UNIVERSITY KASDI MERBAH OUARGLA

Faculty of New Technologies of Information and Communication

Department of Electronic and Telecommunication

ACADEMIC MASTER MEMORY

Domain : Science and Technology

Field : Telecommunication

Specialty : Telecommunication System

Topic

Context-based Emotion Recognition using deep learning model

Presented by :

❑ Ms.BOURABAH Nour Eliman

❑ Ms.BEGGARI Khaoula

Publicly defended on :

before the jury :

❑ Dr. LATI Abdelhai	President	UKM OUARGLA
❑ Dr. CHARGUI Abdelhakim	Supervisor	UKM OUARGLA
❑ Dr. BENSID Khaled	Examiner	UKM OUARGLA

❑ **University year : 2023/2024**

ملخص

يتم تطوير التعرف على المشاعر، وهو أمر حيوي في تطبيقات الذكاء الاصطناعي مثل التفاعل بين الإنسان والحاسوب والرعاية الصحية، من خلال هذه الأطروحة باستخدام الشبكات العصبية التلافيفية (CNN) والشبكات المتبقية (ResNet). يُنتج نموذجنا، الذي تم تدريبه على مجموعة بيانات Emotic، مخرجات انفعالية ذات أبعاد مستمرة وفئوية منفصلة، مما يضمن الكشف الدقيق من خلال دمج الإشارات السياقية. تُظهر النتائج التجريبية أداءً متطورًا، مما يحسن بشكل كبير من مهام التنبؤ بالعاطفة ويمهد الطريق لتطبيقات الذكاء الاصطناعي الواعية بالسياق.

.

الكلمات المفتاحية-- التتعرف على المشاعر، الشبكة العصبية التلافيفية ، الشبكة المتبقية ، إيموتيك , السياق

Resume

La reconnaissance des émotions, qui est essentielle dans les applications d’IA telles que l’interaction homme-machine et les soins de santé, est développée dans cette thèse à l’aide de réseaux neuronaux convolutifs (CNN) et de réseaux résiduels (ResResNet). Notre modèle, entraîné sur l’ensemble de données Emotic, produit des sorties émotionnelles avec des dimensions discrètes, continues et catégorielles, assurant une détection précise en incorporant des indices contextuels. Les résultats expérimentaux montrent une amélioration des performances, ce qui améliore considérablement les tâches de prédiction des émotions et ouvre la voie à des applications d’intelligence artificielle tenant compte du contexte.

- Mots-clés: Reconnaissance des émotions, réseau neuronal convolutif , Resnet , Emotic , Contexte.

ACKNOWLEDGMENTS

First of all, we thank **ALLAH**
who helped us and provided us with patience and courage during these years of study.

I would like to thank my distinguished Dr. **ABDELHAKIM CHERGUI** who kindly supervises this theme were dedicated and did not spare us his valuable advice and guidance.

We would like to thank the members of the jury **Dr.BENSID KHALED** and **Dr.LATI ABDELHAI** for agreeing to review and evaluate our modest work. We would also like to express our sincere thanks to **Dr.MOAD MOHAMMED SAYEH** for all advice on these years of study We take this opportunity to express our gratitude to all our teachers who contributed through their collaboration, availability and sympathy, during our teaching.

DEDICATIONS

We dedicate this modest work as a testimony of affectation, of admiration :

First of all, we thank **ALLAH**
who helped us and provided us with patience and courage during these years of study.

To my support and guardian, my father **ESADDIK**

To my pure angel, my mother **AIDA**

To the twin of my soul, my sister **ASSALA**

To my soul mate and the light of my life, **BRAHIM .e**

To my grandparents who are no longer with us, but their spirit and love are with us
Salah, Rebiha, Bouzid, Hadjira

To my youngest and only family **Bourabah and Neili**

I dedicate the fruit of my success and every step I took in the journey of a thousand miles of hard work and fatigue. I end my journey with joy, sincerity, and gratitude for all mistakes and successes, for moments of strength and weakness, and from moments of loneliness to moments of humanity.

Nour Eliman

DEDICATIONS

We dedicate this modest work as a testimony of affectation, of admiration :

To My Familly and friends

whose unwavering support and encouragement have been my greatest strength.

To my mentors, whose wisdom and guidance have shaped my journey, providing me with the tools and confidence to pursue my dreams. To my colleagues, whose collaboration and camaraderie have enriched this endeavor. To everyone who believed in me even when I doubted myself, your faith has been my driving force. This work is a testament to your love, patience, and dedication. Without you, this achievement would not have been possible.

Thank you for being my rock, my inspiration, and my source of endless motivation. Your contributions are woven into the fabric of this accomplishment, and I am forever grateful for each one of you.

Khaoula

Contents

List of tables

List of figures

General Introduction

Introduction

In our daily lives, the ability to perceive and interpret the emotional states of others is fundamental to effective social interaction and communication. This ability, rooted in human nature, allows us to navigate complex social environments, empathize with others, and respond appropriately to a variety of social cues. Efforts to give machines similar emotion recognition capabilities have attracted significant attention in the fields of artificial intelligence (AI) and computer vision. Achieving this goal holds promise for improved human-computer interactions, more intuitive and responsive AI systems, and advances in areas ranging from mental health monitoring to entertainment.

Traditional automatic emotion recognition methods focus primarily on analyzing facial expressions and, to a lesser extent, posture. These methods have demonstrated impressive accuracy in controlled environments where the subject's face is clearly visible and well-lit. Techniques such as convolutional neural networks (CNNs) have proven particularly effective in these environments because they use large datasets of labeled facial expressions to learn discriminative features. However, the performance of these models often degrades in natural, unconstrained environments, where factors such as occlusion, lighting variations, and complex backgrounds pose significant challenges.

Psychological studies underscore that human perception of emotions extends beyond facial expressions and body language. The context in which an interaction occurs—encompassing the surrounding environment, objects, and activities—plays a crucial role in shaping our emotional interpretations. For instance, a smile in a joyful social gathering conveys a different emotional state compared to a smile in a tense, high-stakes meeting. Despite the importance of contextual information, its integration into automated emotion recognition systems remains underexplored, largely due to the lack of comprehensive datasets that capture this multifaceted aspect of human emotion.

To address this gap, we introduce the EMOTIC dataset, a novel collection of images depicting people in various natural environments, annotated with their apparent emotions. The dataset is unique in that it combines two complementary forms of emotion representation: discrete categories (containing 26 different emotions) and continuous dimensions (valence, arousal, and dominance). Combining these two types of annotations allows for a more nuanced understanding of emotional states and provides a rich resource for developing and evaluating emotion recognition algorithms.

In this work, we deeply investigate the statistical properties and annotation consistency of the EMOTIC dataset and conduct a thorough analysis of the reliability of its structure and emotion labels. In addition, we investigate various CNN-based models that exploit contextual information from both the bounding box containing the individual and the surrounding scene. Our experiments show that incorporating scene context can significantly improve the accuracy of emotion recognition systems, highlighting the value of a holistic approach that reflects the multifactorial nature of human emotion perception.
The findings presented in this work lay the foundation for future research aimed at refining and expanding the capabilities of emotion recognition technology. By improving our understanding of how context influences emotion interpretation, we can get closer to developing AI systems that interact with people in a more natural, empathetic, and context-aware way.

State of art

Emotion recognition has been widely studied in the computer vision community. Most existing works focus on predicting emotions using facial expression analysis [1], [2]. The basis of these methods is the facial action coding system [3], which encodes facial expressions using a sequence of specific local facial movements, so-called action units. These face-based methods [1], [2] typically use features based on facial geometry or appearance features to describe the face. The extracted features are then used to identify action units and the basic emotions proposed by Ekman and Friesen [4]: anger, disgust, fear, happiness, sadness, and surprise. Currently, the most advanced facial expression analysis emotion recognition systems use CNNs to identify emotions or action units [5]. Regarding emotion representation, some recent works based on facial expressions [6] used the continuous dimensions of the VAD emotional state model [7]. The VAD model uses three numerical dimensions to describe emotions: valence (V), which measures the positivity or pleasantness of an emotion, ranging from negative to positive; arousal (A), which measures a person's arousal, ranging from inactive/calm to excited/ready to act; and dominance (D), which measures the degree of control a person has over a situation, ranging from submissive/no control to dominant/controlled. On the other hand, Du et al. [8] proposed a set of 21 facial emotion categories, which are defined as different combinations of basic emotions, such as ”surprise” or ”disgust”. This classification allows the authors to provide detailed information about the emotions expressed. Although research on emotion recognition from a computer vision perspective has mainly focused on facial analysis, some work has also considered other additional visual cues or multimodal approaches. For example, in [9], the position of the shoulders is used as additional information to facial features to identify basic emotions. More generally, Schindler et al. [10] used poses to recognize six basic emotions and experimented on a small dataset of non-spontaneous poses captured under controlled conditions. Mou et al. [11] proposed a system for emotion analysis of still images of crowds, which detects arousal and valence at the group level by combining facial, body, and contextual information.

Contribution

The main contribution of this work is to address the limitations of traditional emotion recognition methods that rely primarily on facial expressions. While facial analysis has been shown to be effective in controlled environments, it often fails to produce accurate results in natural, unobstructed environments due to factors such as occlusion, variable lighting, and complex backgrounds. To overcome these challenges, our approach incorporates contextual information from the surrounding scene, enabling a more comprehensive understanding of emotional states. By leveraging the EMOTIC dataset, which contains diverse images annotated with both discrete emotion categories and continuous dimensions of valence, arousal, and dominance, we show that incorporating scene context can significantly improve the accuracy and robustness of emotion recognition models. This novel integration of contextual cues represents a significant advance in the field of automatic emotion recognition.

Thesis organization

Our work is organized as follows :

The first chapter: provides a comprehensive overview of Artificial intelligence and machine laerning and deep learning , focus specifically on the use of deep learning models such as (convolutional neural networks (CNN)) and (residual network(Resnet)) which was utilized in our research..

The second chapter: The focus of this chapter is. overview of biometrics, including theories, definitions, and terminology that are essential for understanding this field. The chapter also covers biometric system design and introduces the topic of Emotion recognintion. Additionally, Learn about common source information in the field of emotion recognition that will be used in our study, the chapter addresses the challenges associated with Context .

The third chapter: The experimental findings are discussed in the context of the emotion recognition verification analysis. We investigated the ntext of the using a novel approach that utilized a CNN and the backbon (ResNet) and discribe Emotic dataset and our architecture of approach. and dedicated to the experimental results and discussion.

Conclusion and perspectives.

Chapter 1

Generalities In The Artificial intelligence And Deep Learning

1.1 Introduction

Artificial Intelligence (AI) is the field of science and engineering concerned with the theory and practice of developing. systems that exhibit the characteristics we associate with intelligence in human behavior, such as perception, natural language processing, problem solving and planning, learning and adaptation, and acting on the environment. Its main scientific goal is to understand the principles that enable intelligent behavior in humans, animals and artificial agents. This scientific goal directly supports several engineering goals, such as the development of intelligent agents, the formalization of knowledge and the mechanization of reasoning in all areas of human endeavor, making working with computers as easy as working with people, and the development of human-machine systems that exploit the complementarity of human and automated reasoning.

Artificial Intelligence is a very broad interdisciplinary field, with roots in and intersections with many domains, not only all computer science disciplines, but also mathematics, linguistics, psychology, neuroscience, mechanical engineering, statistics, economics,control theory and cybernetics, philosophy, and many others. It has adopted many concepts and methods from these fields, but it has also contributed back.

While some of the systems developed, such as an expert system or a planning system, can be characterized as pure applications of AI, most AI systems are developed as components of complex applications to which they add intelligence in various ways, for example by enabling them to reason with knowledge, to process natural language, or to learn and adapt.[12]

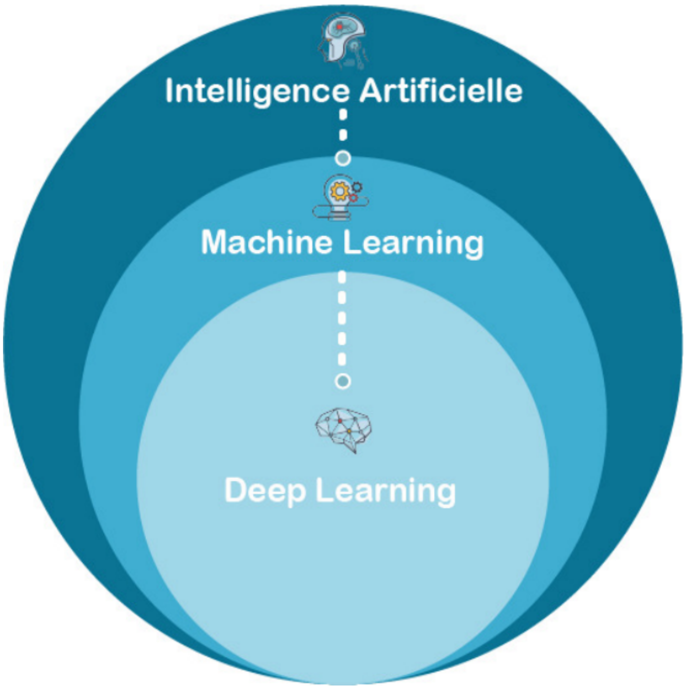


Figure 1.1: Artificial Intelligence

Deep learning, a new field of machine learning research, has been machine learning, has been introduced to bring ML closer to its main goal: artificial intelligence. These are algorithms inspired by the structure and functioning of the brain. They learn multi-level representations to model complex relationships between data.

AI is an umbrella concept that influences and is influenced by many disciplines, such as computer science, engineering, biology, psychology, mathematics, statistics, logic, philosophy, business, and linguistics ([13]; [14]). AI can encompass anything from Apple Siri to Amazon Go, and from self-driving cars to autonomous weapons. Generally, AI can be classified into weak AI and strong AI. Weak AI, also known as narrow AI, excels in specific tasks. Most advancements in AI, that have been achieved to date, can be classified as weak AI, such as Google Assistance and Alpha Go. Researchers from different domains are, however, competing to create a strong AI (also called human-level artificial general intelligence or artificialsuper intelligence), which will process multiple tasks proficiently. A strong AI isthe controversial and contentious concept. Many transhumanists believe that a strong AI can have self-awareness and become the equivalent of human intelligence. Once a strong AI becomes a reality, an intelligence explosion will be precipitated and technologicalsingularity may be unavoidable. Superintelligence may emerge almost immediately after that ([15]). Superintelligence can be loosely defined as “any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest” ([16]). In other words, a strong AI would be able to outperform humans in nearly every cognitive task.

1.2 Application of Artificial Intelligence:

Artificial intelligence, combined with other technologies, has the potential to solve some of the biggest challenges society faces. Artificial intelligence is used in business, manufacturing, medical care, education, military and many other fields. Many innovations have been developed using artificial intelligence-based technologies, Such as facial recognition and self-driving cars. These applications require AI systems to interact with the real world and make automated decisions.

1.2.1 Autonomous vehicles and drones

Self-driving cars are expected to reduce the number of traffic accidents through self-driving technology and death. In addition to land vehicles, autonomous driving technology also makes the development of vehicles possible pilot-less airplane. Drones have become popular among various businesses and government organizations. Especially in areas that are difficult for humans to reach or perform efficiently, such as: B. Scanning Hard to access military bases and fast delivery during peak hours. Goldman Sachs estimates that “global The military will spend 70 billion on drones by 2020, and these drones will play a key role in this effort Resolve future conflicts and replace human pilots ([17]).

1.2.2 Artificial intelligence in education

Artificial intelligence’s natural language processing capabilities will benefit people who cannot read or write People who cannot use computers ([18]).Artificial intelligence will reportedly be used in U.S. education From 2017 to 2021 (March 2018) it will increase by 47.5%. Artificial intelligence can help teachers get rid of duplication They can handle tasks such as grading, allowing teachers to focus more on their professional work ([19]). besides, It’s best to introduce students to this technology as early as possible, as there’s a good chance they’ll use AI future. AI tools can also help make global classrooms accessible to everyone, including students Unable to go to school. Stay current and ensure progress in an ever-changing environment Globally, these applications must also support continuous life-long or never-ending learning ([20]).

1.2.3 Artificial Intelligence in Manufacturing and Factory Automation

Artificial intelligence brings many benefits to the manufacturing industry, such as real-time maintenance Equipment and virtual design.. For example, generative design can be used in manufacturing. Designers enter design goals and the software explores all possible permutations of solutions. Quickly create designs and conduct feasibility testing. possible 50,000 days of work completed in one day ([21]). There is no doubt that artificial intelligence is the key to future production growth.

1.2.4 artificial intelligence in human resources

Artificial intelligence streamlines HR processes in many ways. AI programs compared to humans Thousands of applications can be processed faster, more efficiently and with less mindlessness bias. By identifying the characteristics of successful employees, artificial intelligence can increase a company's chances Hire the most qualified candidates, thereby increasing productivity and retention rates. It also frees human employees from repetitive paperwork and providing answers Frequently Asked Questions. AI can also help improve diversity and inclusion in organizations. one word Caution is advised, though: machine learning often uses large data sets to learn and Everyone has inherent biases.

1.2.5 artificial intelligence in cybersecurity :

Cybersecurity algorithms are critical to combating the tsunami of cyberattacks. Cyber Security Analysis Intelligence can help detect potential attacks before they actually occur. With artificial intelligence and machines Learning to identify and respond to threats can alleviate network workers' fears. In addition, artificial intelligence Can increase efficiency in identifying threats, reduce incident response times, and Real-time reminders of abnormal behavior. Companies have integrated artificial intelligence into every aspect Cybersecurity, including identification, prevention, threat detection and risk analysis of network intruders. This allows human efforts to be diverted to more important activities.

1.2.6 AI in Military

AI benefitsthe military industry on many ways. For instance, combining AI autonomy with computer vision can impact the defense industry by enhancing decision-making and efficiency of soldiers. AI, along with virtual reality, are poised to be a game-changer for military planners, logistics, and military field use. Advances in AI are also enabling significant leaps forward in radio frequency systems. For facial recognition, AI is able to recognize a thermal image captured of a person's face in low-light conditions. Soldiers, who work in covert operations at night, may benefit from this development. AI in the military, however, is a very controversial topic, and many AI researchers and scientists are urging a ban on using AI as a killing machine.

1.2.7 AI at Home

AI is not only changing our workplace, but it is also changing the way we live in our homes. A connected home, combined with AI driven home automation, can take care of almost all daily chores done by humans, such as turning off lights, closing doors, monitoring temperature, playing music, and cleaning floors. Further, AI-powered home automation can reduce the energy consumption by controlling smart thermostats, lighting sensors, and smart plugs. The application of facial recognition algorithms can help detect break-ins and call for emergency services, thereby eliminating the need for human monitoring. Privacy is an issue in this domain.

1.2.8 AI in Health Care

AI-based applications can help improve patients' and elders' health conditions and the quality of their lives. Prime applications of AI in health care include monitoring medicine, treating chronic illness, diagnosing diseases, and surgery support. The National Bureau of Labor Statistics projects that the number of home health aides will grow by 38% over the next 10 years and the length of hospital stays will decrease due personalized rehabilitation and in-home therapy ([22]). From 2012 to 2017, healthcare-AI funding has reached 2.14 billion. AI-based applications in the medical field have achieved many successes, including mining social media to infer health risks, machine learning to predict risky patients, and roboticsto supportsurgery ([23]). In the palliative care field, studies show that only 20 percent of Americans spend their final days at home because of the shortage of palliative care professionals, although 80 percent of Americans prefer to be at home ([24]). AI can also help in predicting and identifying patients who have the most urgent needs for palliative care. From symptom diagnoses to clinical decision support, algorithms that leverage AI have made headlines in terms of accuracy and speed.

1.2.9 AI in Finance

AI in healthcare has a direct impact on finance too. According to Accenture analysis, the AI health market will grow by 6.6 billion by 2021, which can potentially create 150 billion in annual savings for the U.S. healthcare economy by 2026. Unlike traditional finance systemsthat are based on manually set rules and analytics, the AI-based system is shown to be more effective in detecting financial malfeasance with respect to accuracy and completeness. AI can help fight financial crimes, such as credit card fraud, anti-money laundering, and synthetic identity theft ([25]). Financial Technology (FinTech) is emerging as a transformative and strategic initiative for the financial industry ([26])

1.3 Machine Learning :

Arthur Samuel (2000) coined “machine learning” in 1959 and defined it as a field ofstudy that enables computers to learn without being explicitly programmed. Although this definition is rather vague, it indicates a significant feature of machine learning – it does not follow pre-programmed “rules”. In general, machine learning is an automated process that enables machines to analyze a huge data set, recognize patterns, and learn from the data to provide support for predictions and decision making. “Reinforcement machine learning”, like the one used in AlphaGo Zero, starts from scratch with only the rules of the game, Go. It learned by adjusting actions, based on continuous feedback, and AlphaGo Zero achieved unbelievable results (i.e. probably the best Go “player” in the world at the moment) in a very short period of time (i.e. a few days).

It discovered Go moves that were not known in the game’s 3,000-4,000 years history. Reinforcement machine learning can also bypass human biases that may reside in data when using big data for training. Machine learning can be regarded as the automation of cognitive functions ([27]), or the automation of knowledge work (Chui et al., 2016).

AlphaGo and self-driving cars are products of machine learning, especially reinforcement learning. The drawback in machine learning, at the moment, is that the inner working of these self-learning machines is a black box, which makes it difficult to understand and explain the reasoning process and to justify the recommendations. As a result, trusting these machines is a challenge. Further, machine learning using big data can be susceptible to human biases that are inherent in the data ([28]). People tend to only trust a system when the reasoning process is known and interpretable. there are three types of machine learning :

1.3.1 Supervised learning

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance. Under the umbrella of supervised learning fall: Classification, Regression and Forecasting. Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.

- Regression
- Forecasting:

1.3.2 Unsupervised learning

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analysing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organise that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organised. As it assesses more data, its ability to make decisions on that data gradually improves and becomes more refined. Under the umbrella of unsupervised learning, fall:

- Clustering
- Dimension reduction

1.3.3 Reinforcement learning

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result. [29] [30]

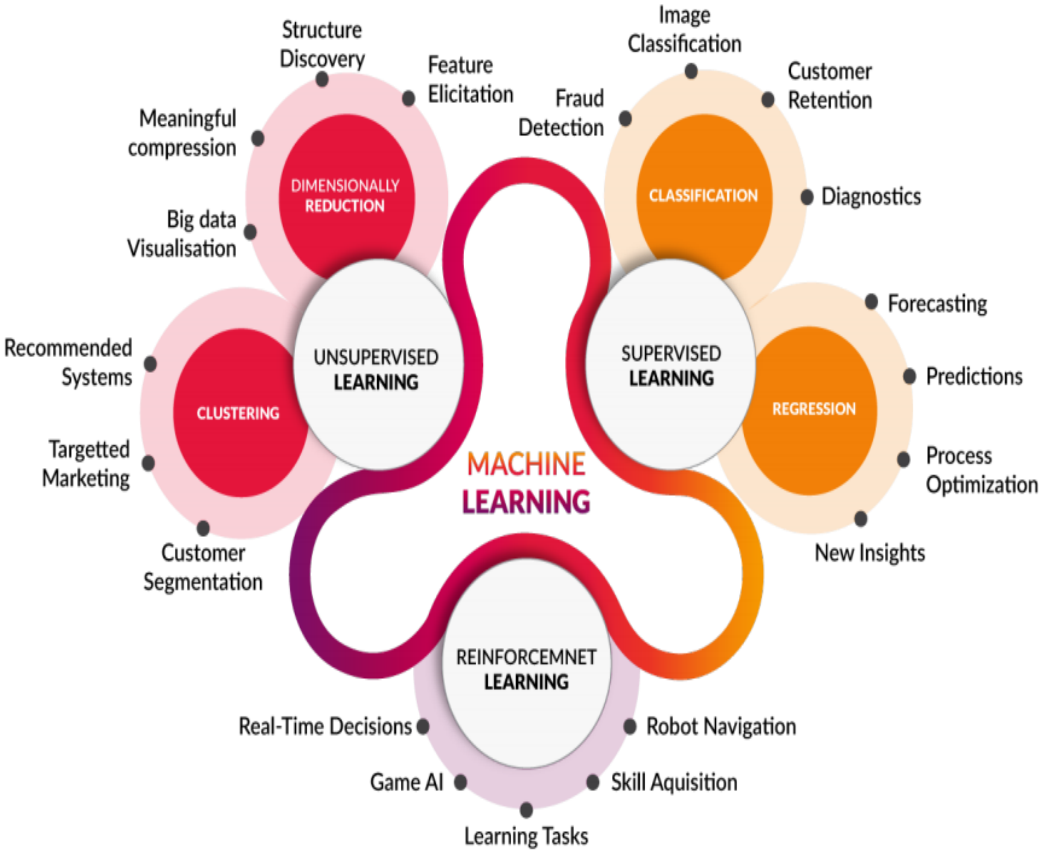


Figure 1.2: Machin Learning types

1.4 History of Deep Learning :

The idea of Deep Learning is not a recent one, but actually dates back to the 1980s, particularly following the work on multi-layer neural networks and the work of pioneers of Machine Learning and Deep Learning, such as the French scientist Yann le Cun.

In collaboration with two other computer scientists, Kunihiko Fukushima and Geoffrey Hinton, they developed a particular type of algorithm called a convolutional neural network.

Although this approach produced results, its progress and evolution were limited limited by technological advances in microprocessors, computational power and lack of and the lack of access to data for training neurons.

However, some researchers have continued to work on this model for around two and, with the help of technological developments and, above all, the ever-increasing availability of data, have been able to improve this technique.

In order to develop an effective learning system, you need to be able to exercise it. This requires a large amount of data to be tested. It was against this backdrop that, in 2007, the STANFORD VISION LAB, with Fei-Fei Li at its head, developed an image aggregator where of millions of photos:

In 2010, Image Net brings together 15,000,000 images, all categorized according to their specific characteristics (vehicles, animals, etc.)

In 2012, Deep Learning was back in the spotlight with a resounding success success at the Image Net Large Scale Visual Recognition Challenge (ILSVRC), an annual image recognition competition founded by Stanford University's STANFORD VISION LAB Several teams of computer science researchers compete against each other in this annual contest to award victory to the program with the program with the lowest failure rate.

And while deep learning algorithms are absent from the competition, in 2012 it was indeed a Deep Learning algorithm that surprisingly won the 2012 edition. [31]

1.5 Deep Learning :

Deep learning is a sub-field of artificial intelligence (AI). The term is used to describe all machine learning techniques (machine learning, i.e. a form of learning based on mathematical approaches used to model data. To better understand these techniques, we need to go back to the origins of artificial intelligence.

in 1950, when Alan Turning became interested in machines capable of thinking. The result was machine learning, a machine that communicates communicates and behaves according to stored information.

Deep learning is an advanced system based on the human brain, featuring a vast network of artificial neurons. These neurons are interconnected to process and store information, compare any given problem or situation with similar ones in the past and solve the problem in the best possible way.

As with human beings, deep learning involves learning from lived experience or, in the case of machines, from recorded information.

Deep learning doesn't mean anything A deeper understanding is gained through this approach; instead, it represents the idea of a continuous hierarchy of representations. How many layers contribute to the data model is called The depth of the model. Other suitable names can be layered for this field Learning representations and learning hierarchical representations. Modern deep learning often include dozens or even hundreds of consecutive presentation layers - and they are Everything is learned automatically by accessing training data. Now there are other ways Machine learning often only focuses on learning one or two layers of representation data; therefore, they are sometimes called shallow learning. In deep learning, these hierarchical representations are (almost always) learned by a model The so-called neural network has a structure similar to the human brain and is composed of artificial neurons. Also called a knot. These nodes are stacked next to each other in three layers:

1. The input layer.
2. The hidden layers.
3. The output layer.

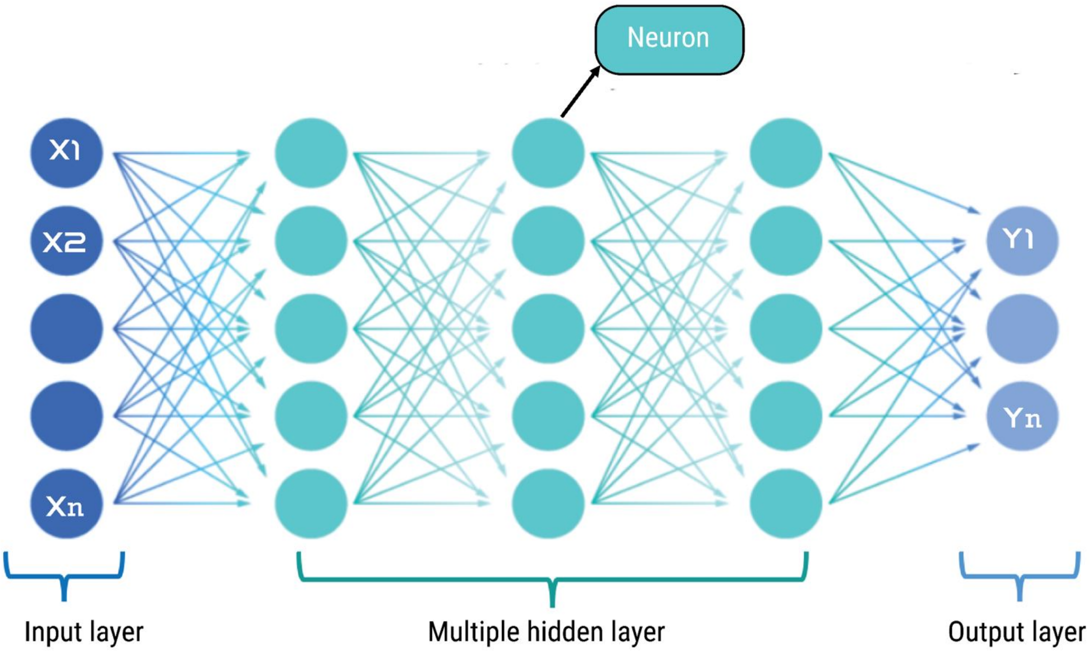


Figure 1.3: Deep Learning

Data provides each node with information in the form of inputs. The node multiplies the inputs with random weights, calculates them, and adds a bias.

Finally, nonlinear functions, also known as activation functions, are applied to determine which neuron to fire. As shown in Figure bellow.

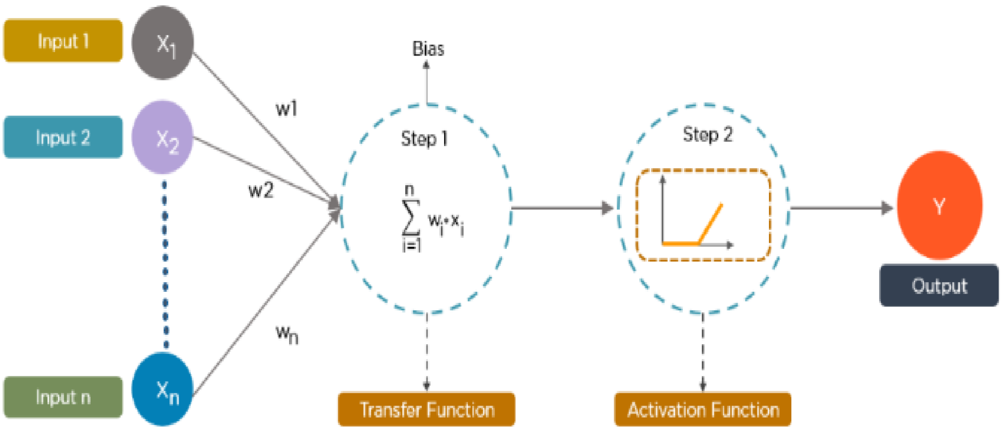


Figure 1.4: Architecture of Neural Network

1.6 The difference between Deep Learning (DL) and Machine Learning ML :

Deep Learning is a specialized branch of Machine Learning. In traditional Machine Learning, the process begins with the manual extraction of relevant features from images. Based on these features, a model is created to categorize the objects in the image. This is illustrated on the left side of Figure 1.5, where features are manually extracted and then used in a classification model to identify objects such as cars, trucks, and bicycles.

In contrast, Deep Learning automates the extraction of relevant features from images, as shown on the right side of Figure 1.5. Deep Learning performs end-to-end learning from raw data, meaning that the network is assigned tasks, such as classification, and learns how to perform them automatically. Another major difference is that Deep Learning algorithms evolve with the data, continually improving as more data is fed into the network. This is unlike Shallow Learning, which refers to Machine Learning methods that converge and stop progressing after reaching a certain level of performance, even with additional training data.

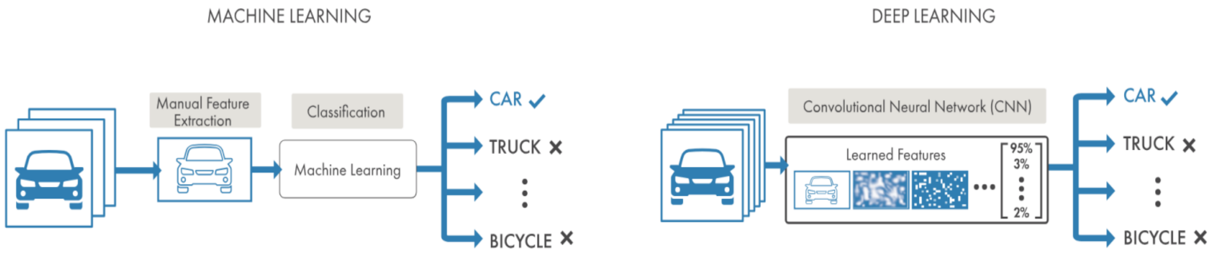


Figure 1.5: The difference between Deep Learning (DL) and Machine Learning ML

To classify images using Machine Learning, the choice of features and classifier must be made manually. With Deep Learning, the feature extraction and modeling processes are automated, enhancing the network's ability to improve as the data volume increases.[32]

1.7 convolutional neural network

Convolutional Neural Networks, also known as CNNs, are most commonly a subtype of ANN Used to analyze visual images [33]. The CNN model consists of a limited set of Processing layers that can learn various features of the input data, such as: B. Images with multiple level of abstraction. CNN means that the network uses mathematical operations So-called convolution [34]. The operation involves convolving the input signal f with a filter g produces an output signal. Convolution is a linear operation where filters and The input signal is passed through a dot product operation to create a scalar for each position The filter is located in the signal. The filter is moved to each signal point and during the run The scalar product then produces a scalar. Convolution creates a complete signal (=image) as the result.

in Figure 1.6 we present Schematic diagram of a basic CNN architecture [35] Schematic diagram of a basic convolutional neural network (CNN) architecture.

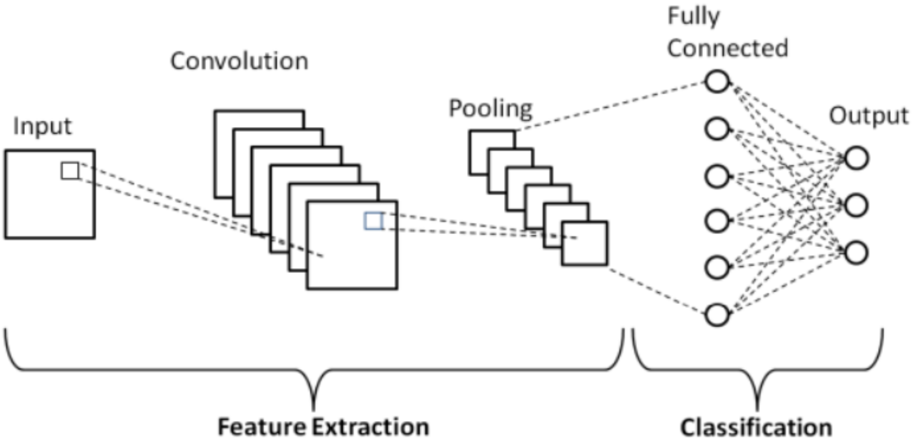


Figure 1.6: The difference between Deep Learning (DL) and Machine Learning ML

1.7.1 Basic CNN components

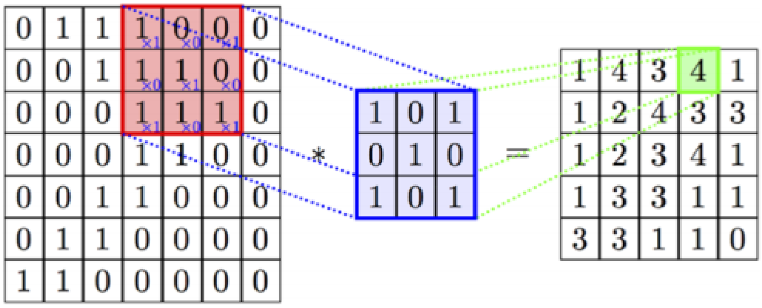
A Convolutional Neural Network (CNN) is composed of several key components, each playing a specific role in the network's ability to process and understand visual data. The basic components of a CNN include:

The Convolution Layer

It acts as the first layer in the process of extracting various features from the source image. In this layer, a mathematical convolution operation is performed between the input image and a filter of a certain size mxn. The dot product between the filter and the portion of the input image proportional to the filter size mxn is calculated by dragging the filter over the input image.

The output is called a feature map and contains detailed information about the image, including its corners and edges.

This feature map is then passed to other layers to teach them other features in the input image.



$$\begin{bmatrix} 0&1&1&1&0&0&0\\ 0&0&1&1&1&0&0\\ 0&0&0&1&1&1&0\\ 0&0&0&1&1&0&0\\ 0&0&1&1&0&0&0\\ 0&1&1&0&0&0&0\\ 1&1&0&0&0&0&0 \end{bmatrix} * \begin{bmatrix} 1&0&1\\ 0&1&0\\ 1&0&1 \end{bmatrix} = \begin{bmatrix} 1&4&3&4&1\\ 1&2&4&3&3\\ 1&2&3&4&1\\ 1&3&3&1&1\\ 3&3&1&1&0 \end{bmatrix}$$

Figure 1.7: 2-D convolution masking.

There are three hyperparameters deciding the spatial of the output feature map:

- **Stride (S) :** is the step we take each time the filter is slid. We move the filters one pixel at a time when the stride is 1. When the stride is two (or, less frequently, three or more; however, this is uncommon in practice), the filters jump two pixels at a time as we move them. This will result in spatially smaller output volumes.
- **Padding (P):** The inputs will be surrounded by a border of the specified size. Zeropadding is most frequently used to pad these areas. The size of this zero-padding is a hyperparameter in neural network frameworks (such as Caffe, TensorFlow, PyTorch, and MXNet). The spatial size of the output volumes can also be controlled by the size of the zero-padding.
- **Depth (D):** The depth of the output volume is a hyperparameter too; this corresponds to the number of filters that are used in a convolution layer.Given w as the width of input, and F as the width of the filter, with P and S as padding, the output width will be: (W + 2P - F) / (S + 1). Typically, set P=(F - 1) / 2 when the stride is S = 1, the spatial size of the input and output volumes will be the same.

Pooling

Pooling layer is usually placed after convolutional layer. The main goal of this layer is to reduce the size of convolutional feature map to save computation. This is done on each feature map independently and by reducing the connections between them in number of layers. There are different types of pooling operations depending on the method used. The largest component in Max Pooling is taken from the feature map. The average value of the components in a part of the image with predefined size is determined by Average Pooling. Sum Pooling calculates the sum of items in a predefined part. It works normally Pooling layer as a connection between FC layer and Convolutional layer.

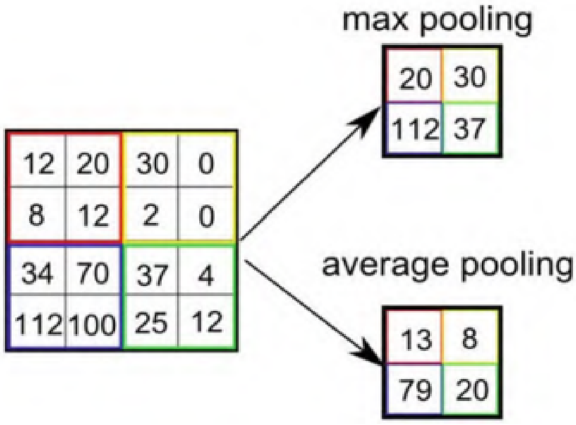


Figure 1.8: Exemple of pooling in Deep learning

- **Max Pooling:** Reduces the spatial dimensions of the data by taking the maximum value in each patch of the feature map.
- **Average Pooling:** Reduces the spatial dimensions by taking the average value in each patch.

Fully-Connected

Layers called “Fully Connected” (FC) include biases, weights, and neurons required to connect neurons between different layers. Usually these are the layers that come before the output layer and form the last few layers of a CNN architecture. The input images from the previous layers are flattened and passed to the FC layers. The flattened vectors then pass through further FC layers where mathematical functions are usually used and operations take place. This marks the beginning of the classification process.

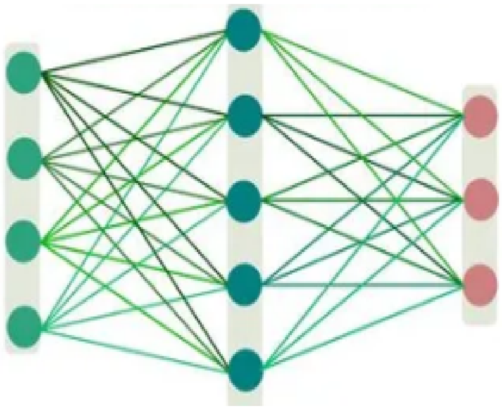


Figure 1.9: Fully-Connected Neural Network

Activation Function

Responsible for determining what information should and should not be activated in the forward pass. Its inclusion also introduces nonlinearity to the network. We use ReLU. ReLU (Rectified Linear Unit) is a commonly used activation function in neural networks that returns the input if it is positive and zero otherwise. In other words, ReLU ”adjusts” negative inputs to zero and leaves positive inputs unchanged according to the following formula:

$$g(x) = \max 0, x \tag{1.1}$$

Batch normalization

Batch normalization is employed to address the issues associated with the internal covariance shift within feature maps. The internal covariance shift is a alteration in the distribution of the hidden unit values, which slows the convergence (by requiring a small learning rate) and requires initializations with care. Batch normalization for a transformed feature map F_i^k is described in this equation :

$$N_l^k = \frac{F_l^k - \mu_B}{\sigma_B^2 + \epsilon} \tag{1.2}$$

N_i^k is the normalized feature-map, F_i^k is the input feature-map, µB and B represent the mean and variance of a feature-map for a mini batch. To prevent division by zero, ϵ is added for numerical accuracy. Batch normalization unifies the distribution of feature map values by setting their mean to zero and variance to one. Additionally, it blurs the gradient's flow and functions as a guiding factor, which, as a result, helps to enhance the network's generalization. [36]

Dropout

Dropout promotes regularization within the network, ultimately improving generalization by skipping certain units or connections with a certain probability. In neural networks, multiple connections that learn nonlinear relationships sometimes coordinate with each other, leading to overfitting. This sporadic omission of connections or entities leads to multiple refined network architectures, and ultimately a representative network with low weights is selected. This random dropping of some connections or units produces several thinned network architectures, and fnally, one representative network is selected with small weights. This selected architecture is then considered as an approximation of all proposed networks.[36] [37]

1.8 Residual network (ResNet)

Residual Neural Networks (ResNet) have become the basis for a variety of cutting-edge deep learning applications. Due to their powerful feature representation capabilities, they are an important feature extractor in many modern networks. As integrating deeper networks into deep learning architectures became popular, models such as AlexNet [38] and VGG Network [39] gained traction. However, increasing the number of layers leads to a common problem known as vanishing/exploding gradients. This problem occurs when the gradients are backpropagated through multiple layers. Due to repeated multiplications, the gradients can become very small. The authors of ResNet [40] found a solution to this problem by implementing so-called "skip connections", which connect the activations of one layer to the next by skipping one or more layers in between (see figure). The residual mapping approach made it possible to train networks with more than a hundred layers.

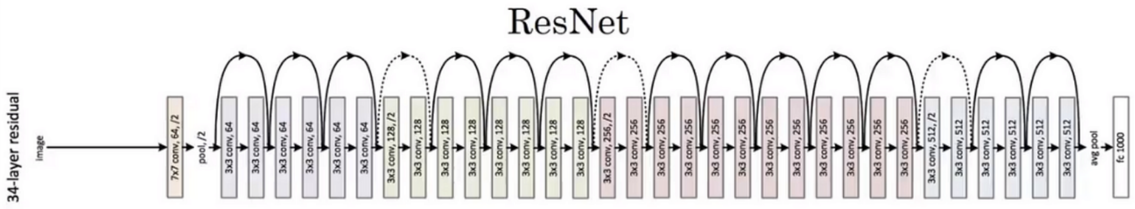


Figure 1.10: ResNet Architecture

1.9 The goal of deep learning :

The ultimate goal of deep learning is to teach computers to recognize an ideal model if given a set of unstructured data. The real world is an example of unordered data, with many of the things that surround us like the sky, trees and people, computers can recognize without human help.

Deep learning has essentially the same aims as machine learning, which is becoming. which is becoming increasingly popular in new technologies. But machine learning is limited in terms of the data it can absorb.

recognize many things in images, but unfortunately it can't adapt to a 3D world, which is not in the interest of building autonomous cars. As for deep learning, it offers an unlimited virtual world for learning, which means that we could theoretically one day surpass the capacities of the human brain. This is due to the set of algorithms used by deep learning called neural networks.

1.10 conclusion :

In summary, this chapter provided a comprehensive overview of artificial intelligence (AI), from its basic principles to its diverse and sophisticated applications. We explored AI's impact across various fields, including autonomous vehicles, education, manufacturing, human resources, cybersecurity, military, home environments, healthcare, and finance. Foundational concepts in machine learning, such as supervised, unsupervised, and reinforcement learning, were discussed alongside the evolution and principles of deep learning. By examining convolutional neural networks (CNNs) and residual networks (ResNets), we highlighted significant advancements in AI capabilities. Overall, this chapter underscores the transformative potential of AI and deep learning in revolutionizing numerous sectors and solving complex problems, setting the stage for future exploration and development in the field.

Chapter 2

Biometrics and emotion recognition

2.1 Introduction

We usually recognize people by their faces, but sometimes also by their voices or handwriting, or by the way they move.

In the past, eye contact was the only way to verify the identity of a traveler traveling from one country to another, a tourist seeking access to a private area, or a shopkeeper withdrawing money from a bank. This approach is no longer realistic given the increase in international travel, the need to ensure workplace security, the increase in electronic banking and the many other changes that now impact our daily lives. Today, a new type of authentication has emerged that uses automated methods and information and communication technologies.

(ICT) Identifying someone based on physical characteristics or behavior – biometrics. Biometrics is a set of technologies used to identify individuals based on physical or behavioral characteristics.

These features are processed through some automated process using devices such as scan engines or cameras.

Unlike information you know or have, biometrics are based on your identity and prevent copying, theft, forgetfulness or loss. In this chapter, we will focus on some basic concepts and definitions related to biometrics and its various technologies. [41] We know that emotions have a significant impact on human life. At different moments, a person's facial expressions reveal his or her feelings or mood. People produce thousands of facial expressions when communicating, each varying in complexity, intensity, and meaning. Emotions or intentions are often conveyed through subtle changes in certain facial features. The presence or absence of certain facial movements can alter the interpretation of facial expressions. Additionally, some facial expressions may look similar but have different meanings depending on their intensity. Therefore, mastering the nuances of facial expressions is crucial to understanding nonverbal communication [13]. In our daily lives and social interactions, we often try to understand the emotional state of others. A lot of research has been done to equip machines with the ability to recognize emotions. From a computer vision perspective posture. While some of these methods work very well in controlled environments, they have limited effectiveness in natural, unrestricted settings. Psychological research shows that scene context, as well as facial expressions and gestures, play a crucial role in emotion perception. However, contextual integration for automatic emotion recognition has not been thoroughly studied, partly due to a lack of sufficient data.

2.2 Difinition of boimitry :

The term refers to the technological sector dedicated to the identification of individuals by their biological characteristics. Automated recognition technologies evaluate the physical and behavioural characteristics of individuals. Common physical takes into account fingerprints, hand geometry, retinal and iris characteristics. characteristics of the retina, iris or face.

2.3 Architecture of biomitric System :

There are always at least two modules in a biometric system: the learning module and the recognition module. The third (optional) module is the adaptation module. adaptation module. During learning, the system acquires one or more biometric measurements which will be used to build a model of the individual. This reference model will serve as a point of comparison during recognition. The model can be re-evaluated after each use, thanks to the adaptation module. [42]

2.3.1 Learning:

In the learning process, biometrics are first measured using sensors, This is known as acquisition or capture. Typically, this capture is not directly stored, but transformed. This is because the signal contains information that is not required.
only the relevant parameters are extracted. The model is a compact representation of the signal, which helps in the identification phase, but also reduces the amount of data to be but also reduces the amount of data to be stored. It's important to note that the quality of the sensor.
can greatly affect the performance of the system. The better the quality of the system, the less pre-processing is required to extract the signal parameters. signal parameters. [42]
However, quality sensors are generally expensive and their use is therefore limited to

high-security applications for a restricted audience.
limited to high-security applications for a restricted audience. The model can be stored in a database, as shown in figure 1, or on a smart card.

2.3.2 Recognition:

During recognition, the biometric feature is measured and a set of parameters. set of parameters is extracted, as during training. The sensor used must have properties as close as possible to those of the sensor used during the learning phase. If the properties of the two sensors are too different, a series of a series of additional pre-processing steps to limit performance degradation. performance degradation. Further recognition will differ according to depending on the system's operating mode: identification or verification. In identification mode, the system must guess the person's identity. It therefore answers a question: "Who am I? In this mode, the system compares the signal signal with the various models contained in the database (problem of type 1 : n). In general, when we talk about identification, we assume that the problem is closed, i.e. everyone who uses the system has a model in the database. model in the database. In verification mode, the system must answer a question such as "Am I who I say I am?
The user proposes an identity to the system, and the system must verify that the identity is indeed the one proposed. All that needs to be done is to compare the signal with of the models in the database (1:1 problem).
In verification mode, we speak of an open problem, since we assume that an individual who has no model in the database (impostor) may seek to be recognized.
Identification and verification are therefore two different problems. Identification can be a daunting task when the database contains thousands or even millions of identities.
identities, especially when there are "real-time" constraints on the system. constraints on the system. These difficulties are analogous to those experienced, for example, by systems for multimedia document indexing systems. [42]

2.4 Caracteristic of biometrics:

A biometric characteristic (or trait) is a measurable physical or behavioral behavioral characteristics of an individual that is distinguishable. It determines how an individual. individual will be recognized. Each biometric modality has its own strengths and weaknesses.

depends on the field of application and sometimes on the population to be be identified. Ross et al have identified certain requirements that a typical biometric system system must meet:

- **Universality:** Any person accessing the system should possess the following For example, we cannot use the iris as a feature to identify blind people. to identify blind people. Uniqueness: To differentiate one individual from another.
- **Permanence (stability):** Biometric characteristics should be resistant to change over time.
- **Measurability:** Biometric characteristics should be quantitatively measurable so that they can be used to compare two individuals.
- **Performance:** A practical biometric system must have acceptable accuracy and a recognition speed that is reasonable in relation to the resources required.
- **Acceptability:** To what extent are the people who are intended to use the application willing to their biometric features to the system.
- **Circumvention:** Reflects how easy it is to deceive the system by fraudulent methods. [43]

2.5 Biometric Modalities

The use of biometric for identification and authentication is a highly efficient technique, as well as a fundamental concept linked to the recognition of individuals biometric characteristics. There are several biometric techniques used in various applications, falling into three categories:

2.5.1 Physical biometrics

This is based on particular physical characteristics which, for everyone, are unique and permanent. Examples of these traits are presented in the following sections :

- **Fingerprinting:** fingerprint recognition is one of the first biometric first biometric techniques based on the fact that every person has unique fingerprints [44]. However, fingerprints are biometric measurements that are poorly accepted by users, due to their frequent association with criminology.

  The term fingerprint refers to the mark left by the grooves and ridges of finger pulps ridges on a sufficiently smooth surface due to the grease and other impurities that we carry on our hands. The fingerprint is definitively traced from the 6th month of pregnancy. It is immutable and unalterable. It is reconstituted identically even in the event of injury or burns. The pattern formed by fingerprints is unique to each of us, which explains why they are used to identification. It's even unique to each finger. Because the process that leads to the formation of what are also known as dermatoglyphs. It depends on the expression of genes, but also on external factors such as the speed growth rate of the fingers or the diet of the foetus. As a result, even the fingerprints of identical twins will be slightly different.

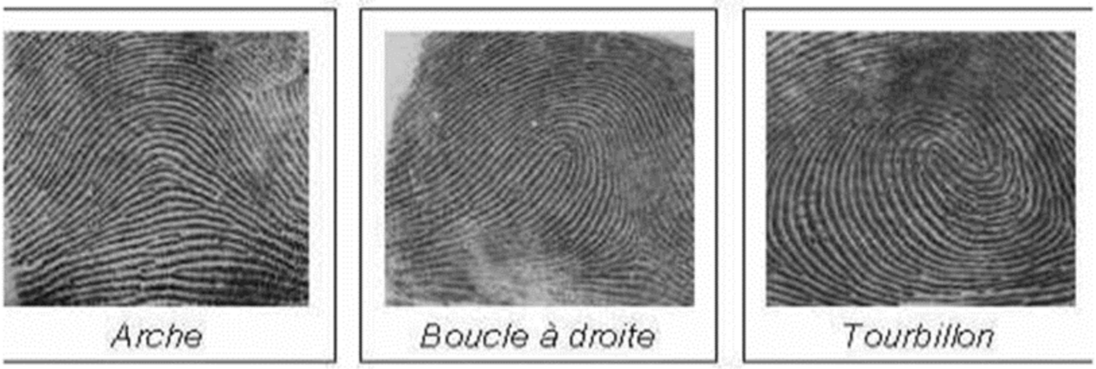


Figure 2.1: Fingerprint shape

- **Hand geometry:** Hand geometry involves analyzing and measuring the shape of the hand. measuring the length, width and height of a user's hand, and creating a and create a 3D image. Infrared LEDs and a digital camera are used to acquire are used to acquire the hand data. [45], this technique offers a reasonable level of precision and is relatively easy to use [46]. fooled by twins or people with similar hand shapes. The most common uses of hand geometry include presence registration and access control.

and access control. On the other hand, hand geometry capture systems are geometry capture systems are relatively bulky and heavy, which limits their use in other applications such as authentication in embedded systems: mobile phones cars, laptops, etc.

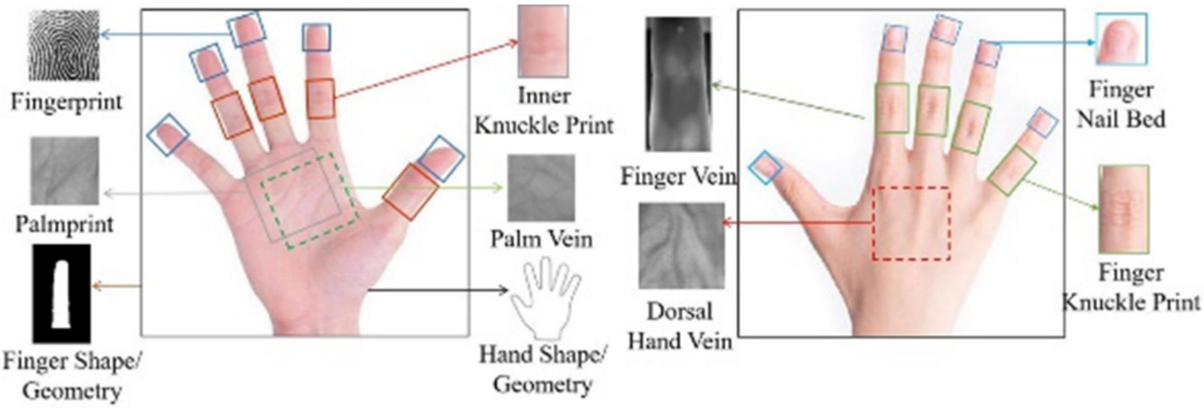


Figure 2.2: The geometry of the hand

- **The face:** Facial authentication is the most natural method of associating humans with visual interaction. [47] Recently, different types of cameras and stills cameras of varying quality and cost have appeared on the market, making it possible to adapt image quality to the conditions of use. Today, however, it is only when the subject is stationary and environmental conditions are normal (uniform background, sufficient lighting) that computer systems give good results. If acquisition takes place in a natural environment, without imposed constraints, performance deteriorates considerably, as personal variations (glasses, hats, moustaches) or environmental variations (lighting, distance) are still poorly taken into account by computer systems. Dans ce cas, une série de prétraitements sont souvent nécessaires avant de procéder à la reconnaissance. [48]

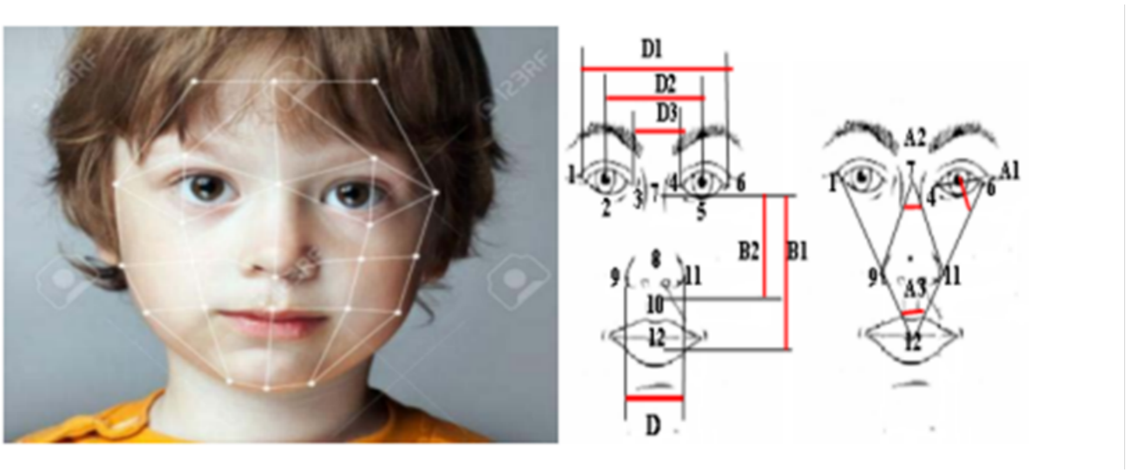


Figure 2.3: Enter Caption

- **The iris:** The iris is the annular region between the pupil and the white of the eye [49]. Iris recognition is a reliable method, as it is always different (even between twins, between left and right eyes) and remains stable throughout life. However, users must place their eyes very close to the acquisition device. [50] Its only drawback is its relatively high cost, which is not as suitable for everyday applications. As a result, its use has been limited to places where security is paramount and even critical, such as nuclear bases. Iris recognition is also used in the financial sector for employees and customers, in hospitals and in major airports. [51] A person wishing to identify himself places his eye a few centimetres from the sensor, and the iris image is taken by a camera. Features are then extracted from the iris image and compared with those stored in the database.

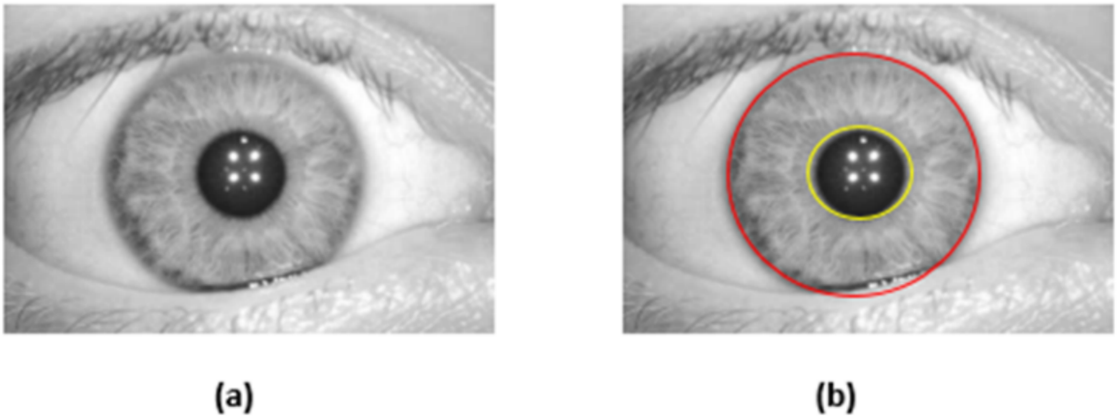


Figure 2.4: Detail of the iris (b) location of the iris

- **The retina:** The retina is a thin layer of cells at the back of the vertebrate eyeball.

It is the part of the eye that converts light into nerve signals.

The principle of retinal biometry is to capture and analyze blood vessel patterns on the thin nerve at the back of the eyeball, which processes light entering the pupil. Retinal biometry also provides a high level of recognition. This technique is used for very high security applications: for example, in military or nuclear applications [52], the characteristics of the retina are linked to the geometric configuration of the blood vessels.

The technique uses specialized equipment and a beam to illuminate the back of the eye.

The systems identify up to one hundred and ninety-two reference points. Certain health risks have been revealed and limit the use of this technique to very sensitive premises. [53]

Retinal acquisition systems are expensive. The eye must be located very close to the reading head, and the user must look at a specific point for several seconds. This method is therefore poorly accepted by the general public.

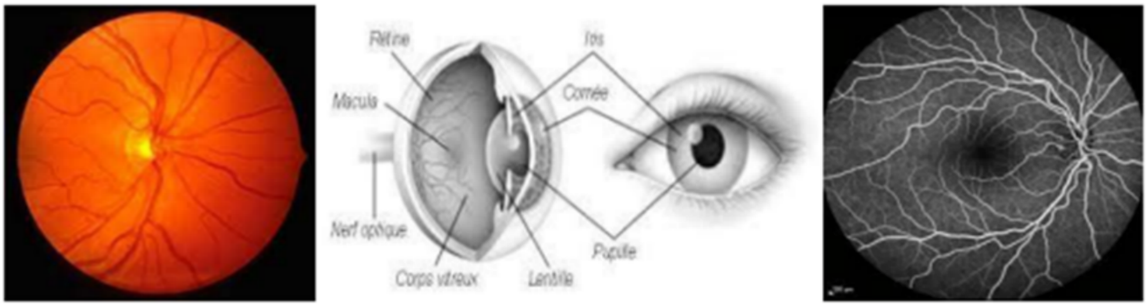


Figure 2.5: Recognising the retina

2.5.2 Behavioral biometrics

It is based on the analysis of certain behaviors of a person, such as the trace of his signature, the imprint of his voice, his gait and the way he types on a keyboard.

- **Voice:** This modality allows us to analyse and recognise the human voice. It is recorded by a microphone and transcribed into machine-readable text. Several characteristics resulting from this modality (such as tone, frequency, harmonics, speed and rhythm, etc.) are then extracted and compared with those already stored in a database in order to confirm or deny the identity of a speaker. [54]

A telephone or microphone can be used as an acquisition device, making this technique relatively cheap and easy to implement. However, it is sensitive to voice changes due to age, cold, etc. In addition, the speech recognition system may fail in noisy environments (low signal-to-noise ratio). It can also be sensitive to attacks (e.g. feedback of the voice signal to the system). [55]
This modality is used in many applications, including computer voice dictation applications, law enforcement by police, spy agencies and telephony. But this modality finds limits because of the big difference between formal language, which includes and uses machines, and the natural language that humans use. The challenge is to find a compromise between these two languages. [56]

Figure 2.6: Voice recognition image

- **Signature:** Signature-based identification is an automatic method for measuring people's signatures. This technique is considered one of the first used in the field of biometrics. It is generally based on the fact that the user signs with an electronic pen on a graphic palette and at the same time it examines the set of dynamics such as the speed, direction, and pressure of the writing, the time during which the pen is in contact with the paper, the time taken to make the signature and the positions where the pen is raised and lowered on the paper. [57] Signing is a behavioral biometric that changes over a period of time and is influenced by the physical and emotional conditions of the signers. Some people's signatures vary considerably.

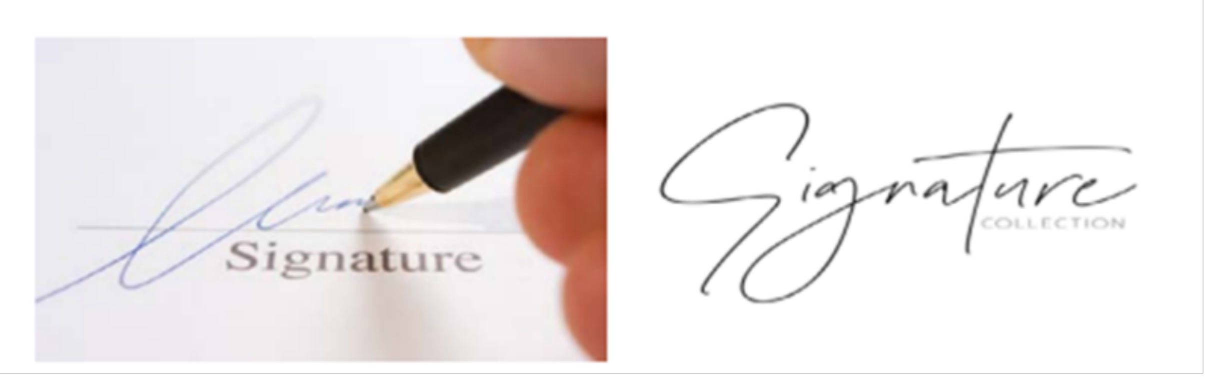


Figure 2.7: Image signature

- **Gait:** The way a person walks can distinguish them from others. In a recognition system using this modality, we seek to identify an individual by the way they walk and move while analyzing video images of the candidate's walk. [58]

  It is therefore a remote mode of identification. People show different traits as they walk, such as body posture, distance between the two feet, position of joints like knees and ankles, and pivoting angles [59], which contribute greatly to their identification.

  This mode is particularly well suited to video surveillance applications. The performance of gait-based systems is not sufficiently acceptable, as it is affected by a number of factors, including choice of footwear, nature of clothing, leg reach, walking surface and so on.s of the signers. Some people's signatures vary considerably.

Figure 2.8: Images of the walking

- **Keyboard typing dynamics:** Keyboard typing dynamics is a characteristic of the individual, and is the transposition of graphology to electronic means. It's a method based on the way you use or type on a keyboard. The parameters taken into account are generally the duration between keystrokes, the frequency of errors and the duration of the keystroke. It also depends on the person's physical and psychological state (age, illness, etc.). This will vary the quality of the f rappe. [60]

Figure 2.9: Enter Caption

2.5.3 Biological biometry

These biometric methods are based on :

- **DNA:** is a biological fluid which is analyzed by simple methods such as blood group, protein or enzyme analysis. Most of these analyses were quickly abandoned in favor of DNA profiling, also known as DNA fingerprinting, given its high discrimination and robustness. The most common DNA analysis is based on short tandem repeat sequences, also known as microsatellites or STRs (Short Tandem Repeats), which are not part of the protein-coding parts and have distinctive characteristics. DNA represents a major advance in forensic science for the identification of unknown persons or for the purpose of determining the source of biological samples left at crime scenes. [61]

Figure 2.10: DNA Representation

- **Body odor:** Every human body gives off an odor that characterizes its chemical composition and could be used to distinguish different individuals [62]. A person's primary odor contains constituents that are stable over time regardless of diet or environmental factors. Secondary odor contains constituents that are present as a result of diet and environmental factors.

  environmental. Third odour contains constituents that are present due to the influence of external sources (i.e. lotions, soaps, perfumes). For individual identification by human odour, the primary odour must have constituents that are stable over time. The compounds present in extracts of male and female axillary secretions containing the characteristic odours present in the axillary region have been isolated and identified.
- **The shape of the ear:** Recognition of the human ear is a new biometric technique. The French criminologist Bertillon was the first to suggest that people could be identified by the shape of their outer ear. A little later, the American policeman Iannarelli proposed the first auditory recognition system based on twelve characteristics.

  Iannarelli experimentally found that ten thousand ears were different even in identical twins . [63] Ear biometrics has been used in many government systems such as forensics, security and law enforcement. [64] In fact, the ear has attracted the interest of the biometrics community because of the following advantages: Firstly, its size speeds up the recognition task and increases its efficiency. Secondly, the ear has a uniform colour distribution, which ensures that the relevant information is retained when converted to a greyscale image. Thirdly, the ear does not require a great deal of collaboration from the user. [64] Ear biometrics is therefore a good choice because it offers a good compromise between accuracy and cost. Ear biometrics has been used in many government and commercial applications, such as forensics and security. For example, in the United States, an ear classification system based on manual measurements has been in use for over 40 years. In addition, the US Immigration and Naturalization Service provides specifications indicating that the right ear must be visible. [64]

- **Lip shape:** The lip is the tactile sensory organ that forms the visible part of the mouth. Studies have shown that the grooves of human lips are unique to each person and can therefore be used for human identification. [65] Lip shape characteristics have been widely used in forensic medicine by experts and by law for human identification. The challenge of using the lip as a biometric lies in the area of uniqueness and circumvention. The use of the lip as a means of human identification was first proposed through the concept of lip prints.

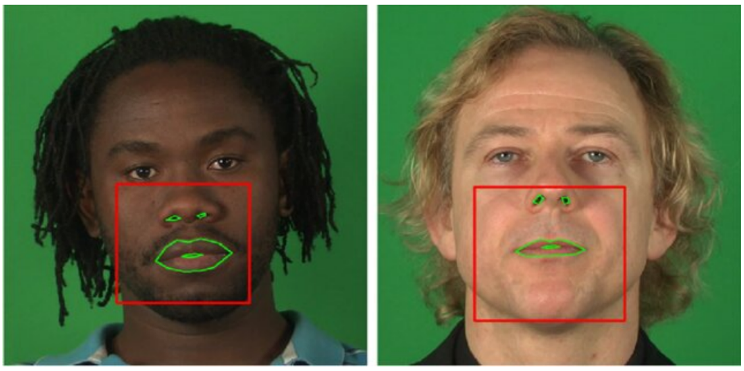


Figure 2.11: Exemple of lips detection [?]

- **Vein biometrics:** It has long been thought that the vein pattern in human anatomy might be unique to individuals. As a result, vein scanning over the years has enjoyed several successes, with wrist network scanning and, more recently, finger scanning. This technique uses a ”palmar venous network scanner”; to be identified, it is necessary to place the affected area above the reader. This is the model used by the vein network to retain characteristic points. [61] Vein biometry is considered a reliable and robust modality, as the overall structure of the veins does not change over time (with the exception of certain changes due to vein dilation or pathological cases). Spoofing biometric systems using veins can be difficult, but recent experiments have shown that the modality can be imperfect, particularly when identification and authentication are carried out in unsupervised mode (absence of human control).

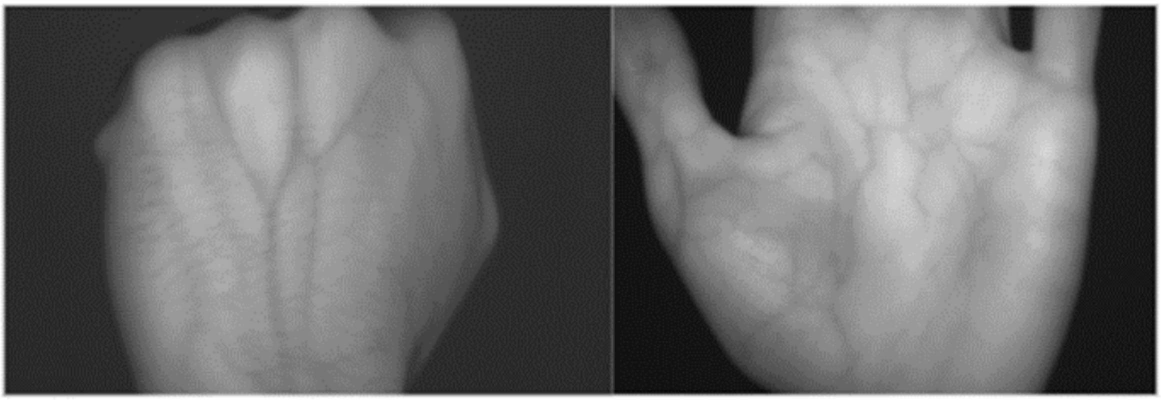
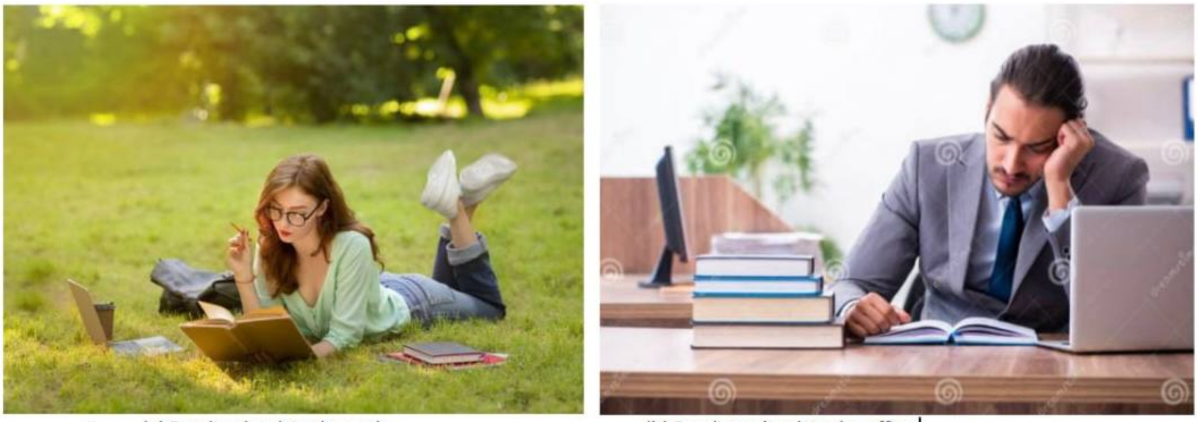


Figure 2.12: Image of the vein configuration system

2.6 Emotion Recognition:

The goal of human emotion recognition is to automatically classify a user's temporal emotional state based on input data. While there are numerous definitions of emotions, it can be broadly defined as a reaction to stimuli lasting seconds or minutes. Mood refers to an emotional state that lasts for hours or days, and personality is an inclination to feel certain emotions. We use the term 'emotional state' to refer to a person's current state, regardless of whether it originates from stimuli, mood, or personality [3]. The question remains: Are feelings hardwired within us, or do they change according to the situation and environment we are in? For example in Figure 1 there are 2 people doing a common activity 'reading a book'. When we pose the question: what are they feeling or thinking?, we get different answers for each of the person. We perceive those differences due to their surroundings. The person in Figure 1.a is reading in the park, and by the looks of his posture and the clothes he is wearing, he seems to be in a relaxed mood; whereas the person in Figure 1.b is reading in an office and his formal clothes along with his posture gives us the impression that he is in work mode. However, observe that both of them are engaged in their respective tasks. We see here that the difference of perception in their emotional state is caused by multiple reasons, but their surrounding environment plays and important part .



(a) Reading book in the park

(b) Reading a book in the office

Figure 2.13: Two people doing the same activity of reading a book in different surroundings. Depending on their surroundings, their perceived emotional states are different

2.7 Sources of Information for Emotion Recognition

In our daily lives, we are surrounded by a multitude of things. Depending on the place and occasion, we encounter various inanimate objects such as utensils, furniture, machines, vehicles, infrastructure, and natural elements. Additionally, we are surrounded by people like colleagues, friends, and family, depending on the circumstances, location, and time of day. As a result, each person's environment is unique and constantly changing. These surroundings can influence a person's emotional state in many ways. Consequently, an observer's perception of someone's emotional state is also significantly affected by these diverse situations. Here, we identify and discuss some sources of context that we can perceive through our vision.

2.7.1 Face Pose:

Facial expressions have always been the primary source for identifying emotions and remain a central focus in emotion recognition research. The visibility and orientation of the face provide varying types of information for recognizing emotions. For example , in Figure 2, the same facial expression is shown with nine different face poses. When viewed separately, each image presents unique visual information.

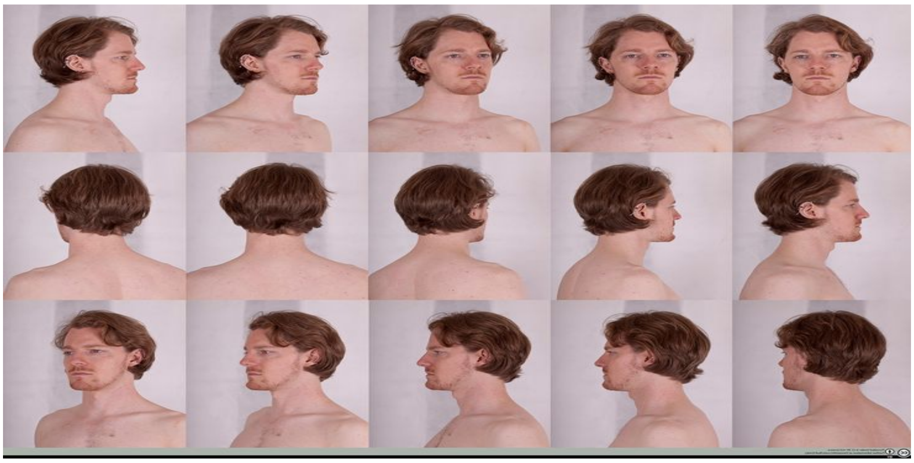
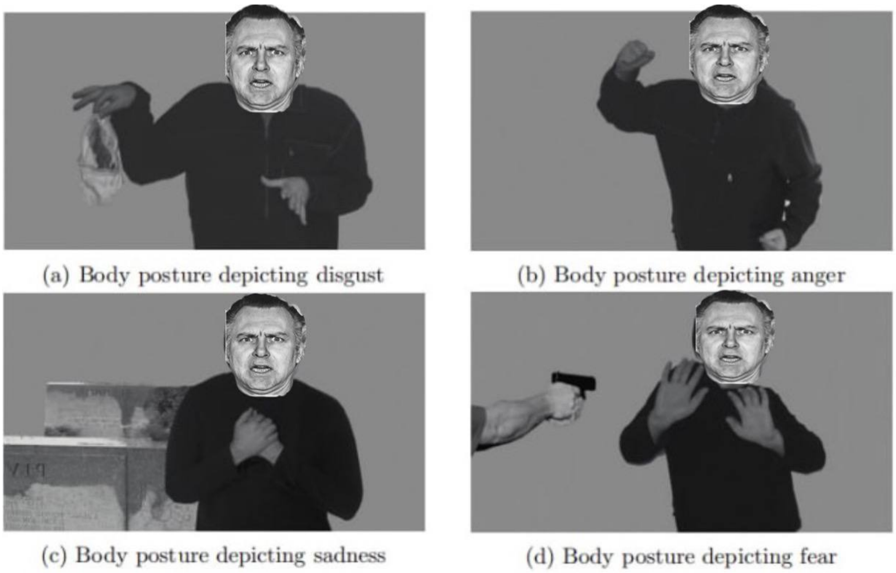


Figure 2.14: Visual information for emotion recognition changes with the change in face poses, for a fixed facial expression

2.7.2 Body Posture:

One of the more effective way of communication includes the way one displays his body posture. For example, when the faces in Figure 3 are seen independent of their associated body pose, the emotion perceived is Disgust.



(a) Body posture depicting disgust

(b) Body posture depicting anger

(c) Body posture depicting sadness

(d) Body posture depicting fear

Figure 2.15: Keeping the facial expression fixed, the body posture influences the perceived emotion of the person

However, the perception of emotion changes based on different body postures. When the same facial expression is combined with different body contexts, emotions such as anger (Figure 3.b), sadness (Figure 3.c), and fear (Figure 3.d) are perceived. This has been experimentally demonstrated, showing the influence of body pose on emotion perception.

These examples suggest that body posture is a crucial context for emotion perception and should be included in the emotion recognition process.

2.7.3 Hand Gestures :

Hand gestures are an essential aspect of body posture. Here, we examine their influence on emotion perception while keeping both facial expression and body posture constant. The gestures we use, often chosen subconsciously through experience, enhance the expression of our feelings. A simple hand movement can significantly affect how we perceive a person's emotions. As illustrated in Figure 4, which is similar to Figure 3, the same facial expression combined with different hand gestures can lead to different emotional interpretations. For example, there is a stark contrast between gestures 1 and 4: gesture 2 indicates positive approval, whereas gesture 5 suggests rejection or disapproval. Gestures provide important context for understanding emotions. Additionally, there are comprehensive surveys on gesture recognition, particularly focusing on hand and facial gestures.

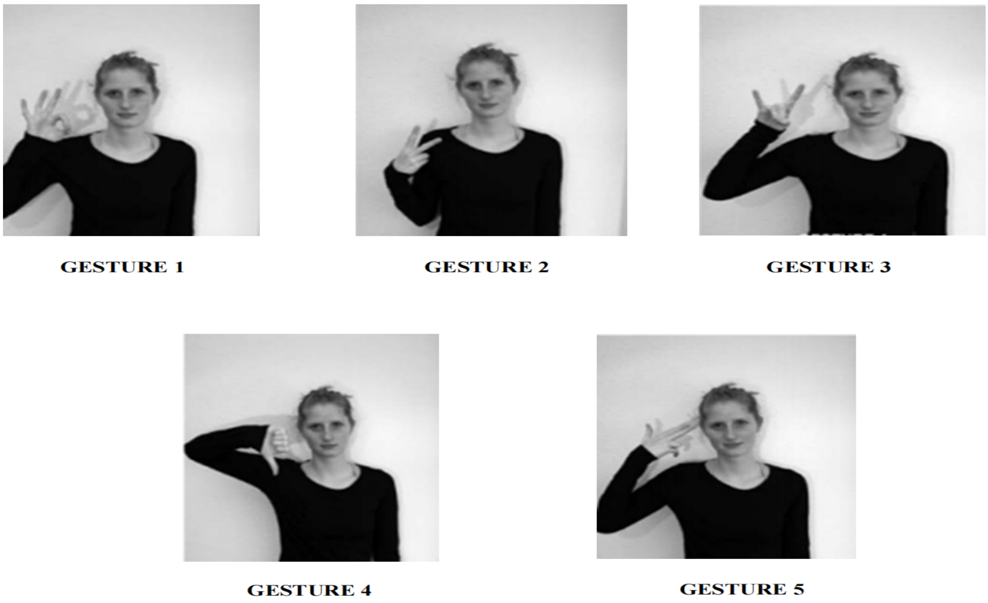


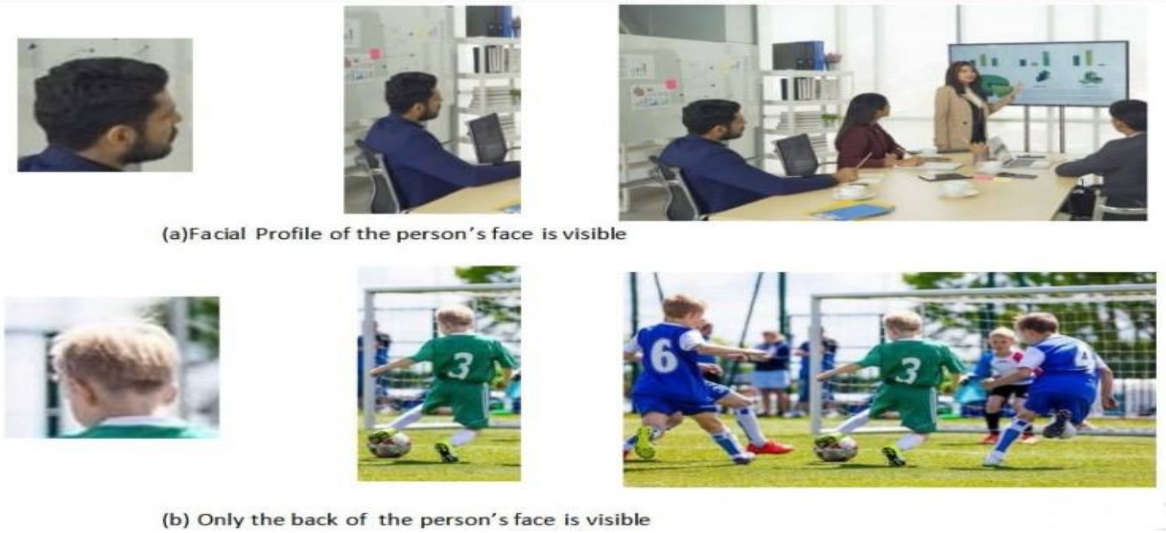
Figure 2.16: Five different hand gestures suggesting distinct emotions (Link)

2.7.4 Visual Scene:

We travel to different places, cities, or countries to explore new towns, their unique features, people, food, and culture. While walking through a new village or neighborhood, we encounter various sights. We might see a building that is unfamiliar or familiar but with a different architectural style. Sometimes, streets or shops remind us of our own neighborhood or something from our past. Such experiences evoke new feelings and generally make us happy. However, there are times when situations or scenes can be disturbing or annoying, causing feelings of sadness, anger, or fear. For instance, walking through a deserted street in an unfamiliar town at night, with complete darkness and silence, can be frightening. This visual scene directly impacts a person's emotions. Our perception of our surroundings strongly influences our emotional state, as the visual environment provides comprehensive information on the contexts affecting our feelings.

2.7.5 Role of Context in Emotion Recognition:

The place and/or the social situation that the person finds himself affects his emotional state and also influences the manner in which his feelings are perceived by an observer. The context of the situation is an important aspect while analysing people's feelings. Despite research focusing on facial expression and body pose, there is ample researchs work that asserts the importance of context in emotion perception. We try to understand the role of context through visual example shown in Figure 2.17



(a)Facial Profile of the person's face is visible

(b) Only the back of the person's face is visible

Figure 2.17: Examples showing people with different facial views, along with their body posture and the surrounding scene.

There are two columns: ”Face/Head, Body” and ”Person in Context.” Figures 2.17.a and 2.17.b illustrate examples of people whose faces and bodies are immersed in different situations.

In Figure 2.17.a, only the profile of the person's face is visible, making it difficult to determine their feelings. When their body posture is visible (second column, Figure 2.17.a), we get more information, indicating that the person is looking away towards something or someone, suggesting they are paying attention, but it's still insufficient to discern their emotions. Only when we see the entire scene (third column, Figure 2.17.a) does it become clearer that the person is in a meeting room, paying attention to someone speaking, likely feeling engaged in the activity. Figure 2.17.b presents an even more challenging scenario.

We see only the back of the person's head (first column), which provides no information about their emotional state. The body posture (second column) reveals part of the story, but it's only the whole image (third column) that gives a fuller, richer picture. We see the boy is playing with other kids, likely in a state of anticipation. This demonstrates that context is essential for predicting emotions when the face is not visible or is partially obscured. Even when the face is completely visible, the contextual cues in the visual scene heavily shape our emotional perception. This clearly shows how context affects a person's emotional state and is equally important in estimating a person's emotions. Context not only alters our perception of someone's emotional state but also impacts what the person is actually feeling.

2.8 Emotion Representation Format :

Automatic emotion recognition algorithms must capture the complex emotional states displayed by humans. Emotions can be described in various ways. Here, we highlight the most commonly used representation formats:

1. **Free Form:** The emotions are described in the form of sentences and paragraphs using words and phrases used in our daily lives. This text carries information about the underlying affective states which is usually reflected in the usage of certain words or grammatical alternatives. There have been attempts in figuring out the best representation form for the emotions described in Free Form by building markup languages [66].

   The HUMAINE database is an example where the authors tried to bridge the gap between emotion elicitation and it's annotation [67].

2. **Affect Dimensions:** The emotions are indicated using scales based on different affect terminologies. For example, Cowen and Keltner [2017] use 12 affect dimensions while gathering emotional experiences.[68] Annotators (or workers who are paid to do such tasks) report their experiences on a given scale (an example of such a scale would be the Likert scale (Likert [1932])) for each of those 12 affect dimensions.[69] Another, very popular approach is to use 3 independent affect dimensions (introduced by Mehrabian [1995]). [70] The three dimensions are viz. Valence, Arousal and Dominance. Valence measures how positive or pleasant an emotion is, ranging from negative to positive; Arousal measures the agitation level of the person, ranging from non-active / calm to agitated / ready-to-act; and Dominance measures the control level of the situation by the person, ranging from submissive / non-control to dominant / in-control [67]
3. **Emotion Categories:**Emotions are represented in a discrete form using specific words that describe their characteristics. Darwin [1998] was the first to propose the permanence of human expressions, suggesting that human emotions are universal and can be categorized into distinct emotion categories like Fear and Anger based on facial expressions [71]. Du et al. [2014] introduced a set of 21 facial emotion categories, defined as various combinations of basic emotions, such as happily surprised or happily disgusted [72]. This categorization allows for a more detailed understanding of expressed emotions. After numerous studies over several decades, modern psychologists and neuroscientists also support the discrete categorization of emotions (Ekman and Friesen [1969] , Izard [1971]).[73] , [74]

2.9 CONCLUSION :

We have seen that biometrics is increasingly becoming an indispensable tool for identifying people in a wide range of applications. It is taking its place as the number one method of authentication. It is gaining its place as the number one means of authentication. However, a number of challenges remain, such as attacks on biometric systems. Nevertheless, the future of biometrics for authentication and personal identification is promising. This chapter provides a general introduction to biometrics. It is an introductory chapter for the following chapters. We have therefore introduced the various concepts of this biometric field and given an overview of the characteristics of a biometric system as well as some widely used modalities.

Chapter 3

Result and discussion

3.1 Introduction

In this section, we present the experimental results of our approach for emotion expression recognition using a convolutional neural network model. Our study aims to evaluate the efficacy of our approach in verifying Emotion expression detection and the relation between object and Context. We present a comprehensive analysis of the performance of our approach in terms of recall, precision, and accuracy. Furthermore, we describe the experimental protocol used, which includes the datasets used, effect hyperparameters, and their fine-tuning. Finally, we present a detailed discussion of the obtained results, highlighting their significance and practical implications.

3.2 Performance Evaluation

Performance evaluation in deep learning refers to the process of assessing how well a deep learning model performs on a given task. It involves various metrics and techniques to measure the effectiveness, efficiency, and generalization ability of the model. Here are the key components involved in our performance evaluation:

- **Accuracy** we can define accuracy as a measure of the capability of classifying the samples correctly, which is expressed as follows:

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{3.1}$$

- **True Positive (TP):** the Emotion nature of sample is recognized by the marker.

- **True Negative (TN) :** for eg : the sample is Angry , and the marker detects it.

- **False Positive (FP):** the sample is happy , but the marker detects it as Angry.

- **False Negative (FN):** the sample is Angry but recognized by the marker as happy.

- **Recall (aka Sensitivity)** is a measure of the proportion of true positive cases that the model correctly In other words, classified as positive is a measure of the percentage of true positive cases that correctly identified a binary classification model using the ratio of true positive cases to the sum of true positive cases and false negative cases used.

$$\text{Sensitivity } = \frac{TP}{TP + FN} \tag{3.2}$$

FN : False negative.

TP : True positive.

- **Precision:** the probability of normal samples being correctly identified.

$$\text{Sensitivity } = \frac{TP}{TP + FP} \tag{3.3}$$

FP : False positive.

3.3 Loss Functions:

In the context of deep learning models, loss functions, also known as cost functions or objective functions, are mathematical constructs that quantify the disparity between the predicted output of the model and the actual target values. They play a crucial role in the training process by providing a metric for optimization algorithms, such as gradient descent, to minimize. This minimization guides the adjustment of model parameters to enhance predictive accuracy. Commonly used loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss for classification tasks, each tailored to specific types of predictive tasks to ensure optimal model performance.

3.3.1 Criterion for Emotion Categories ($\mathbf{L_{disc}}$):

In our emotion recognition task, we have defined 26 distinct emotion categories. Each individual in the EMOTIC dataset can be associated with multiple emotions. This means that there are 26 possible emotion categories, and each input can receive multiple labels, making this a multiclass-multilabel problem. Additionally, there is an inherent class imbalance, as the number of training examples varies across emotion categories and scores on the continuous dimensions. To address this, we employ a weighted Euclidean loss, which we found to be more effective than using Kullback-Leibler divergence or a multi-label multi-classification hinge loss.

The weighted Euclidean loss for our emotion categories is defined as follows:

$$L_{disc} = \frac{1}{N}\sum_{i=1}^{N} w_i \left(\hat{y}_i^{disc} - y_i^{disc}\right)^2 \tag{3.4}$$

3.3.2 Criterions for Continuous Dimensions ($\mathbf{L2_{cont}}$,$\mathbf{SL1_{cont}}$):

In this learning task, there are three dimensions with values that need to be learned. Each dimension has values ranging from 1 to 10, so we approach this task as a regression problem. For the validation and test sets, multiple annotators provide annotations. Since the annotation is based on subjective evaluation, we assess performance using two different robust loss functions:

$$L_{2\ \mathrm{cont}} = \frac{1}{\sharp\mathcal{C}}\sum_{k\in\mathcal{C}} v_k \left(\hat{y}_k^{\mathrm{cont}} - y_k^{\mathrm{cont}}\right)^2 \tag{3.5}$$

where

- $C = (Valence, Arousal, Dominance)$ and $\sharp\mathcal{C} = 3$
- $\hat{y}_k^{\mathrm{cont}}$ and y_k^{cont} are the prediction
- the normalized ground-truth for the k^{th} dimension
- $v_k(= 0, 1)$ is a binary weight to represent the error margin

A margin Euclidean loss $\mathbf{L2_{cont}}$:

The first method establishes a margin of error (v_k) when calculating the loss, within which the error is disregarded. This margin Euclidean loss for the continuous dimension is defined as:

$$\begin{aligned} v_k = 0, &\quad \text{if} \quad \left|\hat{y}_k^{\mathrm{cont}} - y_k^{\mathrm{cont}}\right| < \epsilon \\ = 1, &\quad \text{otherwise} \end{aligned} \tag{3.6}$$

the Smooth L1 SL1$_{\text{cont}}$:

The Smooth L1 loss calculates the absolute error, using the squared error if the error is below a certain threshold (set to 1 in our experiments). This loss function has been widely used in object detection ([75]) and, in our experiments, has proven to be less sensitive to outliers. Specifically, the Smooth L1 loss is defined as follows:

$$SL_{1\ \text{cont}} = \sum_{k=1}^{3} v_k \begin{cases} 0.5x_k^2, if\ |x_k| < 1 \\ |x_k| - 0.5,\ \text{otherwise} \end{cases} \tag{3.7}$$

3.3.3 Combined Criterions (L$_{\text{comb1}}$, L$_{\text{comb2}}$):

We define the combined loss function as a weighted combination of the two distinct losses described above:

$$L_{\text{comb }1} = \lambda_{\text{disc}}\, L_{\text{disc}} + \lambda_{\text{cont}}\, L_{2\ \text{cont}} \tag{3.8}$$

$$L_{\text{comb }2} = \lambda_{\text{disc}}\, L_{\text{disc}} + \lambda_{\text{cont}}\, SL_{1\ \text{cont}} \tag{3.9}$$

The parameters $\lambda(disc, cont)$ determine the importance of each loss and are set empirically using the validation set. After testing various combinations, we found that setting $\lambda_{\text{disc }=}\lambda_{\text{cont }=0.5}$ yields the best performance. Equal weights ensure equal importance to both criteria, preventing bias toward any particular task. L$_{\text{disc}}$ and L$_{\text{cont}}$ (L2$_{\text{cont}}$, SL1$_{\text{cont}}$) correspond to the losses for learning emotion categories and continuous dimensions, respectively. This combined loss criterion is used to implement multi-task training computationally. It calculates the loss for both tasks and backpropagates it through the model during training. This approach allows the model to generalize its performance by learning the parameters in a shared manner.

3.4 Performance Evaluation Metrics:

The task of learning emotion categories is modeled as multi-class classification, and the continuous dimension learning task is modeled as regression. That is why we measure the performance of our fusion model using two different evaluation metrics.

3.4.1 Average Precision (AP):

For this evaluation metric, we first determine the threshold for each emotion class of the validation set predictions. After the model enters evaluation mode, the forward pass generates 26 predictions for samples in the validation set - probabilities for the 26 emotion classes. We collect all such predicted probability values for all validation set samples. Next, for each emotion class, its predictions for all samples are used along with the labels to plot a precision-recall curve [76]. For each of these 26 curves, we find a point where precision = recall, which is the threshold for that particular class. We now have 26 thresholds for the 26 emotion classes. These thresholds are then applied to the model's predicted probabilities for the test set samples. For example, if we iterate forward through the test set samples, the model predicts 26 values for the emotion class. If the predicted value is above its respective threshold, this indicates that the model is more likely to trigger that emotion class. Similarly, we apply the threshold to all the test set samples to find the predicted emotion class for each sample. In addition to this, to measure the performance of the model as a whole, we take the predicted probabilities of the test samples and their respective labels and find the precision-recall curves for each emotion class. The area enclosed by these curves (for each emotion class) represents the so-called average precision. This metric represents the average performance of the model for the emotion class and can take values between 0 and 100, where 0 represents precision. Both precision and recall are zero and 100 represents perfect recall and precision.

3.4.2 Average Absolute Error (AE):

The continuous dimension task is a regression and we use mean absolute error (AE) to measure their performance. The error for a given continuous dimension is the absolute difference between its predicted value and the target value. An error will be considered for calculation only if it is below a predefined error bound. Such errors will be averaged over all inputs for each continuous dimension. This performance measure is similar to the continuous dimension criterion, except that we consider absolute values instead of squares. Hence AE is calculated as follows:

$$AE = \frac{1}{\mathcal{C}} \sum_{k \in \mathcal{C}} v_k \left| \hat{y}_k^{\text{cont}} - y_k^{\text{cont}} \right| \tag{3.10}$$

where $C = (Valence, Arousal, Dominance)$, $\hat{y}_k^{\mathrm{cont}}$ and y_k^{cont} cont are the prediction and the normalized ground-truth for the k^{th} dimension, m is the number of samples and $v_k(= 0, 1)$ is a binary weight to represent the error margin. This error is also weighted in the same manner as the loss criterion.

This AE vector is the metric to represent the average performance of the model for the continuous dimensions. Lower the AE, better the performance.

3.5 Real time Object detection:

3.5.1 YOLO V3:

You Only Look Once (YOLO) could be a viral and broadly utilized calculation [77]. YOLO is popular for its object detection characteristic. In 2015, Redmon et al. gave the presentation of the primary YOLO adaptation [78]. Within the past a long time, researchers have published a few YOLO consequent adaptations depicted as YOLO V2, YOLO V3, YOLO V4, and YOLO V5 [79]-[80]. There are many revised-limited forms, such as YOLO-LITE [81]-[82].

3.5.2 Bounding Box:

our system predicts bounding boxes utilizing measurement clusters as stay boxes [83]. The arrange predicts 4 arranges for each bounding box, $\mathrm{t_x}$, $\mathrm{t_y}$, $\mathrm{t_w}$, $\mathrm{t_h}$. In case the cell is balanced from the best cleared out corner of the picture by ($\mathrm{c_x}$, $\mathrm{c_y}$) and the bounding box earlier has width and tallness $\mathrm{p_w}$, $\mathrm{p_h}$, at that point the forecasts compare to:

$b_x = \sigma(t_x) + c_x$

$b_y = \sigma(t_y) + c_y$

$b_w = p_w e t_w$

$b_h = p_h e t_h$

Amid preparing we utilize whole of squared blunder misfortune. In case the ground truth for a few facilitate forecast is tˆ * our angle is the ground truth esteem (computed from the ground truth box) short our expectation: tˆ * − t* . This ground truth esteem can be effortlessly computed by rearranging the conditions over. YOLOv3 predicts an objectness score for each bounding box utilizing calculated relapse.

This ought to be 1 on the off chance that the bounding box earlier covers a ground truth question by more than any other bounding box earlier. In case the bounding box earlier isn't the finest but does cover a ground truth protest by more than a few limit we disregard the expectation, taking after [84]. We utilize the limit of .5. Not at all like [84] our framework as it were allots one bounding box earlier for each ground truth question. On the off chance that a bounding box earlier isn't alloted to a ground truth question it causes no misfortune for arrange or course expectations, as it were objectness.

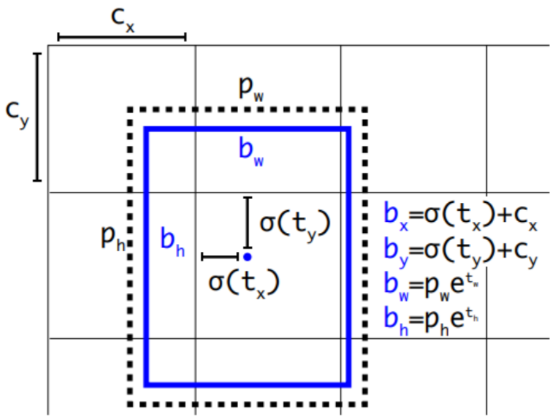


Figure 3.1: YOLO V3 Bounding boxes with dimension priors and location prediction.

3.6 ResNet50 Pytorch:

ResNet (Residual Network) is a deep convolutional neural network architecture developed by Kaiming He et al. in 2015. It is known for its ability to train deep neural networks with up to hundreds of layers without suffering from the vanishing gradient problem that can occur in traditional deep networks. The key innovation of ResNet is the use of residual connections, which allow information to flow directly from one layer to the next, bypassing the intermediate layers. This helps solve the problem of vanishing gradients by providing a shortcut for the gradients to flow back through the network.

ResNet consists of a series of convolutional layers and several residual blocks. Each residual block contains two or more convolutional layers and has shortcut connections that bypass the intermediate layers. The output of each residual block is added to the input of that block before being passed to the next block in the network.

There are several variants of ResNet, including ResNet-18, ResNet-34, ResNet-50, and ResNet-152, which differ in the number of layers and the size of the filters used in the convolutional layers. ResNet-50, as illustrated in the figure, is one of the most popular architectures, with 50 layers and a combination of 3x3 and 1x1 filters. It includes an initial 7x7 convolutional layer followed by a max pooling layer and multiple residual blocks. Each residual block in ResNet-50 is structured with a series of 1x1 and 3x3 convolutional layers.

ResNet has achieved state-of-the-art results in various computer vision tasks, including image classification, object detection, and semantic segmentation. Its ability to train deep networks has made it a popular choice for many applications in both industry and academia.[85]

The figure 3.1 illustrates the basic model of ResNet-50 implemented in PyTorch, showing the detailed arrangement of convolutional layers, residual blocks, and the overall architecture of the network.

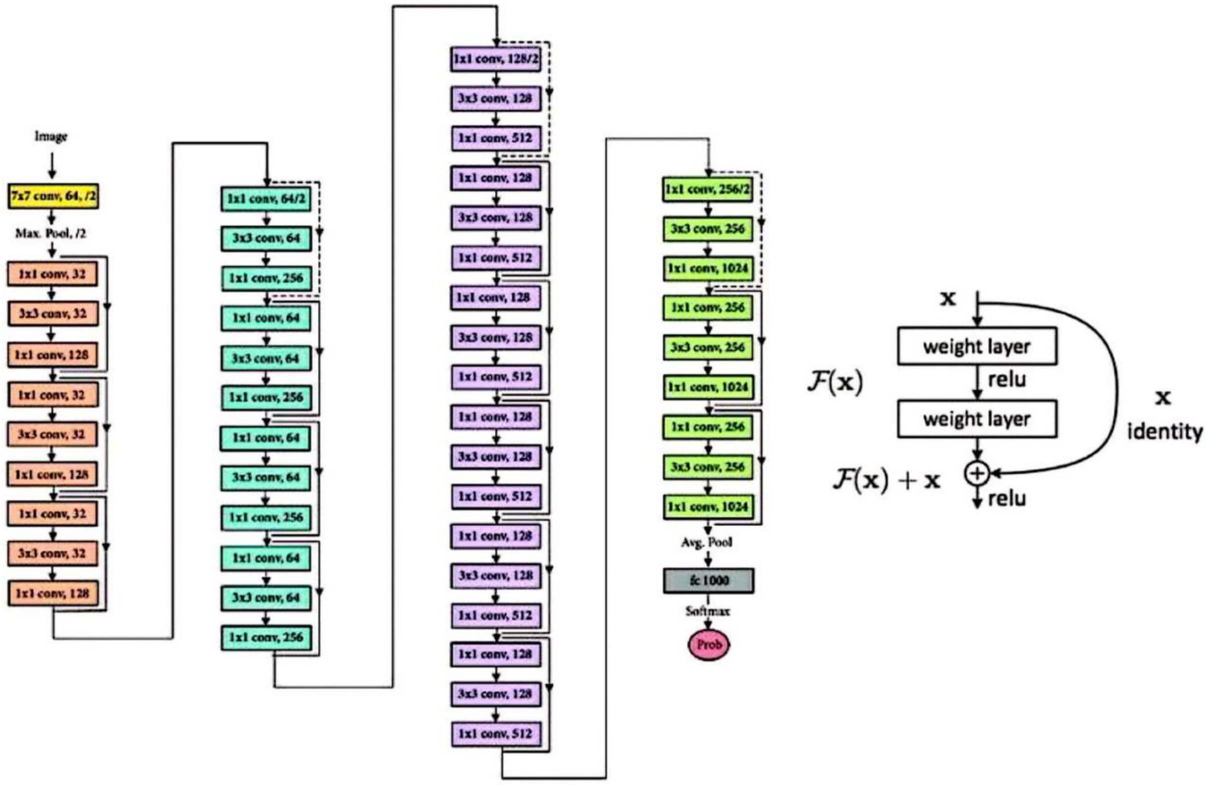


Figure 3.2: Deep architecture of ResNet- 50 Pytorch

3.7 Work Environment:

3.7.1 Hardware Environment:

In this section, we describe the hardware environment used for our experiments. The system configuration includes various components essential for the performance and precision of our results we use high performance graphic card.

.**graphic unit: NVIDIA Quadro RTX A5000 [24GB, 8192 CUDA]**

3.7.2 Software Environment:

- **Tensorflow :** Google's Brain team developed a deep learning framework called TensorFlow, which supports languages such as Python and R and uses data flow graphs to process data. This is very important because as you build these neural networks, you can watch the data flow through the neural network. TensorFlow machine learning models are easy to build, can be used for robust machine learning production, and enable powerful experimentation for research. With TensorFlow, you also get TensorBoard for data visualization, which is a big package that often goes unnoticed. TensorBoard simplifies the process of displaying data visually when working with your stakeholders. You can also use the R and Python visualization packages.
- **Keras:** Keras was originally developed by Francois Chollet, with over 350,000 users and over 700 open source contributors, making it one of the fastest growing deep learning framework packages. Keras supports high-level neural network API, written in Python. What makes Keras interesting is that it runs on top of TensorFlow, Theano, and CNTK. Keras is used in several startups, research labs, and companies, including Microsoft Research, NASA, Netflix, and Cern.
- **Scikit learn:** scikit-learn is an open source machine learning library written in Python [86]. It allows the easy and fast integration of machine learning methods into Python code. The scikit-learn library comprises a wide bandwidth of methods for classification, regression, covariance matrix estimation, dimensionality reduction, data pre-processing, and benchmark problem generation.[87]
- **Paddle Paddle:** PaddlePaddle, short for PArallel Distributed Deep LEarning, is an open-source deep learning platform developed by Baidu. It provides a compre-

hensive suite of tools and libraries to support various AI tasks such as computer vision, natural language processing, and speech recognition. PaddlePaddle is designed to be highly efficient and scalable, making it suitable for both research and industrial applications. The platform includes over 130 pre-trained models, enabling users to quickly build and deploy AI applications.[88]

- **Caffe:** Caffe provides a complete toolkit for training, testing, finetuning, and deploying models, with well-documented examples for all of these tasks. As such, it's an ideal starting point for researchers and other developers looking to jump into state-of-the-art machine learning. At the same time, it's likely the fastest available implementation of these algorithms, making it immediately useful for industrial deployment. [89]
- **Onnx:** The Open Neural Network Exchange (ONNX) [90] is an open ecosystem that allows AI developers to choose the right tools as their projects evolve.
  ONNX provides an open source format for AI models (both deep learning and traditional machine learning). ONNX defines the computational graph of a deep learning model and the various operators used in the model. It provides a set of specifications for converting models to the base ONNX format, and another set of specifications for getting models out of ONNX form. At a high level, ONNX is designed to enable framework interoperability. There are many excellent machine learning libraries in many different languages: PyTorch , TensorFlow , MXNet , and Caffe are just a few that have become very popular in recent years, but there are many more. Machine learning models can be converted to the serialized ONNX format, which can then be run on a range of devices.
  ONNX Runtime is an inference engine written in C++.framework used to deploy ONNX format models into production. It works on diverse hardware and support both deep learning as well as traditional machine learning models. [91]
- **PyTorch:** With the release of PyTorch in 2016, Facebook (Meta) introduced a framework that had the benefits of being both user-friendly and efficient [92]. PyTorch is a Python library that allows the user to have a flexible code base that can be adapted to new features, while still maintaining efficiency through precompiled C++ code. PyTorch also has a straightforward 11 way of converting its models to CoreML by using the Python library coremltools. However, for a framework

where the model can be manually configured, trading speed is acceptable in many cases. Today PyTorch is used in many repositories as the main framework for their implementation.

3.8 Proposed approach :

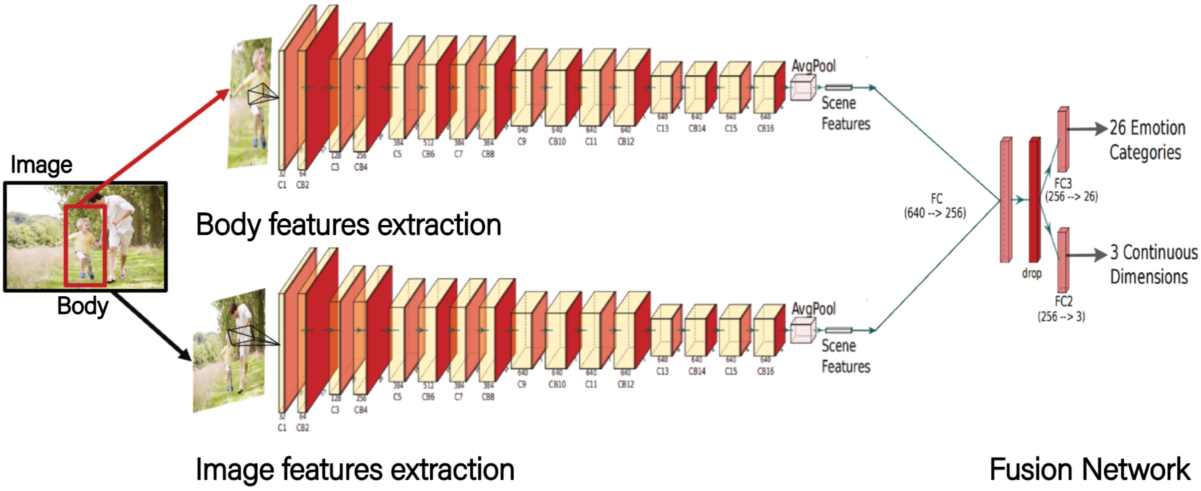


Figure 3.3: Proposed model for emotion recognition in context jointly the discrete categories and the continuous dimensions.

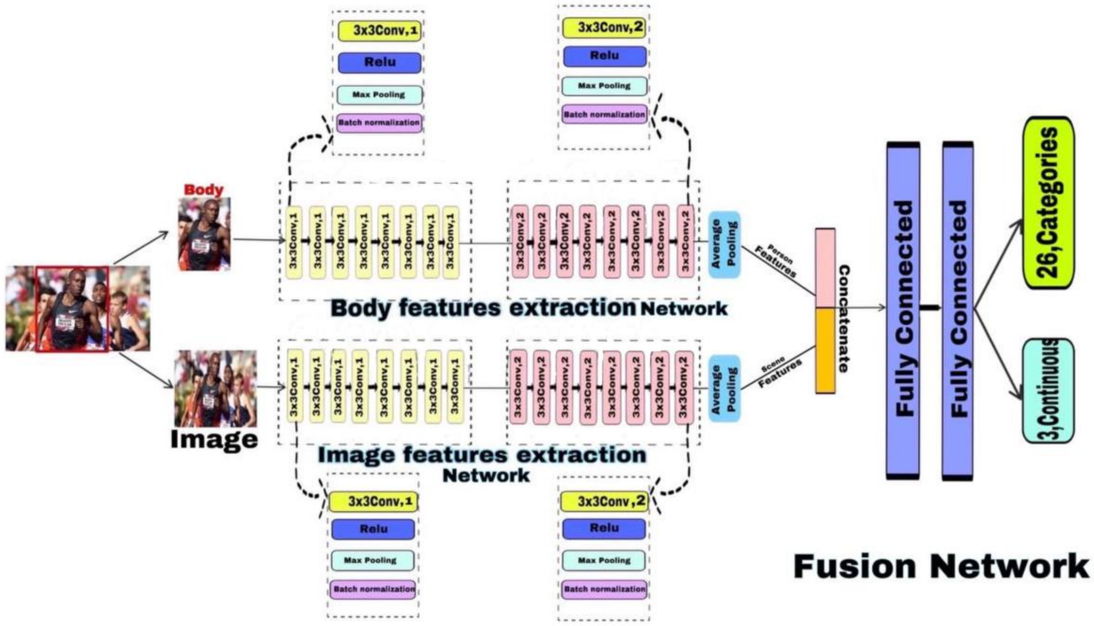


Figure 3.4: Approach model for emotion recognition in context jointly the discrete categories and the continuous dimensions

The proposed CNN architecture for emotion recognition in scene context consists of three main modules: the body feature extraction module, the image feature extraction module, and the fusion network. The body feature extraction module takes the visible part of the person's body as input, generating body-related features that include important cues like face and head aspects, pose, and body appearance. This module is pre-trained on the ImageNet dataset, which is object-centric and includes categories related to people. The architecture of this module is based on a one-dimensional filter CNN with 16 convolutional layers that use 1-dimensional kernels alternating between horizontal and vertical orientations, effectively modeling 8 layers using 2-dimensional kernels. A global average pooling layer is used to reduce the features of the last convolutional layer, and batch normalization layers are added after each convolutional layer to avoid internal covariate shift and speed up training with Rectified Linear Units (ReLU).

The image feature extraction module takes the entire image as input and generates scene-context features. These features encode the scene category, its attributes, objects present in the scene, and the dynamics between other people present. This module used pre-trained with the scene-centric Places(Place-365) dataset to capture contextual features. Its architecture is similar to the body feature extraction module, also based on the one-dimensional filter CNN with 16 convolutional layers, global average pooling, and batch normalization.

The fusion network combines the features from both the body and image feature extraction modules. It performs fine-grained regression for two types of emotion representations: discrete emotion categories and continuous emotion dimensions (Valence, Arousal, Dominance). The fusion network includes two fully connected (FC) layers. The first FC layer reduces the dimensionality of the combined features to 256. The second FC layer learns independent representations for each task, with the output branching into 26 units for discrete emotion categories and 3 units for the continuous dimensions.

Preparing the model for the train:

This step involves modifying the model parameters to adapt to our selected database. To achieve optimal results, we have pretrained the model multiple times. The key parameters set for learning are : batchsize, epoch, activation function, optimizers, loss function .

1. **Batch-size:** Batch size refers to the number of samples used in each iteration. Selecting the appropriate batch size ensures a sufficiently stable estimate of the gradient for the entire dataset. We conducted several experiments to determine the optimal batch size for our context.
2. **Epoch:** The total number of learning iterations, defined as a criterion, is set to 300 epochs.
3. **Activation function:** In a neural network, this determines how the weighted sum of inputs is converted into the output from one or more nodes in a layer.
4. **Optimizer:** An optimizer is an algorithm or method used to adjust the attributes of your neural network, such as weights and learning rate, to minimize losses.

3.9 EMOTIC Dataset :

The EMOTIC dataset was created in two versions:

1. The first version had a total of 18,316 images annotated. There are images with more than one person, and usually these people are also annotated with emotional labels. So there are 23,788 people in total, each with their own annotations. After splitting the dataset into training (70%), testing (20%), and validation (10%), each person in the test set was annotated by 2 additional different annotators. This additional annotation was done to provide a comprehensive test set of images with labels from several different annotators. This created a larger term base created by each person, which made it more efficient to test the model after training.

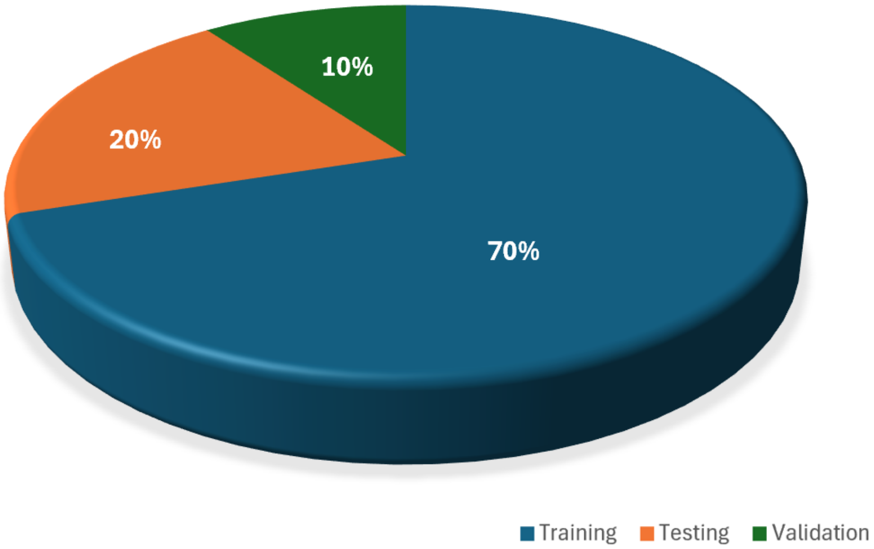


Figure 3.5: EMOTIC Dataset split

2. The second version added 44% more people to the previous set, with a final count of 34,320 people in 23,571 images. All newly added instances in the test set were also annotated with 2 additional annotators, similar to the previous version. In order to thoroughly analyze the annotator agreement, each annotation in the validation set was annotated with 4 additional different annotators. Since the validation set contains more images compared to the test set, it provides more explicit content for performing inter-annotator agreement analysis.

The creation of EMOTIC involved several stages for both versions, some of which were iterative. It started with collecting images while creating the format for expressing emotions. Both stages influenced each other a lot.

3.9.1 Image Data Collection :

EMOTIC is an image-based dataset whose main subject is people. Therefore it is natural for us to look at currently available datasets that have similarities to the features we need in our dataset. Another important aspect is to collect only images where the location of the subject (in our case, the person) in the image is available. Figure 3.1 shows an example of this. When generating annotations, the subject must be located and the image must be localized so that it can be distinguished from the context to avoid ambiguity.

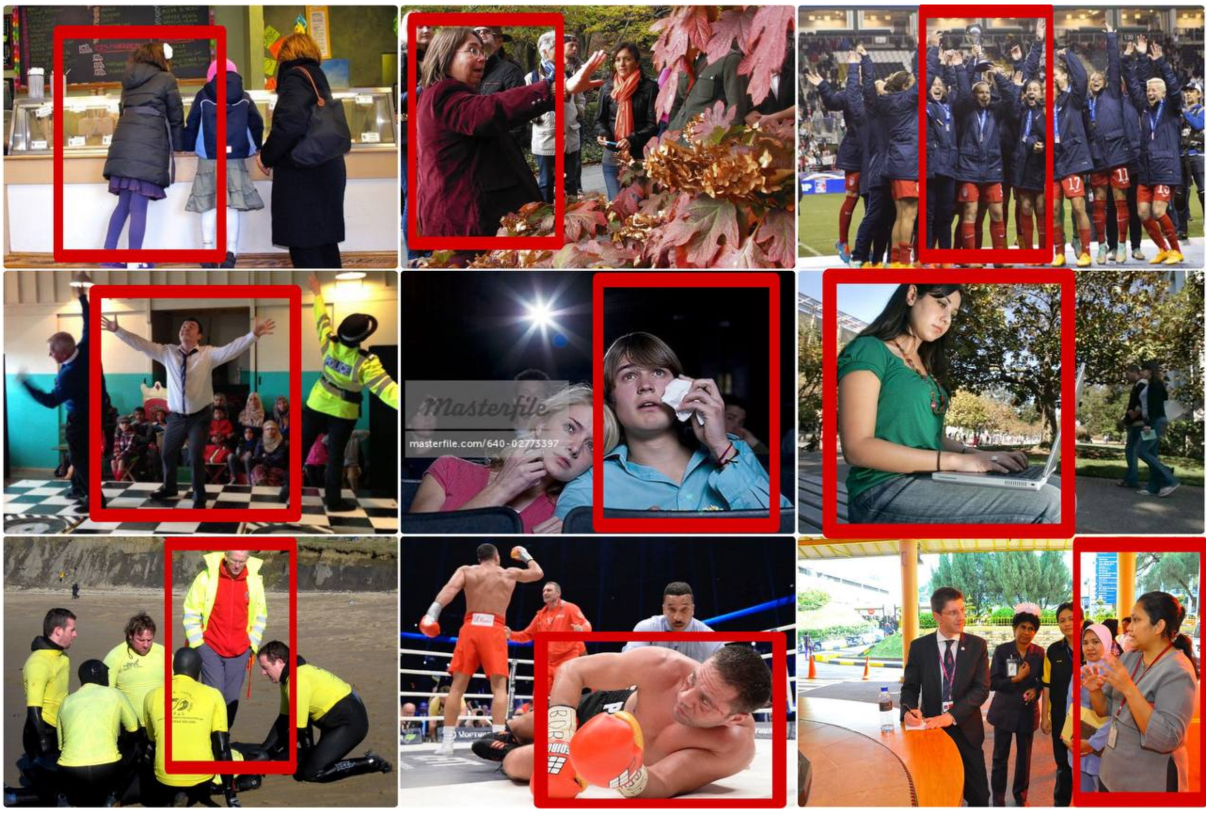


Figure 3.6: Example images from EMOTIC: The person-in-context is enclosed in a rectangular bounding-box.

The images in EMOTIC from the following 3 sources :

- **COCO (COmmon objects in COntext) :** COCO is an extensive dataset for object detection, segmentation, and captioning. It includes images with corresponding bounding boxes for the objects within them. There are 80 annotated object categories, complete with bounding boxes and segmentation masks, including people. We filtered COCO for images containing only people and added these, along with their associated metadata (including bounding boxes), to our collection.
- **the Internet using the search engine :** We used search engine Like Google to search with terms from our curated list of emotion categories and collected images that met the desired criteria. The people in these images were manually localized using bounding boxes and then included in our collection.
- **ADE20K (Scene Parsing Benchmark)** ADE20K is a scene parsing dataset with detailed annotations of objects and their parts. We extracted images containing people from ADE20K and added them to our collection, including their respective bounding boxes. they divided the images into three sets: Training (70%), Validation (10%), and Testing (20%) maintaining a similar affective category dis-

tribution across the different sets.[93]

3.9.2 Caracteristic of EMOTIC dataset :

1. **Appearance of Subjects:** The images of people in EMOTIC are natural and not staged (as they might be in laboratory settings). They are not limited by facial expressions, head pose, or body postures.
2. **Presence of Context:** The background or surrounding environment is included in the images and is not confined to any specific location or setting. The images can depict any place, viewpoint, or social situation. People may be engaged in various activities and can have any objects around them, including other people.
3. **Extensive Emotion labels:** The subjects are labeled with a comprehensive range of emotions. These emotion representations are clear and cover the broadest spectrum of human emotions. The apparent emotions of individuals are depicted through a combination of 26 detailed discrete emotion categories and 3 continuous dimensions. [94]

Transforming DATA :

These techniques help to simulate different lighting conditions and orientations that the model may encounter in the real world, thus aiding in the development of a more robust and accurate emotion recognition system.

- **Random Horizontal Flip:** This augmentation technique mirrors the image along the vertical axis, effectively doubling the dataset size with mirrored versions of each image. It is particularly useful for datasets that do not have a fixed horizontal orientation.
- **Color Jitter:** This technique randomly alters the brightness, contrast, and saturation of images within specified ranges. For this project, all three attributes are adjusted by a factor of 0.4. This helps the model become robust to variations in lighting and color saturation that may occur in real-world scenarios.

3.10 Emotion representation Format in EMOTIC

The EMOTIC dataset employs a dual approach to emotion representation , Continuous Dimensions and Emotion Categories.
on Continuous Dimensions utilizes the VAD model which quantifies emotions through three continuous dimensions: Valence, Arousal, and Dominance. Each dimension is scored on an integer scale ranging from 1 to 10. the dataset includes examples of individuals annotated with varying levels of these dimensions.

Affect Dimensions (Continuous Dimensions)

are straightforward to understand and apply. They involve continuous dimensions, as the recorded values are real numbers. Specifically, the Valence, Arousal, and Dominance (VAD) dimensions are used in emotion representation, known as the Emotional State Model, originally proposed by Mehrabian in 1995. In this model, emotions are depicted as a tuple (V, A, D) with values ranging from 1 to 10, forming a 3D Cartesian coordinate system.

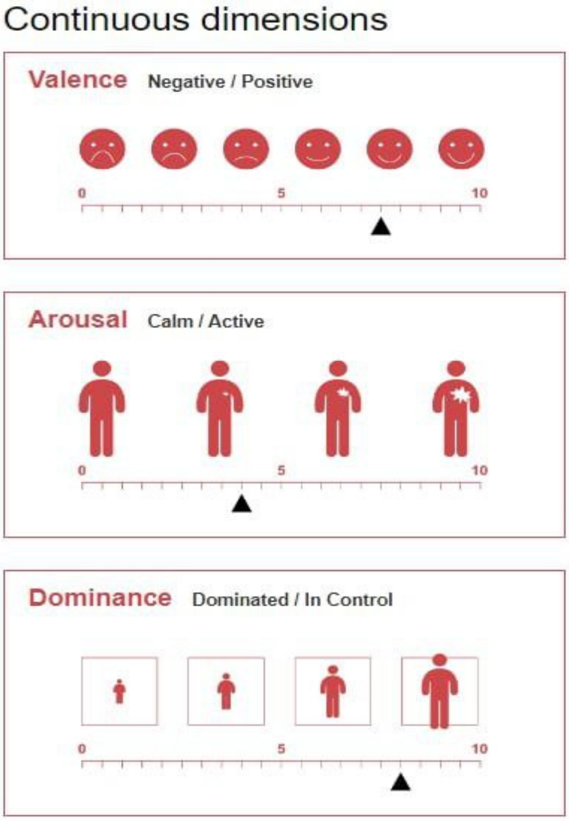


Figure 3.7: Examples of the 3 continuous dimensions , Valence, Arousal Dominance.

- **Valence** indicates the positivity or pleasantness of an emotion, with lower values signifying negative emotions and higher values indicating positive ones.
- **Arousal** measures the level of activity or excitement, where low values correspond to calmness and high values to heightened activity.
- **Dominance** reflects the degree of control a person feels over a situation, with lower values indicating feelings of powerlessness and higher values suggesting confidence and control.

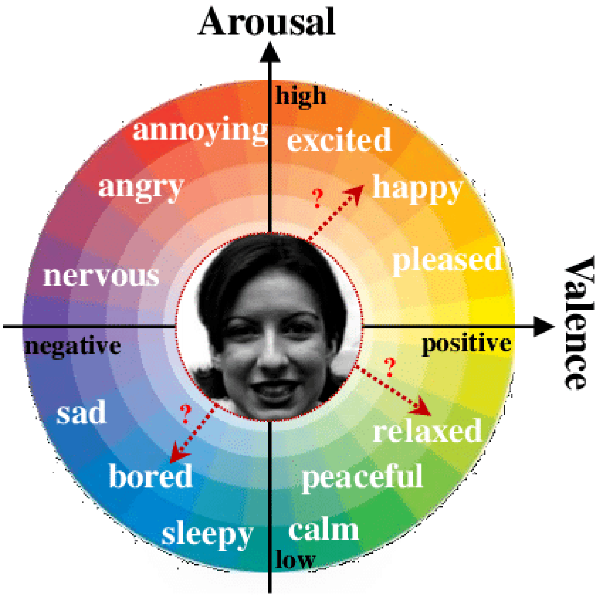


Figure 3.8: The Concept chart of (V,A,D) Model

Emotion Categories (discrete Dimensions)

the dataset incorporates a set of 26 discrete emotion categories. These categories were meticulously developed from an extensive affective lexicon sourced from psychological literature, encompassing around 400 terms that describe a broad spectrum of emotional states. Through a detailed analysis of definitions and semantic similarities, words were clustered into categories that are visually distinguishable within the context of a single image. The final selection of categories was guided by the principle of Visual Separability, ensuring that each category is uniquely identifiable despite closely related meanings. For example, the category of ‘Anger’ encompasses a spectrum of states such as rage, fury, and resentment, which, while distinct, may not always be visually discernible from one another.[93]

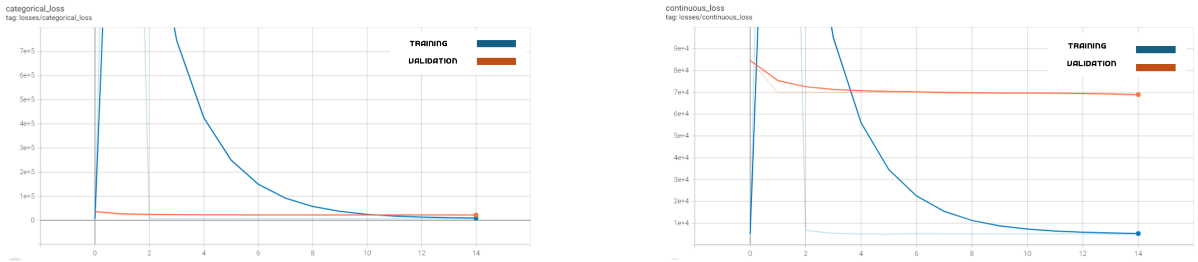
Emotion	**Description**
1.Affection	fond feelings; love; tenderness
2.Anger	intense displeasure or rage; furious; resentful
3.Annoyance	bothered by something or someone; irritated; impatient; frustrated
4.Anticipation	state of looking forward; hoping on or getting prepared for possible future events
5.Aversion	feeling disgust, dislike, repulsion; feeling hate
6.Confidence	feeling of being certain; conviction that an outcome will be favourable; encouraged; proud
7.Disapproval	feeling that something is wrong or reprehensible; contempt; hostile
8.Disconnection	feeling not interested in the main event of the surrounding; indifferent; bored; distracted
9.Disquietment	nervous; worried; upset; anxious; tense; pressured; alarmed
10.Doubt/Confusion	difficulty to understand or decide; thinking about different options
11.Embarrassment	feeling ashamed or guilty
12.Engagement	paying attention to something; absorbed into something; curious; interested
13.Esteem	feelings of favorable opinion or judgment; respect; admiration; gratefulness
14.Excitement	feeling enthusiasm; stimulated; energetic
15.Fatigue	weariness; tiredness; sleepy
16.Fear	feeling suspicious or afraid of danger, threat, evil or pain; horror
17.Happiness	feeling delighted; feeling enjoyment or amusement
18.Pain	physical suffering
19.Peace	well being and relaxed; no worry; having positive thoughts or sensations; satisfied
20.Pleasure	feeling of delight in the senses
21.Sadness	feeling unhappy, sorrow, disappointed, or discouraged
22.Sensitivity	feeling of being physically or emotionally wounded; feeling delicate or vulnerable
23.Suffering	psychological or emotional pain; distressed; anguished
24.Surprise	sudden discovery of something unexpected
25.Sympathy	state of sharing others' emotions, goals or troubles; supportive; compassionate
26.Yearning	strong desire to have something; jealous; envious; lust

Table 3.1: Descriptions of 26 Emotions Categories

3.11 Result and Analysis :

In this section, we present the results and analysis of the model's performance using two different loss functions: categorical loss and continuous loss SL1. The performance of the model is evaluated over a set number of epochs, and the results are depicted in the figures below. The categorical loss function is typically used for classification problems, whereas the continuous loss SL1 is used for regression tasks involving continuous values. The visualizations help to illustrate how well the model has learned from the training data and how it performs on the validation set. By analyzing these results, we can draw conclusions about the effectiveness of the training process and the suitability of each loss function for the given task.

3.11.1 Epoch15,Batch size 26:



(a) Categorical loss

(b) Continious loss SL1

Figure 3.9: Results of loss (Categorical and Continious SL1)

The effect of batch size on the total loss during training is illustrated in three distinct scenarios, each demonstrating unique characteristics and implications for model performance.

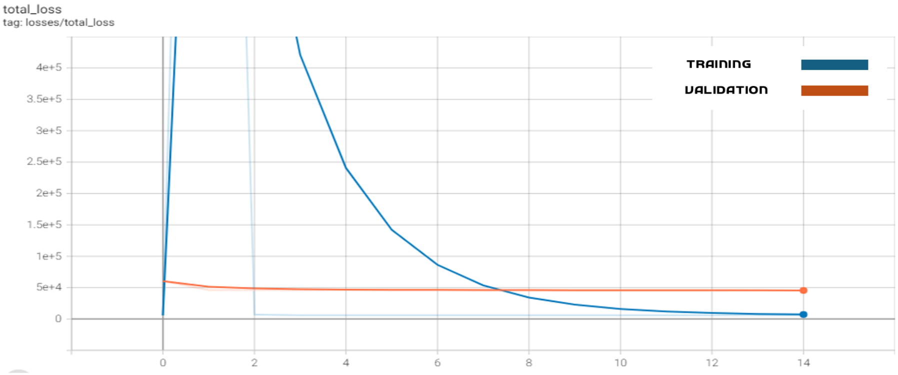
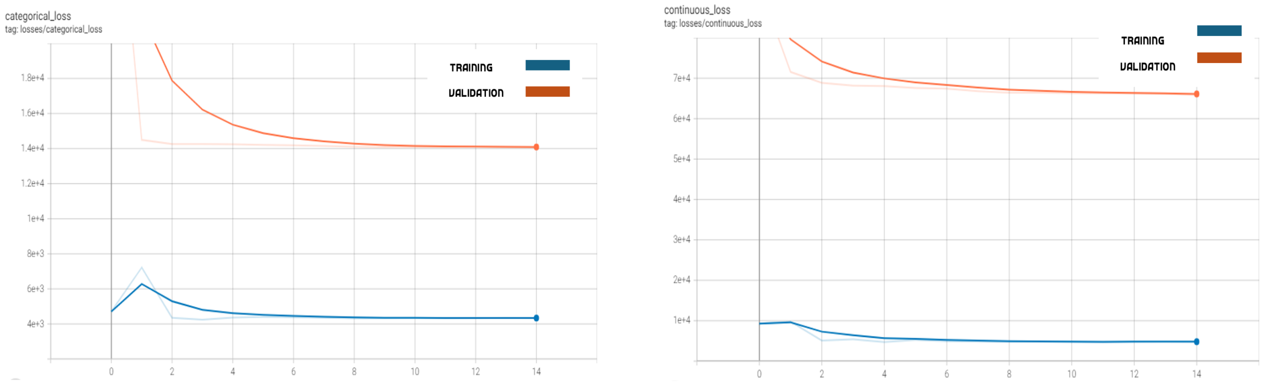


Figure 3.10: Total loss for(Categorical + Continious SL1)

Small Batch Size (26): As shown in Figure 3.9, the validation loss remains relatively stable and low from the beginning of the training process, even as the number of epochs increases. This indicates that with a small batch size of 26, the model is able to maintain a consistent level of performance early on. This stability suggests that the model can generalize well to new data from the outset. However, the lack of significant reduction in loss over time might imply that the model is not learning substantially more as training progresses. This could be beneficial in scenarios where quick and stable training is preferred, but it may limit the model's capacity to achieve deeper insights from the data.

3.11.2 Epoch15,Batch size 104:

Figure 3.10 illustrates the training loss curves for two different loss functions. Subfigure (a) presents the categorical loss used for classification tasks, and subfigure (b) shows the continuous SL1 loss used for regression tasks. Both plots demonstrate how the loss decreases over epochs, indicating the model's learning and convergence.



(a) Categorical loss

(b) Continious loss SL1

Figure 3.11: Results of loss (Categorical and Continious SL1)

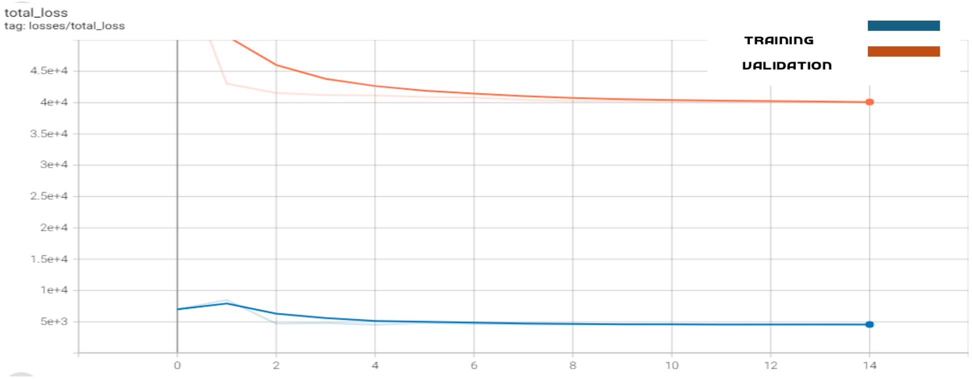
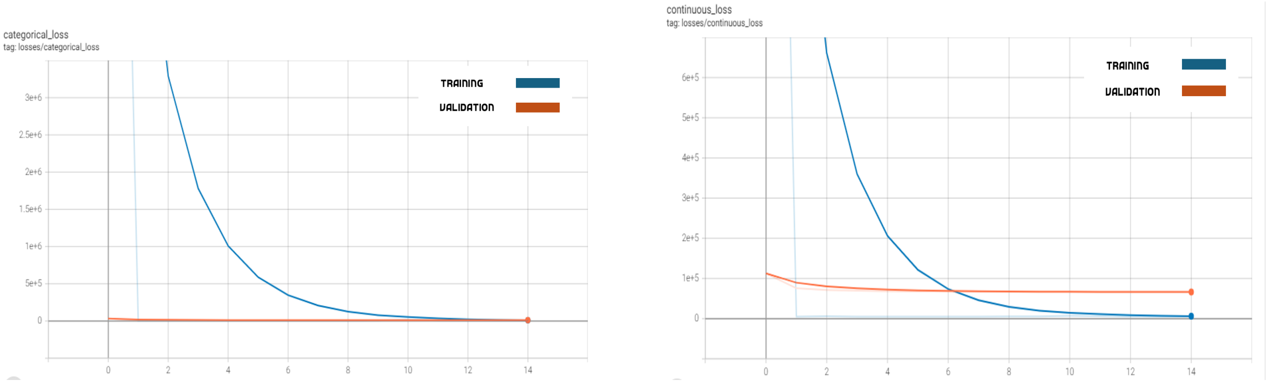


Figure 3.12: Total loss for(Categorical + Continious SL1)

Medium Batch Size (104): Figure 3.11 presents the scenario with a batch size of 104, where the total loss curve decreases slowly but steadily until it reaches a stable value. This gradual reduction in loss signifies that the model is learning more progressively with each epoch. The steady decline and eventual stabilization indicate a balanced learning process, where the model improves its performance incrementally. This approach can lead to a more robust model as it allows for thorough training without rapid fluctuations. Such a batch size is often chosen when a balance between computational efficiency and effective learning is needed.

3.11.3 Epoch15,Batch size 182:

Figure 3.12 shows the training loss curves for two types of loss functions. Subfigure (a) displays the categorical loss used in classification tasks, while subfigure (b) depicts the continuous SL1 loss used in regression tasks. Both plots highlight the reduction in loss over epochs, reflecting the model's improvement and convergence.



(a) Categorical loss

(b) Continious loss SL1

Figure 3.13: Results of loss (Categorical and Continious SL1)

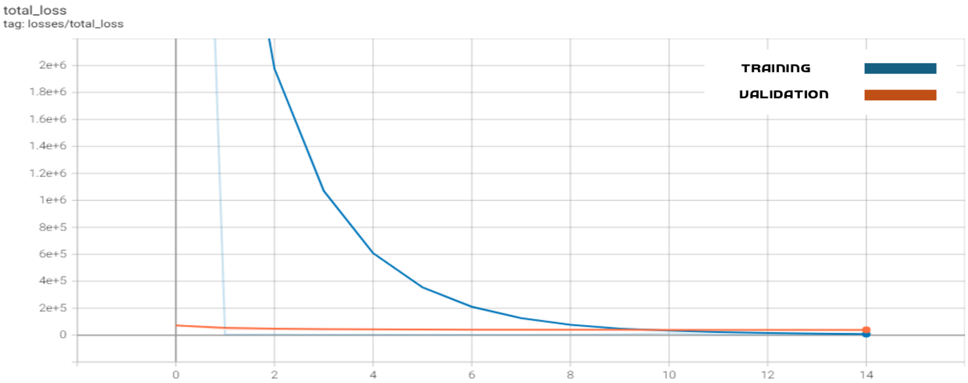


Figure 3.14: Total loss for(Categorical + Continious SL1)

Large Batch Size (182): In Figure 3.13, the training with a batch size of 182 shows the total loss remaining almost constant and near zero throughout the training process after the initial epochs. This rapid convergence to a very low loss value suggests that the model reaches a stable and optimal state very quickly. The minimal fluctuation in loss indicates a highly stable training process, which is desirable for applications requiring reliable and consistent performance. However, achieving this stability requires more computational resources and memory, as larger batch sizes demand greater processing power.

Summary on Batch Size Impact:

The analysis of different batch sizes reveals a trade-off between initial stability, learning dynamics, and resource requirements. Smaller batch sizes lead to quick stabilization but may limit further learning improvements. Medium batch sizes offer a balanced approach with gradual and steady learning, while larger batch sizes ensure rapid convergence and stability at the cost of increased computational demand. The choice of batch size should therefore be guided by the specific goals of the training process, the nature of the data, and the available computational resources.

Effect of Loss Function

The loss function is a critical component in deep learning models. It quantifies how well a model's predictions match the actual data, serving as a measure of error. The goal during training is to minimize this loss, thereby improving the model's precision.

3.11.4 Comparison Performance of Average precision (AP) for L2 cont and SL1cont:

the corresponding table shown a results of precision with two loss functions :

CNN Input and $\mathbf{L_{cont}}$ type		
Category	**SL1**	**L2**
Affection	**28.997**	27.616
Anger	**06.060**	5.908
Annoyance	11.352	**12.095**
Anticipation	93.297	**93.609**
Aversion	08.490	**8.878**
Confidence	72.837	**76.574**
Disapproval	10.229	**11.556**
Disconnection	**32.023**	30.928
Disquietment	**16.281**	15.319
Doubt/Confusion	13.339	**15.217**
Embarrassment	05.295	**5.503**
Engagement	**96.934**	95.835
Esteem	**21.791**	20.460
Excitement	72.734	73.126
Fatigue	**08.350**	7.817
Fear	**05.565**	5.286
Happiness	70.966	**71.572**
Pain	**07.369**	6.903
Peace	20.822	**21.826**
Pleasure	**40.468**	39.949
Sadness	06.303	**8.266**
Sensitivity	04.894	**5.240**
Suffering	06.720	**8.022**
Surprise	**11.643**	11.419
Sympathy	**26.769**	26.688
Yearning	10.649	**10.747**
Mean AP	27.315	**27.552**

Table 3.2: Average precision (AP) for L2 cont and SL1cont

In Table 3.2 The table presents the precision of various emotion categories measured using two different loss functions: SL1 and L2. Below is a comparison of how each emotion category performs with these loss functions.We noticed Annoyance, Doubt/Confusion, Sadness, Suffering: These categories have higher precision with the L2 loss function. Anticipation, Confidence, Disapproval, Excitement, Happiness, Peace: These categories also show slightly higher precision with the L2 loss function.Engagement: Shows a slight decrease in precision with the L2 loss function.Affection, Anger, Disconnection, Disquietment, Esteem, Fatigue, Fear, Pain, Pleasure: Precision values for these categories are relatively similar for both loss functions, with minor differences.The L2 loss function generally provides slightly higher precision overall, and it performs better in certain emotion categories such as Annoyance, Sadness, and Suffering. The choice of loss function affects the training process, influencing the convergence rate, stability, and overall performance of the model.

3.11.5 Comparison Performance Average absolute Error(AE) for L2 cont and SL1cont:

CNN Input and $\mathbf{L_{cont}}$ type		
Continious Dimenssion	**ESL1**	**EL2**
Valence	0.71471	0.69993
Arousal	0.91511	0.85898
Dominance	0.89981	0.81298
Mean VAD Error	0.84321	0.79063

Table 3.3: Average absolute Error (AE) for comparing Performance of each with L2 cont and SL1cont

The tabel Shows that the less value of the Mean VAD Erorr is Less is more better ,This demonstrates that the Lowest value is in the Mean VAD Erorr of the loss (**L2**)

3.11.6 Comparison Results of Average precision (AP) for B(SL1) and B+I(SL1):

Input to the network		
Category	**Body** **B(SL1)**	**Body+ Context** **B+I(SL1)**
Affection	16.55	**27.85**
Anger	04.67	**09.49**
Annoyance	05.54	**14.06**
Anticipation	56.61	**58.64**
Aversion	03.64	**07.48**
Confidence	72.57	**78.35**
Disapproval	05.50	**14.97**
Disconnection	16.12	**21.32**
Disquietment	13.99	**16.89**
Doubt/Confusion	28.35	**29.63**
Embarrassment	02.15	**03.18**
Engagement	84.59	**87.53**
Esteem	**19.48**	17.73
Excitement	71.80	**77.16**
Fatigue	06.55	**09.70**
Fear	12.94	**14.14**
Happiness	51.56	**58.26**
Pain	02.71	**08.94**
Peace	17.09	**21.56**
Pleasure	40.98	**45.46**
Sadness	06.19	**19.66**
Sensitivity	03.60	**09.28**
Suffering	04.38	**18.84**
Surprise	17.03	**18.81**
Sympathy	09.35	**14.71**
Yearning	07.40	**08.34**
Mean AP	22.36	**27.38**

Table 3.4: Average precision (AP) for just Body =B(SL1) and Bodycontext =B+I(SL1)

In the corresponding table results of experiments conducted as prior works where the model was divided into parts to see how important the visual context for emotion recognation we noticed The mean average precision (Mean AP) improved from 22.36 to 27.38 when using Body + Context compared to Body alone.

Adding context to body language significantly improves the precision of emotion detection across most categories, especially for emotions like sadness, disapproval, and suffering.

However, for some emotions like anger and aversion, the precision remains relatively low even with the added context.

The overall mean average precision indicates a notable improvement when context is included.herewe see the importance of context for emotion recognation .

3.12 Comparison of our proposed approach with last proposed approach

	Input to the network			
CTEGORIES	**SL1**	**Our SL1**	**L2**	**Our L2**
Affection	27.85	28.997	21.16	27.616
Anger	09.49	06.060	06.45	5.908
Annoyance	14.06	11.352	11.18	12.095
Anticipation	58.64	93.297	58.61	93.609
Aversion	07.48	08.490	06.45	8.878
Confidence	78.35	72.837	77.97	76.574
Disapproval	14.97	10.229	11.00	11.556
Disconnection	21.32	32.023	20.37	30.928
Disquietment	16.89	16.281	15.54	15.319
Doubt/Confusion	29.63	13.339	28.15	15.217
Embarrassment	03.18	05.295	02.44	5.503
Engagement	87.53	96.934	86.24	95.835
Esteem	17.73	21.791	17.35	20.460
Excitement	77.16	72.734	76.96	73.126
Fatigue	09.70	08.350	08.87	7.817
Fear	14.14	05.565	12.34	5.286
Happiness	58.26	70.966	60.69	71.72
Pain	08.94	07.369	04.42	6.903
Peace	21.56	20.822	19.43	21.826
Pleasure	45.46	40.468	42.12	39.949
Sadness	19.66	06.303	10.36	8.266
Sensitivity	09.28	04.894	04.82	5.240
Suffering	18.84	06.720	07.65	8.022
Surprise	18.81	11.643	16.42	11.419
Sympathy	14.71	26.769	11.44	26.688
Yearning	08.34	10.649	08.34	10.747
Mean AP	27.38	27.315	24.88	27.552

the table comparison demonstrates that our results have substantially improved precision across most emotion categories compared to previous works, especially for emotions like

Anticipation, Engagement, and Sensitivity. This indicates that there are some factors that affect the precision of the model , we will mention some of them :

Computational Resources:

Hardware: Efficient utilization of computational resources can reduce training time and improve model performance.

The graphics card (GPU): plays a critical role in the precision and overall performance of machine learning models, particularly in deep learning tasks.The use of a powerful graphics card can significantly enhance the precision of machine learning models by enabling faster, more efficient training processes, handling larger datasets, and allowing for more complex and well-tuned model architectures. This results in better convergence, reduced overfitting, and ultimately, higher precision in model predictions.

Hyperparameter Tuning:

Optimal Hyperparameters: Carefully tuning hyperparameters (e.g., learning rate, number of layers, number of neurons) can significantly impact model precision.

Search Techniques: Grid search, random search, and Bayesian optimization are methods to find the best hyperparameters.

Training Process:

Training Time: Sufficient training time ensures that the model learns well from the data.

Batch Size: The size of the data batches used in training can affect model performance. Larger batch sizes can speed up training but might lead to less precise models.

Early Stopping: This technique prevents overfitting by stopping the training when the model's performance on validation data starts to degrade

3.13 Application

To bring the advancements of our research into practical use, we have developed a comprehensive application that leverages the enhanced emotion recognition capabilities derived from our approach.

This application integrates the context-aware models trained on the EMOTIC dataset, providing a robust platform for detecting and interpreting human emotions in real-time, across diverse environments.

The application is designed to function seamlessly in natural settings, overcoming the limitations of traditional facial expression analysis by incorporating contextual information from the scene. It utilizes advanced Convolutional Neural Networks (CNNs) to process both the facial features and the surrounding context, ensuring a more accurate and nuanced understanding of emotional states.

Key features of the application include real-time emotion detection, user-friendly interface, and adaptability to various use cases such as mental health monitoring, customer service enhancement, and interactive entertainment. By combining cutting-edge AI technology with practical usability, this application exemplifies the potential of context-aware emotion recognition to transform human-computer interactions and provide valuable insights across different domains.

3.14 Hardware Used:

System Processor: Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50 GHz.

Installed memory (RAM): 8.00 GB (7.88 GB usable)

System type:64-bit Operating System, x64- based processor

Pen and Touch: No Pen or Touch Input is available for this Display Computer name, domain and workgroup settings.

Computer name: DESKTOP-QK358HB

Full computer name: DESKTOP-OK358HB

3.15 Software Requirements:

Programming language python3.

Libraries: pytorch / opencv / numpy.

YOLO v3.

3.15.1 Input Requirements:

Images, video, Gif.

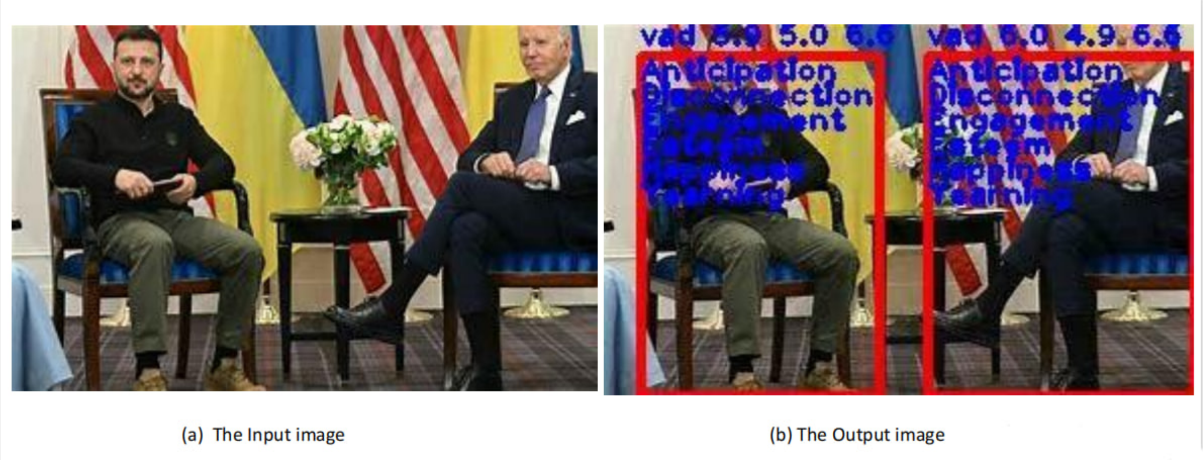
Gif: Graphics Interchange Format.

3.15.2 Output Requirements:

Image, AVI.

AVI: Audio Video Interleave.

3.15.3 A practical example of our app :

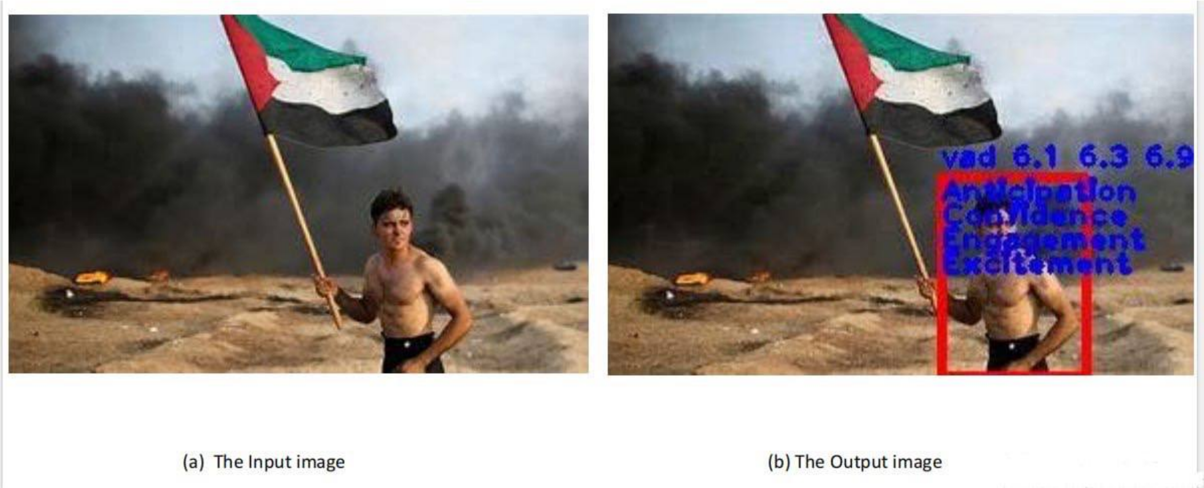


(a) The Input image

(b) The Output image

Figure 3.15: image of Two seated individuals in official discussion before and after applying the app.

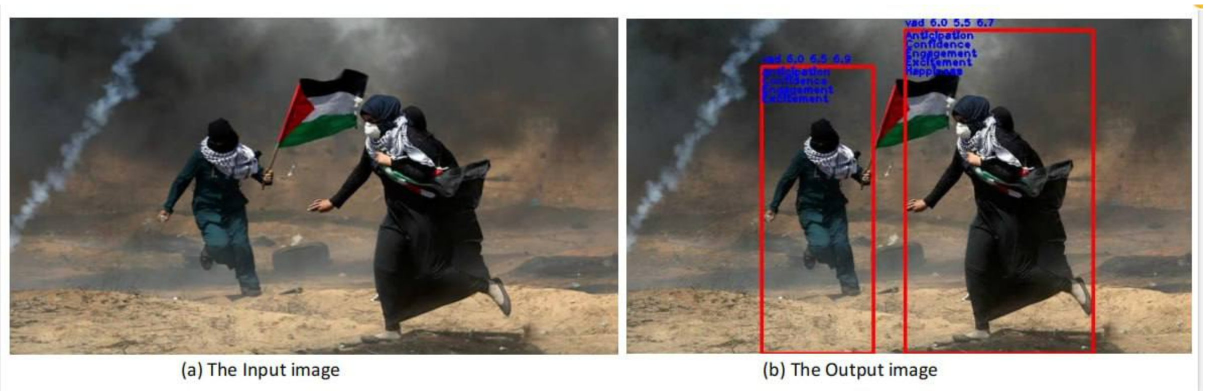
The image (a) shows two seated individuals in official discussion. the image (b) shows the image (a) After applying the app ,Both individuals are outlined in red boxes (Bonding Box) with text indicating discrete categories (various emotions are the same :Anticipation,Disconnection ,Engagement,Esteem,Happiness and Yearning) and values of continuous dimensions (VAD between 1 to 10).



(a) The Input image (b) The Output image

Figure 3.16: image of a person holding a Palestinian flag in what appears to be a protest or conflict zone.

The image (a) shows a person holding a Palestinian flag in what appears to be a protest or conflict zone, as indicated by the smoke and desolate landscape. the image (b) shows the image (a) After applying the app ,the person are outlined in red box (Bonding Box) with text indicating discrete categories (various emotions are :Anticipation ,Confidence , Engagement and Excitement) and values of continuous dimensions(VAD between 1 to 10).

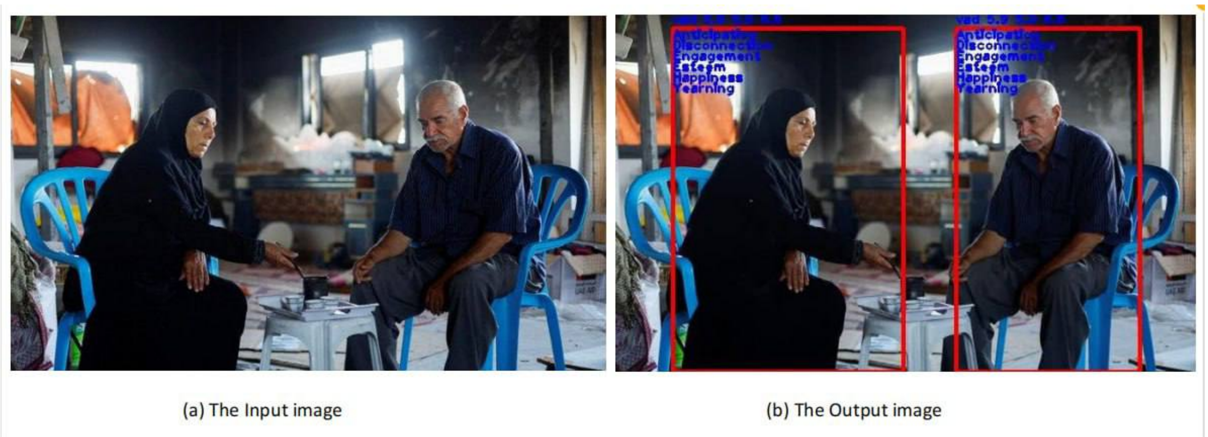


(a) The Input image (b) The Output image

Figure 3.17: image of two individuals running,in a protest or conflict situation

The image (a) shows two individuals running, possibly in a protest or conflict situation. They are carrying Palestinian flags and appear to be wearing scarves over their faces, likely for protection against smoke or tear gas seen in the background. the image (b) shows the image (a) After applying the app ,Both individuals are outlined in red boxes (Bonding Box) with text indicating discrete categories (various emotions are: Anticipation ,Confidence , Engagement and Excitement) and values of continuous dimensions

(VAD between 1 to 10).



(a) The Input image

(b) The Output image

Figure 3.18: image of two individuals, a woman and a man, sitting in an indoor setting.

The image (a) shows two individuals, a woman and a man, sitting in what appears to be an indoor setting. the image (b) shows the image (a) After applying the app ,Both individuals are outlined in red boxes (Bonding Box) with text indicating discrete categories (various emotions are: Anticipation ,Disconnection , Engagement ,Esteem ,Happiness and Yearning) and values of continuous dimensions(VAD between 1 to 10).

3.16 Conclusion:

In this chapter, we analyzed the performance of the proposed model using various loss functions and discussed the implications of the results. The categorical loss effectively demonstrated the model's ability to learn and classify the data accurately, with a noticeable reduction in loss over the training epochs. Similarly, the continuous SL1 loss highlighted the model's proficiency in handling regression tasks, with a consistent decrease in loss indicating improved predictive accuracy.

The results from these experiments confirm that the model is robust and capable of learning from both classification and regression tasks. The decreasing loss curves across epochs signify effective training and convergence of the model. These findings underscore the model's versatility and potential application in various real-world scenarios.

General Conclusion

This thesis has presented a comprehensive approach to emotion recognition by integrating contextual information using Convolutional Neural Networks (CNN) and Residual Networks (ResNet) architectures, trained on the Emotic dataset. Our model, capable of producing dual outputs—continuous dimensional emotion representation and discrete categorical emotion classification—has demonstrated significant improvements in accurately detecting human emotions.

The incorporation of contextual cues has been shown to enhance the robustness and precision of emotion recognition systems, making them more reliable for practical applications. The Emotic dataset, with its extensive annotations, has served as a crucial resource for developing and evaluating our models. By leveraging the deep feature extraction capabilities of CNN and ResNet, our approach has successfully captured nuanced emotional cues from visual data, achieving state-of-the-art performance.

The implications of this research are vast, offering significant advancements in various fields. Enhanced emotion recognition systems can lead to more natural and effective human-computer interactions, improved mental health monitoring and intervention, more empathetic social robots, and more engaging and responsive educational tools. The ability to accurately detect and respond to human emotions can transform the way technology interacts with users, making it more intuitive and human-centric.

Future Work

While this research has laid a strong foundation, there are several promising directions for future work that can further enhance the applicability and impact of emotion recognition systems:

- **Multimodal Integration:** Future research should explore integrating additional data modalities such as audio, text, and physiological signals. Combining these with visual data can provide a more comprehensive understanding of emotional states, improving the accuracy and reliability of emotion recognition systems in diverse applications.
- **Real-time Implementation:** Developing real-time emotion detection capabilities is crucial for applications requiring immediate emotional feedback, such as interactive gaming, live customer support, and real-time therapeutic interventions. Optimizing models for low-latency performance will be essential.
- **Personalization and Adaptation:** Emotion recognition systems can benefit from personalization, adapting to individual differences in emotional expression. Future work could focus on creating adaptive models that learn from user-specific data, enhancing accuracy and user satisfaction in applications like personalized mental health support and adaptive learning environments.
-
- Improving the contextual analysis capabilities of emotion recognition systems by incorporating advanced techniques such as scene understanding and activity recognition can provide deeper insights into emotional states, enhancing applications in surveillance, security, and human behavior analysis.
- **Ethical Considerations and Bias Mitigation:** Addressing ethical concerns and mitigating biases in emotion recognition systems is paramount. Future research should develop fair and unbiased models that respect user privacy and ensure equitable treatment across different demographic groups, which is essential for applications in sensitive areas such as healthcare and law enforcement.
- **Deployment in Real-world Scenarios:** Implementing and testing emotion recognition systems in real-world applications will provide valuable feedback for further refinement. This includes applications in healthcare monitoring systems, educational tools, customer service bots, and social robots, where accurate emotion

detection can significantly enhance user experience and outcomes.

By pursuing these directions, future research can build upon the advancements made in this thesis, creating more accurate, robust, and ethically sound emotion recognition systems. These systems have the potential to revolutionize interactions across various domains, making technology more responsive and attuned to human emotional needs.

Bibliography

[1] Maja Pantic and Leon JM Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[2] Zisheng Li, Jun-ichi Imai, and Masahide Kaneko. Facial-component-based bag of words and phog descriptor for facial expression recognition. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 1353–1358. IEEE, 2009.

[3] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[4] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[5] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 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, 2016.

[6] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2015.

[7] Albert Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 121(3):339–361, 1995.

[8] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

[9] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE*

Transactions on Affective Computing, 2(2):92–105, 2011.

[10] Konrad Schindler, Luc Van Gool, and Beatrice De Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9):1238–1246, 2008.

[11] Wenxuan Mou, Oya Celiktutan, and Hatice Gunes. Group-level arousal and valence recognition in static images: Face, body and context. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 5, pages 1–6. IEEE, 2015.

[12] Gheorghe Tecuci. Artificial intelligence. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4, 03 2012.

[13] Bruce G Buchanan. A (very) brief history of artificial intelligence. *Ai Magazine*, 26(4):53–53, 2005.

[14] Narendra Kumar, Nidhi Kharkwal, Rashi Kohli, and Shakeeluddin Choudhary. Ethical aspects and future of artificial intelligence. In *2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*, pages 111–114. IEEE, 2016.

[15] Vincent C Müller and Nick Bostrom. Future progress in artificial intelligence: A survey of expert opinion. *Fundamental issues of artificial intelligence*, pages 555–572, 2016.

[16] Bostrom Nick. Superintelligence: Paths, dangers, strategies. 2014.

[17] J Philip Craiger and Diane Maye Zorri. Current trends in small unmanned aircraft systems: Implications for us special operations forces. 2019.

[18] GM Del Prado. 18 artificial intelligence researchers reveal the profound changes coming to our lives. *Business Insider, available at: www. businessinsider. com/researchers-predictionsfuture-artificial-intelligence-2015-10 (accessed 10 September 2018)*, 2015.

[19] Bernard Marr. How is ai used in education–real world examples of today and a peek into the future. *Forbes, Forbes Magazine*, 25, 2018.

[20] Keng Siau. Education in the age of artificial intelligence: how will technology shape learning? 2018.

[21] Ion Stoica, Dawn Song, Raluca Ada Popa, David Patterson, Michael W Mahoney, Randy Katz, Anthony D Joseph, Michael Jordan, Joseph M Hellerstein, Joseph E

Gonzalez, et al. A berkeley view of systems challenges for ai. *arXiv preprint arXiv:1712.05855*, 2017.

[22] Forbes Insights and AI Intel. How ai builds a better manufacturing process. *Forbes, July*, 17, 2018.

[23] Forbes Insights and AI Intel. How ai builds a better manufacturing process. *Forbes, July*, 17, 2018.

[24] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, 2017.

[25] Kris Newby. Compassionate intelligence: Can machine learning bring more humanity to health care?, 2018.

[26] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. Crime data mining: a general framework and some examples. *computer*, 37(4):50–56, 2004.

[27] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[28] Colin Lewis and Dagmar Monett. Getting clarity by defining artificial intelligence—a survey. 2017.

[29] K Wakefield. A guide to the types of machine learning algorithms and their applications. *URL: https://www. sas. com/en_gb/insights/articles/analytics/machine-learning-algorithms. html [Accessed on 10 February 2021]*, 2021.

[30] Tom M Mitchell et al. Machine learning. 1997.

[31] IBM. Que sont les réseaux neuronaux ? https://www.ibm.com/fr-fr/topics/neural-networks. Accessed: 2024-06-02.

[32] MathWorks. What is deep learning? https://fr.mathworks.com/discovery/deep-learning.html, 2024. Accessed: 2024-06-04.

[33] Maria V Valueva, NN Nagornov, Pavel Alekseevich Lyakhov, Georgii V Valuev, and Nikolay I Chervyakov. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and computers in simulation*, 177:232–243, 2020.

[34] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017 international conference on engineering and technology (icet). *Understanding of a Convolutional Neural Network, Antalya*, pages 1–6, 2017.

[35] Van Hiep Phung and Eun Joo Rhee. A deep learning approach for classification of cloud image patches on small datasets. *Journal of information and communication convergence engineering*, 16(3):173–178, 2018.

[36] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.

[37] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[38] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Hervé Rostaing. *Veille technologique et bibliométrie: concepts, outils, applications.* PhD thesis, Université Paul Cézanne d'Aix-Marseille, 1993.

[42] Florent Perronnin and Jean-Luc Dugelay. Introduction à la biométrie authentification des individus par traitement audio-vidéo. *Traitement du signal*, 19(4), 2002.

[43] Larbi NOUAR. *Identification biométrique par fusion multimodale.* PhD thesis, Université de Sidi Bel Abbès-Djillali Liabes, 2018.

[44] Davide Maltoni, Dario Maio, Anil K Jain, Salil Prabhakar, et al. *Handbook of fingerprint recognition*, volume 2. Springer, 2009.

[45] Nesrine Charfi. *Biometric recognition based on hand schape and palmprint modalities.* PhD thesis, Ecole nationale supérieure Mines-Télécom Atlantique, 2017.

[46] Florent Perronnin and Jean-Luc Dugelay. Introduction à la biométrie authentification des individus par traitement audio-vidéo. *Traitement du signal*, 19(4), 2002.

[47] Souhila Guerfi Ababsa. Authentification d'individus par reconnaissance de caractéristiques biométriques liées aux visages 2d/3d. *Evry-Val d'Essonne*, 2008.

[48] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20,

2004.

[49] Nesrine Charfi. *Biometric recognition based on hand schape and palmprint modalities.* PhD thesis, Ecole nationale supérieure Mines-Télécom Atlantique, 2017.

[50] Lamiaa A Elrefaei, Doaa H Hamid, Afnan A Bayazed, Sara S Bushnak, and Shaikhah Y Maasher. Developing iris recognition system for smartphone security. *Multimedia Tools and Applications*, 77:14579–14603, 2018.

[51] Ying Li Han, Tae Hong Min, and Rae-Hong Park. Efficient iris localisation using a guided filter. *IET Image Processing*, 9(5):405–412, 2015.

[52] Tripti Rani Borah, Kandarpa Kumar Sarma, and Pran Hari Talukdar. Retina recognition system using adaptive neuro fuzzy inference system. In *2015 International Conference on Computer, Communication and Control (IC4)*, pages 1–6. IEEE, 2015.

[53] Bijay K Ekka, NB Puhan, and Rashmi Panda. Retinal verification using point set matching. In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 159–163. IEEE, 2015.

[54] Preeti Saini and Parneet Kaur. Automatic speech recognition: A review. *International Journal of Engineering Trends and Technology*, 4(2):1–5, 2013.

[55] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004.

[56] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li. Speech emotion recognition using fourier parameters. *IEEE Transactions on affective computing*, 6(1):69–75, 2015.

[57] Margit Antal, László Zsolt Szabó, and Tünde Tordai. Online signature verification on mobisig finger-drawn signature corpus. *Mobile Information Systems*, 2018:1–15, 2018.

[58] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer vision and image understanding*, 167:1–27, 2018.

[59] Anil K Jain, Ruud Bolle, and Sharath Pankanti. *Biometrics: personal identification in networked society*, volume 479. Springer Science & Business Media, 2006.

[60] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20,

2004.

[61] Stan Z Li and Anil Jain. *Encyclopedia of biometrics.* Springer Publishing Company, Incorporated, 2015.

[62] P Inbavalli and G Nandhini. Body odor as a biometric authentication. *International Journal of Computer Science and Information Technologies*, 5(5):6270–6274, 2014.

[63] Alfred Victor Iannarelli. *The Iannarelli system of ear identification.* Foundation Press, 1964.

[64] Lamis Ghoualmi, Amer Draa, and Salim Chikhi. An ear biometric system based on artificial bees and the scale invariant feature transform. *Expert Systems with Applications*, 57:49–61, 2016.

[65] MP Reshmi and VJ Arul Karthick. Biometric identification system using lips. *Int. J. Sci. Res.(IJSR)*, 2(4):304–307, 2013.

[66] Marc Schröder, Laurence Devillers, Kostas Karpouzis, Jean-Claude Martin, Catherine Pelachaud, Christian Peter, Hannes Pirker, Björn Schuller, Jianhua Tao, and Ian Wilson. What should a generic emotion markup language be able to represent? In *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*, pages 440–451. Springer, 2007.

[67] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1667–1675, 2017.

[68] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017.

[69] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[70] Albert Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 121(3):339–361, 1995.

[71] C Darwin and P Prodger. The expression of the emotions in man and animals. oxford university press, usa. 1998.

[72] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus.

Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems*, 27, 2014.

[73] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[74] Carroll E Izard. The face of emotion. 1971.

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[76] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[77] Farhana Sultana, Abu Sufian, and Paramartha Dutta. A review of object detection models based on convolutional neural network. *Intelligent computing: image processing based applications*, pages 1–16, 2020.

[78] Wang Zhiqiang and Liu Jun. A review of object detection based on convolutional neural network. In *2017 36th Chinese control conference (CCC)*, pages 11104–11109. IEEE, 2017.

[79] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[80] Junyan Lu, Chi Ma, Li Li, Xiaoyan Xing, Yong Zhang, Zhigang Wang, and Jiuwei Xu. A vehicle detection method for aerial image based on yolo. *Journal of Computer and Communications*, 6(11):98–107, 2018.

[81] Rachel Huang, Jonathan Pedoeem, and Cuixian Chen. Yolo-lite: a real-time object detection algorithm optimized for non-gpu computers. In *2018 IEEE international conference on big data (big data)*, pages 2503–2510. IEEE, 2018.

[82] Bo Gong, Daji Ergu, Ying Cai, and Bo Ma. A method for wheat head detection based on yolov4. 2020.

[83] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[84] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information*

processing systems, 28, 2015.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[86] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[87] Oliver Kramer and Oliver Kramer. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53, 2016.

[88] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Domputing*, 1(1):105–115, 2019.

[89] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

[90] Junjie Bai, Fang Lu, Ke Zhang, et al. Onnx: Open neural network exchange, 2019.

[91] Ayush Shridhar, Phil Tomson, and Mike Innes. Interoperating deep learning models with onnx. jl. In *Proceedings of the JuliaCon Conferences*, volume 1, page 59, 2020.

[92] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[93] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1667–1675, 2017.

[94] Ronak Kosti. *Visual scene context in emotion perception.* PhD thesis, Universitat Oberta de Catalunya, 2019.