in the field of facial emotion recognition, and we also touched on some of the work done in this field using machine learning and deep learning algorithms, and how facial emotions are recognized in the two approaches (machine learning and deep learning). Finally, we presented some common databases related to facial expressions that consist of videos and images.

CHAPTER 3

EXPERIMENT

3.1 Introduction

This chapter presents our experiment with a model of facial emotion recognition using deep learning. Among the methods of deep learning, we have chosen temporal convolution network (TCN). We also train a convolution neural network on the extracted frames of videos. Below, we will talk about the selected network architectures and the obtained results.

3.2 Network Architecture

From what we saw in the proposed model in (Gund et al., 2021) which was used to classify the facial emotions in image, physiological signals (Yang and Liu, 2019) and to extract a specific human emotion (Bargshady et al., 2020; Feng, 2019). These works are based on temporal convolution network and we found that it gave good results, which prompted us to use this structure in our first experiment.

What makes a temporary convolutional network different from LSTM / GRU:

- TCNs exhibit longer memory than recurrent architectures with the same capacity.
- Performs better than LSTM/GRU on a vast range of tasks (Seq. MNIST, Copy Memory, ...).
- Parallelism (convolutional layers), flexible receptive field size (possible to specify how far the model can see), stable gradients (backpropagation through time, vanishing gradients)...

The Visualization of a stack of dilated causal convolutional layers in Figure 3.1.

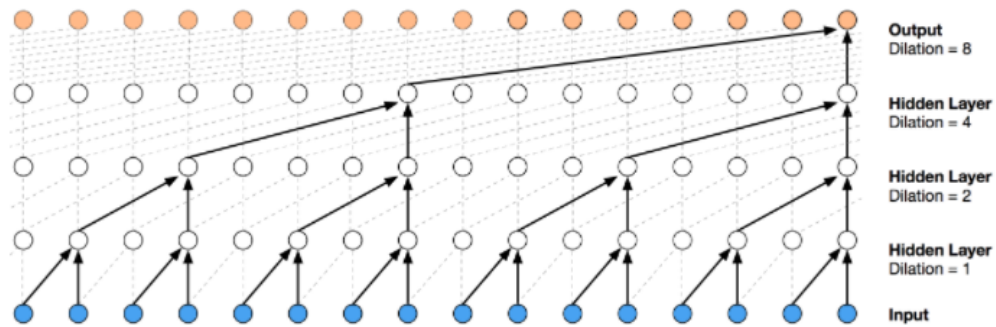


Figure 3.1: Visualization of a stack of dilated causal convolutional layers

Arguments

The TCN model is characterized by the following parameters.

- `nb-filters`: Integer. The number of filters to use in the convolutional layers. Would be similar to units for LSTM, these filters can be a list.
- `kernel-size`: Integer. The size of the kernel to use in each convolutional layer.
- `dilations`: List/Tuple. A dilation list. Example is: `[1, 2, 4, 8, 16, 32, 64]`.
- `nb-stacks`: Integer. The number of stacks of residual blocks to use.
- `padding`: String. The padding to use in the convolutions. 'causal' for a causal network (as in the original implementation) and 'same' for a non-causal network.
- `use-skip-connections`: Boolean. If we want to add skip connections from input to each residual block.
- `return-sequences`: Boolean. Whether to return the last output in the output sequence, or the full sequence.
- `dropout-rate`: Float between 0 and 1. Fraction of the input units to drop.
- `activation`: The activation used in the residual blocks $o = \text{activation}(x + F(x))$.
- `kernel-initializer`: Initializer for the kernel weights matrix (Conv1D).
- `use-batch-norm`: Whether to use batch normalization in the residual layers or not.
- `use-layer-norm`: Whether to use layer normalization in the residual layers or not.
- `use-weight-norm`: Whether to use weight normalization in the residual layers or not.
- `kwargs`: Any other set of arguments for configuring the parent class Layer. For example `"name=str"`, Name of the model. Use unique names when using multiple TCN.

Non-causal TCN:

Making the TCN architecture non-causal allows it to take the future into consideration to do its prediction as shown in the figure 3.2.

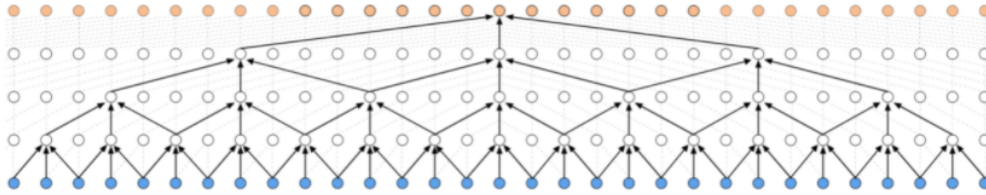


Figure 3.2: Non-Causal TCN - $ks = 3$, dilations = $[1, 2, 4, 8]$, 1 block

When initializing the TCN layers, use `padding='valid'` or `padding='same'` to utilize a non-causal TCN.

The arguments of the TCN architecture that we try to use it in our experiment is shown in the figure 3.3.

```

[10]: from tcn import TCN, tcn_full_summary
from tensorflow.keras.layers import Input, Embedding, Dense, Dropout, SpatialDropout1D
from tensorflow.keras.layers import concatenate, GlobalAveragePooling1D, GlobalMaxPooling1D
from tensorflow.keras.models import Model

model0 = compiled_tcn(14*16384, # type: int
7, # type: int
128, # type: int
7, # type: int
[1,2,4,8], # type: List[int]
1, # type: int
None, # type: int
output_len=1, # type: int
padding='causal', # type: str
use_skip_connections=False, # type: bool
return_sequences=True,
regression=False, # type: bool
dropout_rate=0.05, # type: float
name='tcn', # type: str,
kernel_initializer='he_normal', # type: str,
activation='relu', # type: str,
opt='adam',
lr=0.002,
use_batch_norm=False,
use_layer_norm=False,
use_weight_norm=False)

model0.summary()

```

Figure 3.3: TCN Architecture arguments

Since, the experiment using TCN was not finished, we make another one using simple convolution network, its architecture is shown in the figure 3.4

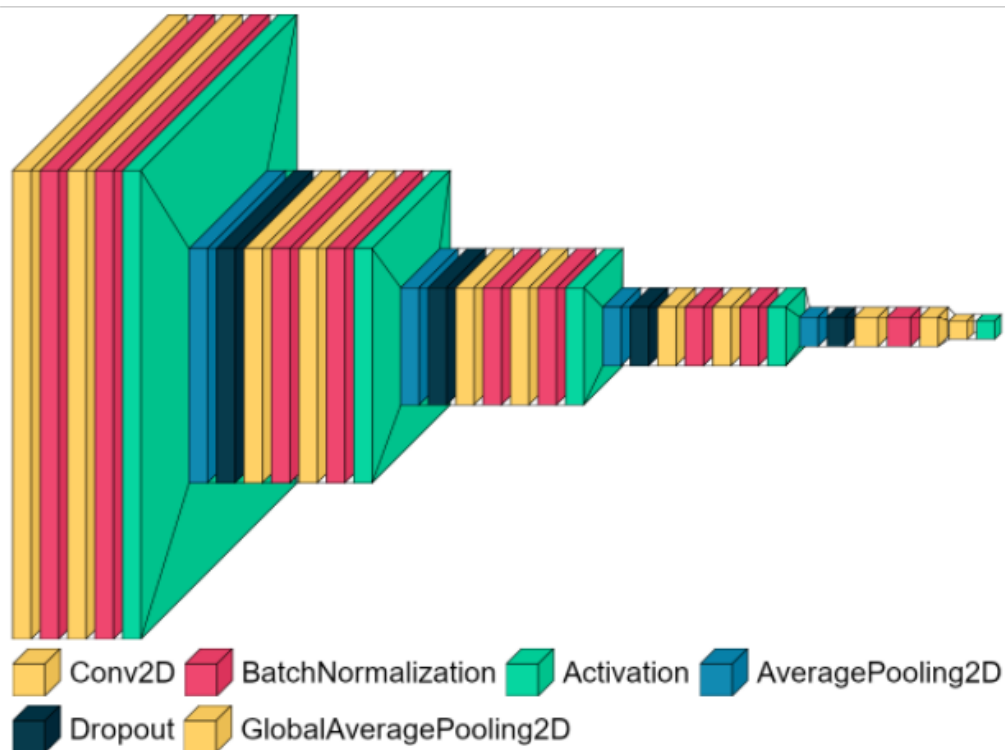


Figure 3.4: CNN Architecture

The information about layers of the CNN architecture is shown in the following Figures 3.5, 3.6, 3.7 and 3.8.

```
Model: "sequential"
```

Layer (type)	Output Shape	Param #
image_array (Conv2D)	(None, 128, 128, 16)	2368
batch_normalization (BatchNo	(None, 128, 128, 16)	64
conv2d (Conv2D)	(None, 128, 128, 16)	12560
batch_normalization_1 (Batch	(None, 128, 128, 16)	64
activation (Activation)	(None, 128, 128, 16)	0
average_pooling2d (AveragePo	(None, 64, 64, 16)	0
dropout (Dropout)	(None, 64, 64, 16)	0
conv2d_1 (Conv2D)	(None, 64, 64, 32)	12832

Figure 3.5: CNN architecture part 1

batch_normalization_2 (Batch Normalization)	(None, 64, 64, 32)	128
conv2d_2 (Conv2D)	(None, 64, 64, 32)	25632
batch_normalization_3 (Batch Normalization)	(None, 64, 64, 32)	128
activation_1 (Activation)	(None, 64, 64, 32)	0
average_pooling2d_1 (Average Pooling)	(None, 32, 32, 32)	0
dropout_1 (Dropout)	(None, 32, 32, 32)	0
conv2d_3 (Conv2D)	(None, 32, 32, 64)	18496
batch_normalization_4 (Batch Normalization)	(None, 32, 32, 64)	256
conv2d_4 (Conv2D)	(None, 32, 32, 64)	36928

Figure 3.6: CNN architecture part 2

batch_normalization_5 (Batch Normalization)	(None, 32, 32, 64)	256
activation_2 (Activation)	(None, 32, 32, 64)	0
average_pooling2d_2 (Average Pooling)	(None, 16, 16, 64)	0
dropout_2 (Dropout)	(None, 16, 16, 64)	0
conv2d_5 (Conv2D)	(None, 16, 16, 128)	73856
batch_normalization_6 (Batch Normalization)	(None, 16, 16, 128)	512
conv2d_6 (Conv2D)	(None, 16, 16, 128)	147584
batch_normalization_7 (Batch Normalization)	(None, 16, 16, 128)	512
activation_3 (Activation)	(None, 16, 16, 128)	0

Figure 3.7: CNN architecture part 3

average_pooling2d_3 (Average (None, 8, 8, 128))		0
dropout_3 (Dropout)	(None, 8, 8, 128)	0
conv2d_7 (Conv2D)	(None, 8, 8, 256)	295168
batch_normalization_8 (Batch Normalization)	(None, 8, 8, 256)	1024
conv2d_8 (Conv2D)	(None, 8, 8, 7)	16135
global_average_pooling2d (Global Average Pooling)	(None, 7)	0
predictions (Activation)	(None, 7)	0
=====		

Figure 3.8: CNN architecture part 4

3.3 Implementation Setup

3.3.1 Dataset

In this experiment, we want to recognize human facial expressions in the video. After looking to various datasets, we choose Acted Facial Expressions In The Wild dataset (Dhall et al., 2012). In this dataset, various facial emotions, natural head posture movements, individuals of different genders and ages, and many subjects in the scene are captured. The video clips have been labelled for six basic expressions anger, disgust, fear, happiness, sadness, surprise and the neutral class. Figure 3.9 shows the different classes of AFEW databases, whereas Table 3.1 shown the attributes of AFEW dataset.

Table 3.1: AFEW database attributes (Dhall et al., 2012)

Attribute	Description
Length of sequences	300-5400 ms
Number of sequences	1426
Total number of expressions (incl. multiple subjects)	1747
Video format	AVI
Maximum number of clips of a subject	134
Minimum number of clips of a subject	1
Number of labelers	2
Number of subjects	330
Number of clips per expression	Anger: 194, Disgust: 123 Fear: 156, Sadness: 165 Happiness:387,Neutral:257, Surprise:144



Figure 3.9: One sampled frame from each of the 7 classes in AFEW

3.3.2 Data processing

Before using the AFEW dataset, we processed it as follows:

- We divided the videos into a training group and a test group.
- In each group, we converted each video into frames, the number of which varies from one

video to another.

- We specify the name of the video, the frames to which it belongs, as well as the category to which each category belongs, as shown in the figures 3.10, 3.11 below.

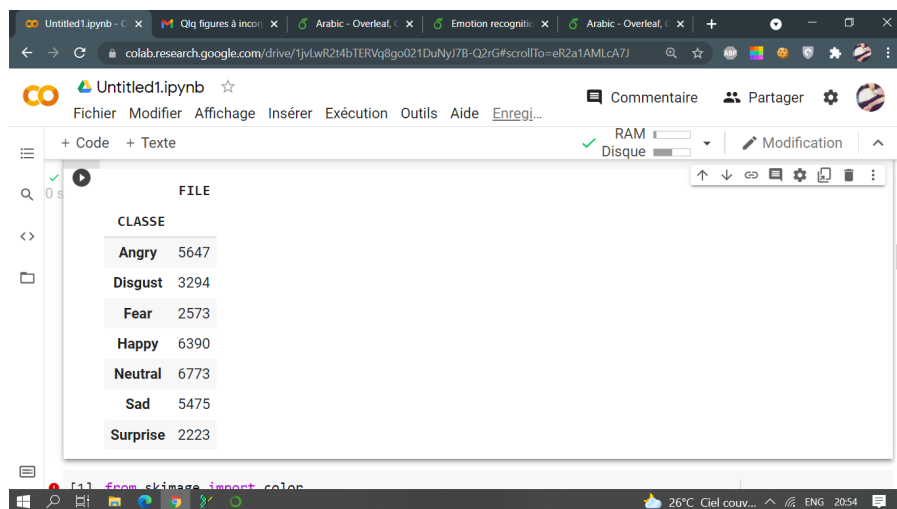


Figure 3.10: Train dataset

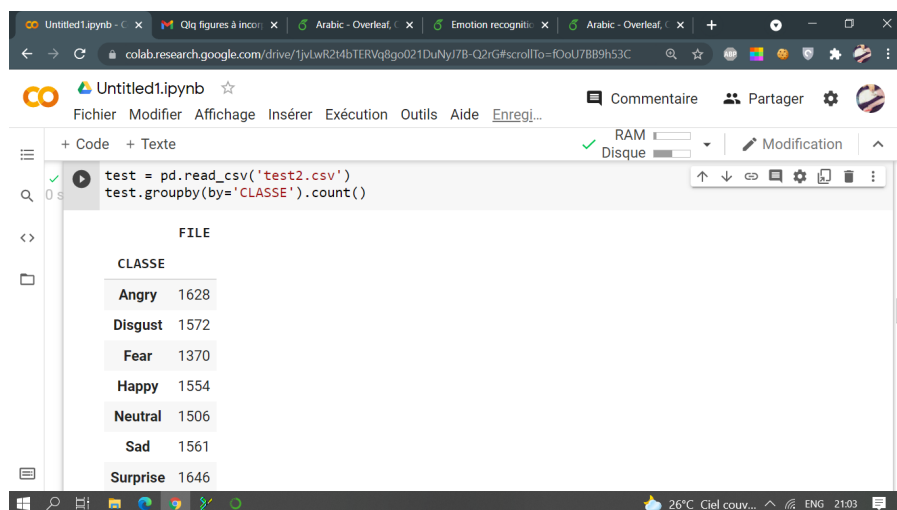
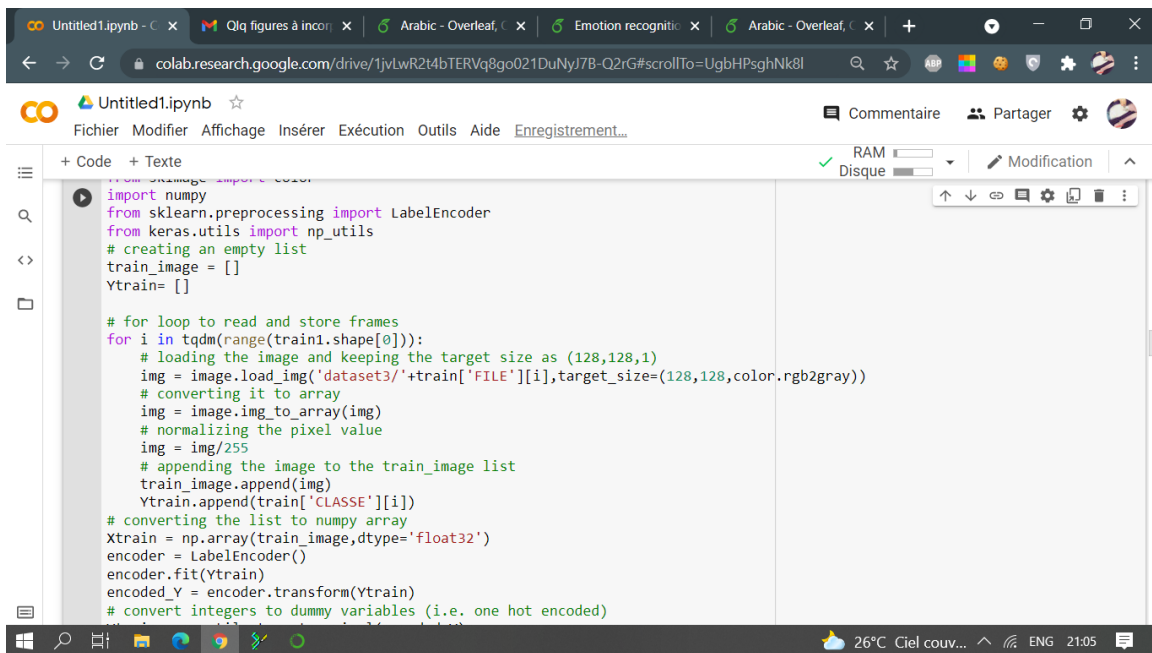


Figure 3.11: Test dataset

- We converted the extracted frames to gray-scale.
- We converted the gray-scales frames into a numpy matrix.
- We normalize the weights in the interval $[0,1]$, the Figure 3.12 shows the code.



```

import numpy
from sklearn.preprocessing import LabelEncoder
from keras.utils import np_utils
# creating an empty list
train_image = []
Ytrain= []

# for loop to read and store frames
for i in tqdm(range(train1.shape[0])):
    # loading the image and keeping the target size as (128,128,1)
    img = image.load_img('dataset3/'+train['FILE'][i],target_size=(128,128,color.rgb2gray))
    # converting it to array
    img = image.img_to_array(img)
    # normalizing the pixel value
    img = img/255
    # appending the image to the train_image list
    train_image.append(img)
    Ytrain.append(train['CLASSE'][i])
# converting the list to numpy array
Xtrain = np.array(train_image,dtype='float32')
encoder = LabelEncoder()
encoder.fit(Ytrain)
encoded_Y = encoder.transform(Ytrain)
# convert integers to dummy variables (i.e. one hot encoded)

```

Figure 3.12: Data processing code

3.3.3 Environment

In order to implement our model, we used **Anaconda** which is a free and open-source Python and R programming language distribution. The distribution comes with the Python interpreter and various packages related to machine learning and data science¹. Figure 3.13 represent the applications available by default in the anaconda browser:

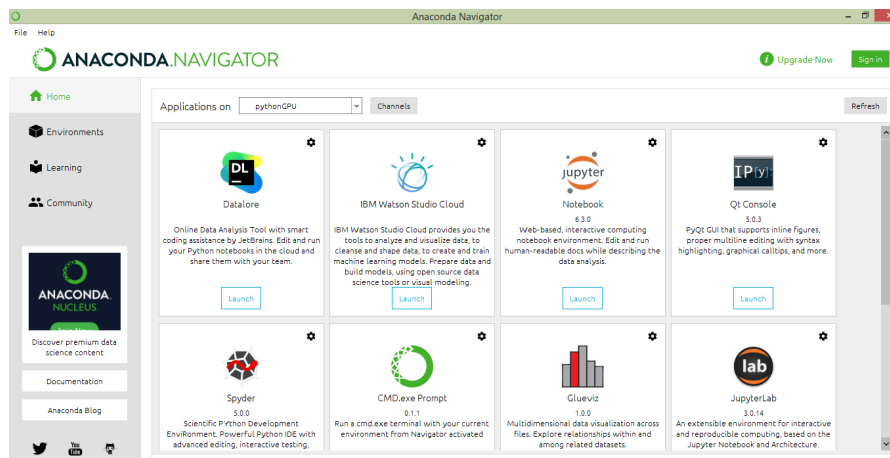


Figure 3.13: anaconda navigator

since we have to deal with big data We had to move to Google Colaboratory¹, which is a

¹<https://www.venturelessons.com/what-is-anaconda>,consulted in July 2021/

¹<https://colab.research.google.com,/intro.ipynb>,consulted in July 2021/

free Jupyter notebook environment that runs on Google cloud servers and allows users to take advantage of backend technology such as GPUs and TPUs. This allows you to accomplish anything you can in a Jupyter notebook on your local machine without having to install or set up a notebook on your local machine. The interface of jupyter notebook in Google Colaboratory environment is shown in the Figure 3.14.

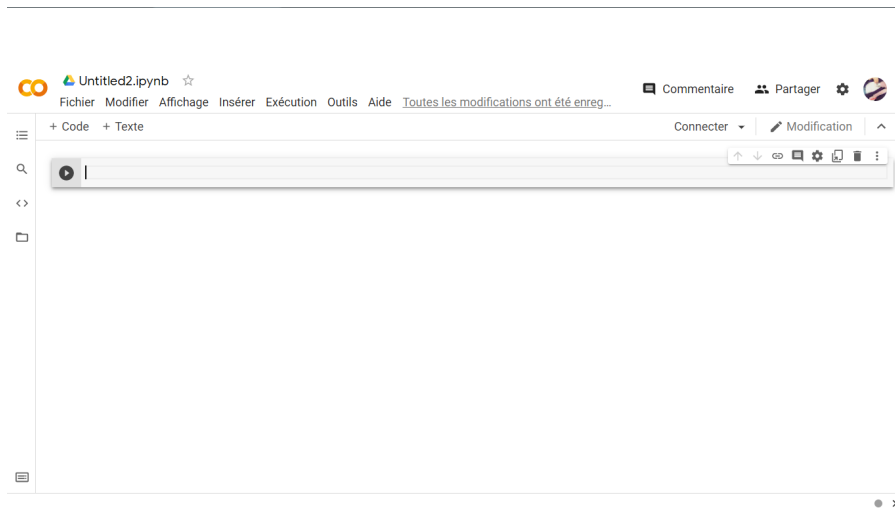


Figure 3.14: jupyter notebook in Google Colab

We used the Python3 programming language under the jupyter notebook (see Figure 3.15) to write the code.

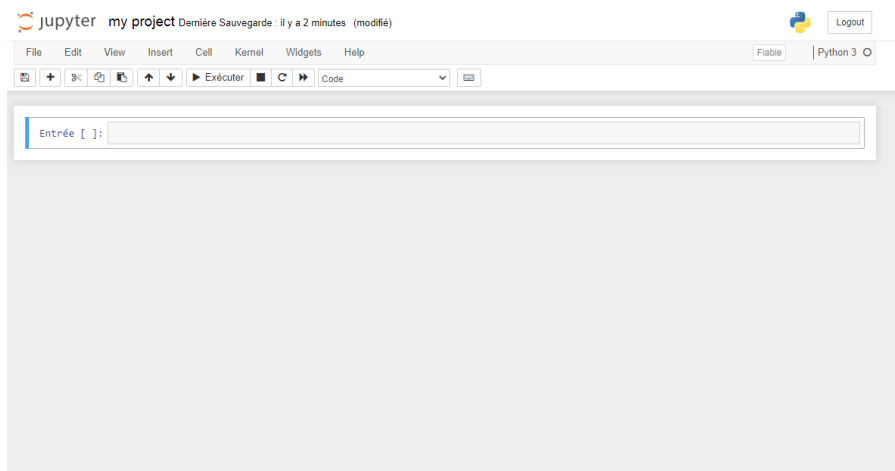


Figure 3.15: jupyter notebook interface

python is a programming language that has a large set of libraries that facilitate programming in deep learning.

Tensorflow² is an open source core library used to create deep learning models.

Keras³ is an Open Source Neural Network library.

Numpy⁴ is the fundamental library for scientific computing in Python. It is also used in procedures for quick operations on matrices, basic linear algebra, basic statistical operations and much more.

OpenCV⁵ is a python library which are widely used in operations on images such as shape / resolution, displaying, resizing, etc.

Pandas⁶ Library for data processing and analysis.

3.4 Results and Discussion

The experiment of emotion recognition in videos using temporal convolution network was not terminated because of the lack of resources and information, since this architecture is still new. We did not have enough time and required computing resources to complete it for the whole dataset. So, we have used a convolution neural network to detect the emotions in the collection of frames extracted from the videos. When creating the convolutional neural network model, we changed the work environment where we worked in a Google Collaboration environment(see Figure 3.14).

The curves 3.16 and 3.17 represents accuracy and loss, respectively.

²<https://www.tensorflow.org>,consulted in July 2021/

³<https://keras.io>,consulted in July 2021/

⁴<https://numpy.org>,consulted in July 2021/

⁵<https://opencv.org>,consulted in July 2021/

⁶<https://pandas.pydata.org>,consulted in July 2021/

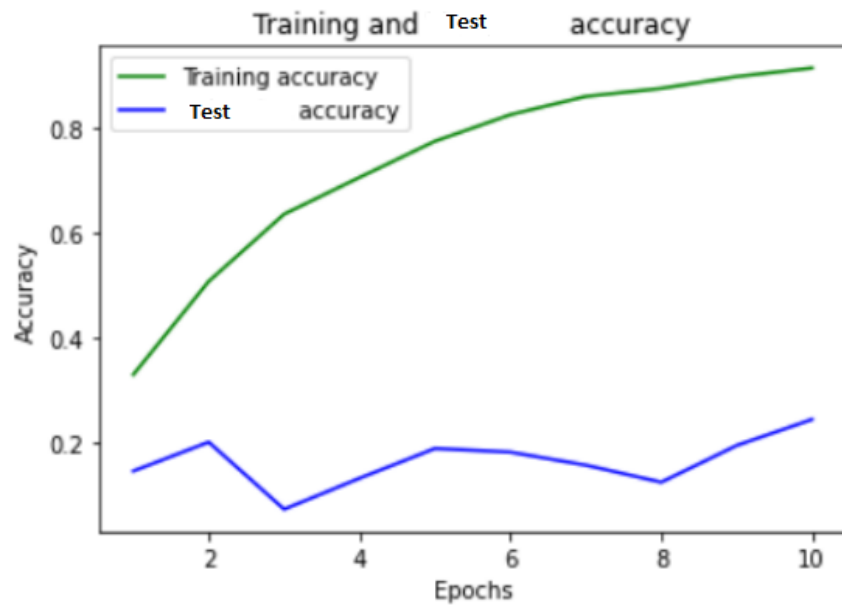


Figure 3.16: The training and test accuracy curve

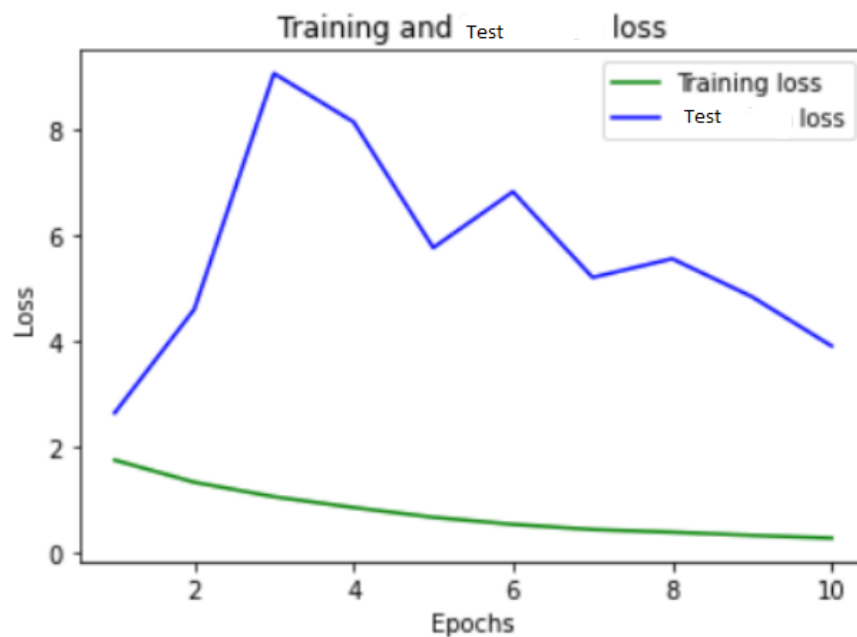


Figure 3.17: The training and test loss curve

The results were generally acceptable.

The factor affecting the results is that the data set was not distributed in an optimal manner. It includes noise in many samples. The figure 3.18 represents the confusion matrix which enables us to determine the accuracy of the model. The matrix shows the classification of emotion for each category.

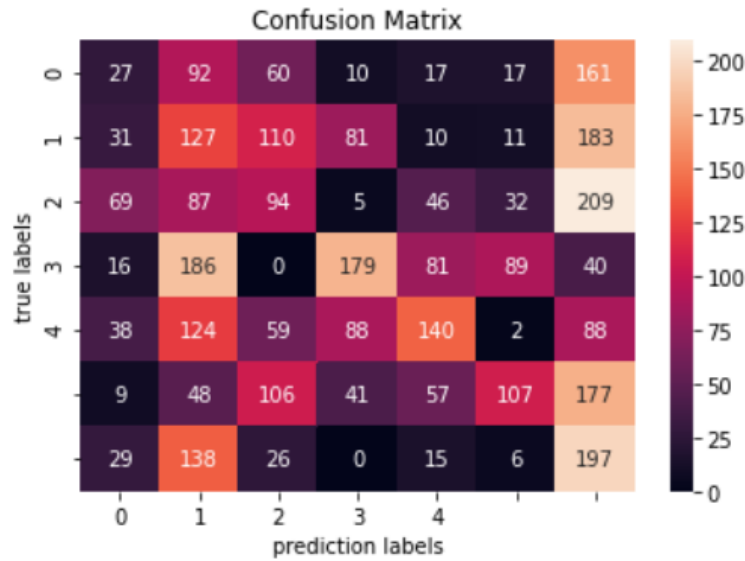


Figure 3.18: The confusion matrix for the

The Figure 3.19 presents the recall, the precision, f1-score and the support of our model.

	precision	recall	f1-score	support
0	0.09	0.16	0.12	384
1	0.13	0.16	0.14	553
2	0.24	0.11	0.15	542
3	0.34	0.32	0.33	591
4	0.19	0.21	0.20	539
5	0.37	0.11	0.17	545
6	0.18	0.26	0.21	411
accuracy			0.19	3565
macro avg	0.22	0.19	0.19	3565
weighted avg	0.23	0.19	0.19	3565

Figure 3.19: Precision, Recall, F1-score, Support

3.5 Conclusion

In this chapter, we created an emotion recognition model and used a dataset containing videos expressing different facial emotions, where we explained the steps involved in that.

We would have liked to build our model using a temporal convolutional network, but we did not complete it due to lack of resources and information as this field is still new.

So, we used a simple convolutional neural network in our experiment and our model achieved results that reached 91% on Acted Facial Expressions In The Wild(AFEW) dataset.

CHAPTER 4

CONCLUSION

Our work belongs in the field of affective computing; it deals with the topic of emotion recognition using deep learning approaches. Emotions can be expressed through social behaviors and physiological signals such as: facial expressions, speech, and Electrocardiography (ECG), Electromyography (EMG), Electroencephalography (EEG) signals. Emotion recognition is critical, it becomes an active research area where a great deal of work is ongoing in order to find the best results.

We studied this topic about emotion through facial expressions in the video. We started with a brief reminder about artificial intelligence, machine learning, and deep learning. Then we moved on to defining the meaning of emotions and mentioned the different theories that classify them and the different areas of use of emotions. We touched also on some studies related to emotion recognition where we mentioned some of the first studies on this topic using traditional methods, machine learning methods and deep learning methods.

We proposed a model for recognizing emotions by facial expressions using a convolutional neural network on the Acted Facial Expressions In The Wild dataset and our model achieved a good results.

Due to the lack of capabilities and resources, we could not apply the temporal neural network, as this network has several characteristics over other neural networks. Therefore, it must be used in future work. Some other interesting extension of the present work are:

- Investigate the incorporation of other signals for emotion recognition such as: speech, ECG, etc.
- Recognition of the emotion in video of a group of person
- Exploration of multidimensional TCN models

BIBLIOGRAPHY

- Abdulrahman, M. and Eleyan, A. (2015). Facial expression recognition using support vector machines. In *2015 23rd Signal Processing and Communications Applications Conference (SIU)*, pages 276–279. IEEE.
- Abdulsalam, W. H., Alhamdani, R. S., and Abdullah, M. N. (2019). Facial emotion recognition from videos using deep convolutional neural networks. *International Journal of Machine Learning and Computing*, 9(1):14–19.
- Ahonen, T., Hadid, A., and Pietikäinen, M. (2004). Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer.
- Aleksic, P. S. and Katsaggelos, A. K. (2006). Automatic facial expression recognition using facial animation parameters and multistream hmms. *IEEE Transactions on Information Forensics and Security*, 1(1):3–11.
- Alionte, E. and Lazar, C. (2015). A practical implementation of face detection by using matlab cascade object detector. In *2015 19th international conference on system theory, control and computing (ICSTCC)*, pages 785–790. IEEE.
- Anderson, K. and McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part*

B (Cybernetics), 36(1):96–105.

- Bae, S. H., Choi, I., and Kim, N. S. (2016). Acoustic scene classification using parallel combination of lstm and cnn. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pages 11–15.
- Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271.
- Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., and Wang, H. (2020). The modeling of human facial pain intensity based on temporal convolutional networks trained with video frames in hsv color space. *Applied Soft Computing*, 97:106805.
- Bartlett, M. S., Littlewort, G., Fasel, I., and Movellan, J. R. (2003). Real time face detection and facial expression recognition: development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53. IEEE.
- Boda, R., Priyadarsini, M. J. P., and Pemeena, J. (2016). Face detection and tracking using klt and viola jones. *ARPN journal of Engineering and Applied Sciences*, 11(23):13472–13476.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.
- Cohen, I., Sebe, N., Garg, A., Chen, L. S., and Huang, T. S. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187.
- Cohen, I., Sebe, N., Garg, A., Lew, M. S., and Huang, T. S. (2002). Facial expression recognition from video sequences. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 2, pages 121–124. IEEE.
- Cohn, J. F., Zlochow, A. J., Lien, J. J., and Kanade, T. (1998). Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–

401. IEEE.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

Cristinacce, D., Cootes, T. F., and Scott, I. M. (2004). A multi-stage approach to facial feature detection. In *Bmvc*, volume 1, pages 277–286.

Dang, K. and Sharma, S. (2017). Review and comparison of face detection algorithms. In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, pages 629–633. IEEE.

Darwin, C. (2015). *The expression of the emotions in man and animals*. University of Chicago press.

Deng, L. and Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4):197–387.

Dettmers, T. (2015). Understanding convolution in deep learning. *Retrieved March, 25:2018*.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multim.*, 19(3):34–41.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. (2015). Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474.

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Ekman, P. (1977). Facial action coding system.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion.

Journal of personality and social psychology, 17(2):124.

Eleftheriadis, S. (2016). *Gaussian processes for modeling of facial expressions*. PhD thesis, Imperial College London.

Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570.

Fan, Y., Lu, X., Li, D., and 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*, pages 445–450.

Fehr, B. and Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of experimental psychology: General*, 113(3):464.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Feng, S. (2019). Dynamic facial stress recognition in temporal convolutional network. In *International Conference on Neural Information Processing*, pages 698–706. Springer.

Ghimire, D. and Lee, J. (2013). Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734.

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

Gund, M., Bharadwaj, A. R., and Nwogu, I. (2021). Interpretable emotion classification using temporal convolutional models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6367–6374. IEEE.

- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48.
- Gupta, T. K. and Raza, K. (2019). Optimization of ann architecture: a review on nature-inspired techniques. *Machine learning in bio-signal analysis and diagnostic imaging*, pages 159–182.
- Hasani, B. and Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 30–40.
- Hasnul, M. A., Alelyani, S., Mohana, M., et al. (2021). Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. *Sensors*, 21(15):5015.
- He, Z., Jin, T., Basu, A., Soraghan, J., Di Caterina, G., and Petropoulakis, L. (2019). Human emotion recognition in video using subtraction pre-processing. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pages 374–379.
- Hendy, N. A. and Farag, H. (2013). Emotion recognition using neural network: A comparative study. In *Proceedings of World Academy of Science, Engineering and Technology*, number 75, page 791. World Academy of Science, Engineering and Technology (WASET).
- Hill, D. (2009). *Emotionomics: Winning Hearts and Minds*. Kogan Page Publishers.
- Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.
- Hossain, M. S. and Muhammad, G. (2019). An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework. *IEEE Wireless Communications*,

26(3):62–68.

- Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.
- Hu, Y. and Lu, X. (2018). Learning spatial-temporal features for video copy detection by the combination of cnn and rnn. *Journal of Visual Communication and Image Representation*, 55:21–29.
- Huang, C.-W., Chiang, C.-T., and Li, Q. (2017). A study of deep learning networks on mobile traffic forecasting. In *2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, pages 1–6. IEEE.
- Huang, Y., Chen, F., Lv, S., and Wang, X. (2019). Facial expression recognition: A survey. *Symmetry*, 11(10):1189.
- Hubel, D. H. and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, 28(2):229–289.
- Jain, D. K., Zhang, Z., and Huang, K. (2020). Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognition Letters*, 139:157–165.
- Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. (2011). Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pages 410–415.
- Jones, C. M. and Jonsson, M. (2007). Performance analysis of acoustic emotion recognition for in-car conversational interfaces. In *International Conference on Universal Access in Human-Computer Interaction*, pages 411–420. Springer.
- Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991.
- Kalfaoglu, M. E., Kalkan, S., and Alatan, A. A. (2020). Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*,

pages 731–747. Springer.

Kambi Beli, I. L. and Guo, C. (2017). Enhancing face identification using local binary patterns and k-nearest neighbors. *Journal of Imaging*, 3(3):37.

Khan, R. A. (2013). *Detection of emotions from video in non-controlled environment*. PhD thesis, Université Claude Bernard-Lyon I.

Kim, D. H., Baddar, W. J., Jang, J., and Ro, Y. M. (2017). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236.

Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., and Wrobel, M. R. (2014). Emotion recognition and its applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 51–62. Springer.

Kölbl, L. (2017). *Deep Convolution Neural Networks for Image Analysis*. PhD thesis.

Kotsia, I., Buciu, I., and Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7):1052–1067.

Kundu, P. and Singh, R. K. (2013). Geometric feature based recognition of facial expressions using ann. In *2013 IEEE International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 1–6. IEEE.

Lara-Benítez, P., Carranza-García, M., Luna-Romera, J. M., and Riquelme, J. C. (2020). Temporal convolutional networks applied to energy-related time series forecasting. *applied sciences*, 10(7):2322.

- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Liu, G. (2020). It may be time to improve the neuron of artificial neural network.
- Liu, M., Li, S., Shan, S., Wang, R., and Chen, X. (2014). Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian conference on computer vision*, pages 143–157. Springer.
- Liu, Y., Wang, J.-d., and Li, P. (2011). A feature point tracking method based on the combination of sift algorithm and klt matching algorithm. *Journal of Astronautics*, 32(7):1618–1625.
- Lucas, B. D., Kanade, T., et al. (1981). An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia.
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE.
- Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE.
- Mallick, T., Goyal, P., Das, P. P., and Majumdar, A. K. (2016). Facial emotion recognition from kinect data-an appraisal of kinect face tracking library. In *VISIGRAPP (4: VISAPP)*, pages 525–532.
- Mallya, A. (2017). Introduction to rnns.
- Margineantu, D. D. and Dietterich, T. G. (1997). Pruning adaptive boosting. In *ICML*, volume 97, pages 211–218. Citeseer.
- Mattela, G. and Gupta, S. K. (2018). Facial expression recognition using gabor-mean-dwt feature extraction technique. In *2018 5th International Conference on Signal Processing*

and *Integrated Networks (SPIN)*, pages 575–580. IEEE.

Michel, P. and El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264.

Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., and Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1973–1976.

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. In *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, pages 1–6. IEEE.

Panchal, G. and Pushpalatha, K. (2017). A local binary pattern based facial expression recognition using k-nearest neighbor (knn) search. *Int. J. Eng. Res. Technol.*, 6(5):525–530.

Pantic, M., Pentland, A., Nijholt, A., and Huang, T. S. (2007). Human computing and machine understanding of human behavior: A survey. In *Artificial intelligence for human computing*, pages 47–71. Springer.

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE.

Pardàs, M. and Bonafonte, A. (2002). Facial animation parameters extraction and expression recognition using hidden markov models. *Signal Processing: Image Communication*, 17(9):675–688.

Picard, R. W. (1999). Affective computing for HCI. In Bullinger, H. and Ziegler, J., editors, *Human-Computer Interaction: Ergonomics and User Interfaces, Proceedings of HCI International '99 (the 8th International Conference on Human-Computer Interaction)*, Munich,

Germany, August 22-26, 1999, Volume 1, pages 829–833. Lawrence Erlbaum.

Piórkowska, M. and Wrobel, M. (2017). *Basic Emotions*.

Plutchik, R. (1982). A psychoevolutionary theory of emotions.

Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Saad, M. M., Jamil, N., and Hamzah, R. (2018). Evaluation of support vector machine and decision tree for emotion recognition of malay folklores. *Bulletin of Electrical Engineering and Informatics*, 7(3):479–486.

Saatci, Y. and Town, C. (2006). Cascaded classification of gender and facial expression using active appearance models. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 393–398. IEEE.

Salmam, F. Z., Madani, A., and Kissi, M. (2016). Facial expression recognition using decision trees. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, pages 125–130. IEEE.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Sánchez, A., Ruiz, J. V., Moreno, A. B., Montemayor, A. S., Hernández, J., and Pantrigo, J. J. (2011). Differential optical flow applied to automatic facial expression recognition. *Neurocomputing*, 74(8):1272–1282.

Saud, A. S. and Shakya, S. (2020). Analysis of look back period for stock price prediction with rnn variants: A case study on banking sector of nepse. *Procedia Computer Science*, 167:788–798.

Scherer, K. R. and Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology*, 70(1):719–745. PMID: 30110576.

- Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2):81.
- Sethi, N. and Aggarwal, A. (2011). Robust face detection and tracking using pyramidal lucas kanade tracker algorithm. *International Journal of Computer Technology and Applications*, 2(5):1432–1438.
- Shen, L., Wang, M., and Shen, R. (2009). Affective e-learning: Using “emotional” data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society*, 12(2):176–189.
- Singh, D. (2012). Human emotion recognition system. *International Journal of Image, Graphics and Signal Processing*, 4(8):50.
- Singh, M. and Fang, Y. (2020). Emotion recognition in audio and video using deep neural networks. *arXiv preprint arXiv:2006.08129*.
- Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., and Gulyás, B. (2020). 3d deep learning on medical images: a review. *Sensors*, 20(18):5097.
- Skelley, J. P. (2005). *Experiments in expression recognition*. PhD thesis, Massachusetts Institute of Technology.
- Sloboda, J. A. and Juslin, P. N. (2001). Psychological perspectives on music and emotion. *Music and emotion: Theory and research*, pages 71–104.
- Socher, R., Huval, B., Bath, B., Manning, C. D., and Ng, A. (2012). Convolutional-recursive deep learning for 3d object classification. *Advances in neural information processing systems*, 25:656–664.
- Song, H. A. and Lee, S.-Y. (2013). Hierarchical representation using nmf. In *International conference on neural information processing*, pages 466–473. Springer.
- Sown, M. (1978). A preliminary note on pattern recognition of facial emotional expression. In *The 4th International Joint Conferences on Pattern Recognition, 1978*.

- Spiers, D. L. (2016). Facial emotion detection using deep learning.
- Steffens, J., Elagin, E., and Neven, H. (1998). Personspotter-fast and robust system for human detection, tracking and recognition. In *Proceedings Third IEEE International Conference on automatic face and gesture recognition*, pages 516–521. IEEE.
- Suk, M. and Prabhakaran, B. (2014). Real-time mobile facial expression recognition system—a case study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 132–137.
- Sumpeno, S., Hariadi, M., and Purnomo, M. H. (2011). Facial emotional expressions of life-like character based on text classifier and fuzzy logic. *Scientific Article of IAENG International Journal of Computer Science*, 38: 2, *IJCS_38_2_04*, 2(38):122–133.
- Tao, H. and Huang, T. S. (2002). A piecewise bezier volume deformation model and its applications in facial motion capture. In *Advances In Image Processing And Understanding: A Festschrift for Thomas S Huang*, pages 39–56. World Scientific.
- Terzopoulos, D. and Waters, K. (1990). Analysis of facial images using physical and anatomical models. In *Proceedings Third International Conference on Computer Vision*, pages 727–728. IEEE Computer Society.
- Thorat, B., Manza, G., and Yannawar, P. (2015). Automatic classification of facial expressions from video stream using decision tree. *International Journal of Computer Applications*, 121(22).
- Tian, Y.-I., Kanade, T., and Cohn, J. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115.
- Tian, Y.-L., Kanade, T., and Cohn, J. F. (2005). Facial expression analysis. In *Handbook of face recognition*, pages 247–275. Springer.
- Tie, Y. and Guan, L. (2012). A deformable 3-d facial expression model for dynamic human emotional state recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(1):142–157.

- Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee.
- Wagener, D. W. and Herbst, B. (2002). Face tracking: An implementation of the kanade-lucas-tomasi tracking algorithm. *South Africa*.
- Walczak, S. (2019). Artificial neural networks. In *Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction*, pages 40–53. IGI Global.
- Wang, X., Jiang, W., and Luo, Z. (2016). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2428–2437.
- Whitehill, J. and Omlin, C. W. (2006). Haar features for faces au recognition. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 5–pp. IEEE.
- Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98.
- Wrobel, M. R. (2013). Emotions in the software development process. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 518–523. IEEE.
- Xu, Z., Li, S., and Deng, W. (2015). Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 141–145. IEEE.
- Yacoob, Y. and Davis, L. S. (1994). Labeling of human face components from range data. *CVGIP: Image Understanding*, 60(2):168–178.
- Yacoob, Y. and Davis, L. S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on pattern analysis and machine intelligence*, 18(6):636–642.

- Yang, L. and Liu, J. (2019). Eeg-based emotion recognition using temporal convolutional network. In *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*, pages 437–442. IEEE.
- Yang, P., Liu, Q., and Metaxas, D. N. (2010). Exploring facial expressions with compositional features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2638–2644. IEEE.
- Yao, Q. (2014). Multi-sensory emotion recognition with speech and facial expression.
- Youssef, A. E., Aly, S. F., Ibrahim, A. S., and Abbott, A. L. (2013). Auto-optimized multimodal expression recognition framework using 3d kinect data for asd therapeutic aid. *International Journal of Modeling and Optimization*, 3(2):112.
- Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.
- Yu, J. and Bhanu, B. (2006). Evolutionary feature synthesis for facial expression recognition. *Pattern Recognition Letters*, 27(11):1289–1298.
- Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.
- Zhang, R., Yuan, Z., and Shao, X. (2018). A new combined cnn-rnn model for sector stock price analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 546–551. IEEE.